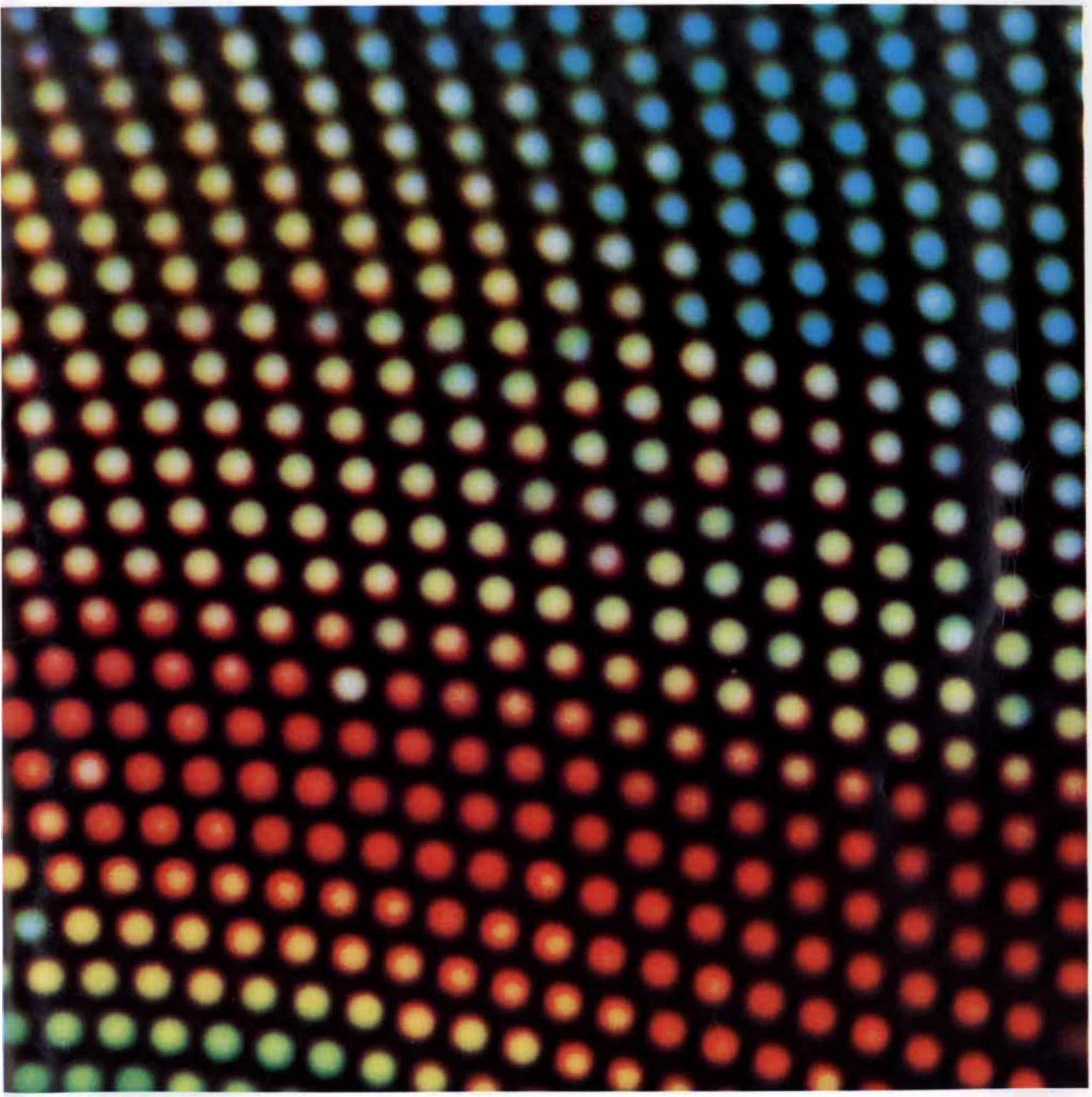
SCIENTIFIC AMERICAN



INSECT EYE

$1.50

July 1977


"WHAT DOES THE MAN WHO DRIVES A PORSCHE BUY FOR HIS FAMILY?"
Twenty-five years of racing and research have come down to this. Porsche 911S Targa. Ultimate perfection of the engineer's art.
Evidence of Porsche's achievement is abundant within the car, and within the record books. The 911 series is both cause and effect of one of the most successful racing campaigns ever waged. And they are perhaps the most elegant grand touring automobiles ever built.
Audi
FOX


What does the Porsche owner buy for his family? What can meet his family's needs, yet satisfy a personal demand for engineering, performance, and handling?
Nothing comes closer than the Audi Foxwagon.
Audi, with Porsche, shares a dedication to engineering excellence that is all too rare. Both cars have high-performance overhead cam engines with CIS fuel injection. Both have rack-and-pinion steering. And both have high strength, lightweight construction.
In addition, the Foxwagon boasts 51.6 cubic feet of space with the rear seat folded down. Plus orthopedically designed leatherette seats, flow-through ventilation, and child-proof rear door locks for safety.
The 1977 Audi Foxwagon.
There's no better way for a Porsche driver to fill a two-car garage.
DRIVE ONE OF OUR LEGENDS
PORSCHE
+AUDI

White rum and Canada Dry® tonic. The making of a "classic."

Through the years the summer heat has inspired an endless variety of drinks, but few have ever attained the rank of "classic."

So it pleases us immensely to observe that white rum and Canada Dry tonic shows all the signs of becoming an enduring favorite.

We base that feeling on the fact that a national taste test has shown us that Puerto Rican white rum is preferred over gin and vodka for taste and smoothness. And that white Puerto Rican rum is the fastest growing distilled spirit in America.

In addition we know that more drinkers prefer Canada Dry than any other tonic.

The two of them together are an unbeatable team—with all the ingredients for the making of a "classic."

The reasons that lie behind these preferences are clear enough. Drinkers appreciate quality.

Breeding shows

White rum, for example, is the beneficiary of a great deal of care and attention. More so than either gin or vodka. Only white rum, of the three, is aged for smoothness, spending a minimum of a year in oak casks, by Puerto Rican law. By the time it emerges, not even the faintest trace of a rough edge remains.

A perfect combination

Canada Dry, for its part, is the perfect partner. Providing its own distinctively smooth taste, created by a unique blending of citrus flavors and quinine.

We suggest that you sample the pleasures of white rum and Canada Dry tonic for yourself. When you taste this new "classic" you'll discover what a tonic drink was meant to be.

PUERTO RICAN RUMS

For free "White Rum Classics" recipes, write: Puerto Rican Rums, Dept. SC-4, 1290 Avenue of the Americas, N.Y., N.Y. 10019

Established 1845 SCIENTIFIC AMERICAN July 1977 Volume 237 Number 1

ARTICLES

DEPARTMENTS

PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $18 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $22 PER YEAR, ALL OTHER COUNTRIES.

Physics...analytical math... astronomy and cosmology... chemistry...computer sciences... structural geology...undoubtedly the best books in every scientific area that interests you.

Take any 3 books, values to $67.50, all for only $3.95 as a member of

The Library of Science

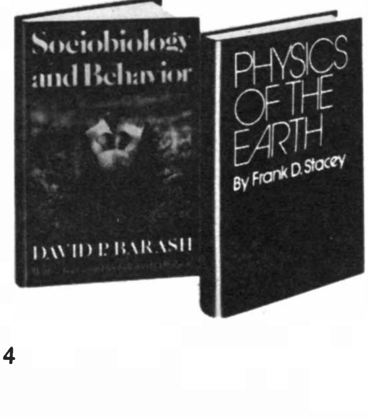
if you will join now for a trial period and agree to accept only 3 more books—at members' discount prices—over the next 12 months

(Publishers' Prices shown)

79140. SOCIOBIOLOGY AND BEHAVIOR. *David T. Barash*. New, highly controversial way of looking at behavior. A groundbreaking, thought provoking discussion of the possible genetic influences on human behavior. **$9.95**

72995. PULSARS. *F. G. Smith, F.R.S.* A concise look into the major topics in pulsar astrophysics. Contains 89 line diagrams and graphs. **$19.95**

68780. PHYSICS OF THE EARTH. *Frank D. Stacey.* Second edition of this highly acclaimed work includes the first comprehensive look at rotation and excitation of the wobble, tidal friction, earthquake prediction, thermodynamics of convection and melting and anelasticity. **$16.95**



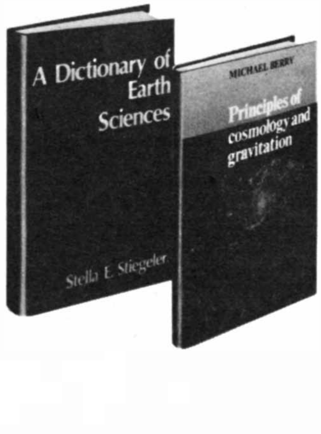
36785. THE BOWELS OF THE EARTH. *John Elder.* The first systematic text on the dynamics of the earth's interior, which joins together crustal geology and geophysics. **$14.00**

42205. A DICTIONARY OF EARTH SCIENCES. *Edited by Stella Stiegler.* An all-in-one work covering geophysics, structural geology, geomorphology, oceanography, stratigraphy, paleontology, sedimentology, climatology, meteorology, cartography, mineralogy, petrology and more. **$16.50**

42235. A DICTIONARY OF PHYSICAL SCIENCES. *Edited by John Daintith.* Physics, chemistry, astronomy, electronics, computer sciences and mathematics are covered by definitions, formulas and diagrams in this 352-page work of 3,000 entries. **$16.50**

36512. THE BOG PEOPLE/THE MOUND PEOPLE. *P. V. Glob.* Perfectly preserved 2,000-year-old bodies buried in peat bogs, all murdered . . . 6-ft-long oak coffins hidden within mounds of earth . . . two ancient mysteries solved, revealing two cultures, quite different but equally fascinating. Richly illustrated. *The set counts as 2 of your 3 books.* **$25.00**

66570. AN OUTLINE OF STRUCTURAL GEOLOGY. *Bruce E. Hobbs, Winthrop D. Means and Paul F. Williams.* Up-to-date, thorough account of every aspect—from the nature of stress and strain to the origin of folds to the latest information on plate tectonics. **$16.95**



61530. MATHEMATICS DICTIONARY. Fourth Edition. *Edited by Robert C. James and Edwin F. Beckenbach.* Revised and expanded edition of this leading mathematics dictionary offers clear, accurate definitions of more than 8,000 terms. **$17.95**

64145. A NEW DICTIONARY OF PHYSICS. *H. J. Gray and Alan Isaacs.* More than 9,000 entries provide a comprehensive, and, where necessary, detailed focus on key concepts, formulas, men and machines that are crucial to understanding modern physics. Updated reference contains hundreds of illustrations, graphs, charts and tables. *Counts as 2 of your 3 books.* **$35.00**

55000. THE ILLUSTRATED ENCYCLOPEDIA OF ASTRONOMY AND SPACE. *Ian Ridpath.* Covers the full range of astronomical knowledge and space exploration, from ancient times to now. **$16.95**

34670. ASTRONOMY AND COSMOLOGY. *Sir Fred Hoyle.* One of the giants of modern astronomy clearly explains today's astronomy, and the physical processes and relationships underlying basic astronomical phenomena ranging from the earth and the sun to the origin of the universe. Over 600 illustrations. **$16.50**

42210. A DICTIONARY OF LIFE SCIENCES. *Edited by Elizabeth Martin.* From *Abscisic acid* to *Zymogen,* this work covers taxonomy, zoology, botany, anatomy, physiology, cytology, genetics and ecology with 3,300 entries. **$16.50**

55190. THE INCREDIBLE DR. MATRIX. *Martin Gardner.* The delightful, incredibly brilliant Dr. Matrix returns with more amazing feats and adventures in number theory, sleight of word, numerological analysis and puzzles. **$8.95**

70190. PRINCIPLES OF OPERATIONS RESEARCH. Second Edition. *Harvey M. Wagner.* Based on the author's twenty years of experience applying operations research to a wide variety of problems, this 1039-page book offers a complete overview of all the fundamental concepts. **$19.95**

34140. APPLIED ANALYTICAL MATHEMATICS FOR PHYSICAL SCIENTISTS. *James T. Cushing.* This 650-page work for scientists deals with topics ranging from spectral analysis of linear operators to complex variables, to group theory. **$19.95**

45430. ENVIRONMENTAL CHEMISTRY. *Edited by J. O'M. Bockris.* A veritable encyclopedia of what's known about the environment, as well as an 800-page text covering the full spectrum of environmental issues. *Counts as 3 of your 3 books.* **$49.50**

59800. MAN DISCOVERS THE GALAXIES. *Richard E. Berendzen, Richard C. Hart and Daniel Seeley.* The history-making discoveries by the band of ten great astronomers who led the way between 1915 and 1940 are recaptured through original letters, photographs and archival treasures. Profusely illustrated. **$15.95**

70157. PRINCIPLES OF COSMOLOGY AND GRAVITATION. *Michael Berry.* A fascinating description of the universe as revealed by observation, and the current theoretical framework for its construction. **$16.95**

46510. EVOLUTION AND THE DIVERSITY OF LIFE. *By Ernst Mayr.* A most refined work probing a lifelong study of: evolution, speciation, the history of biology, theory of systematics, man, biogeography, behavior. **$20.00**

39865. COMPUTER DATA-BASE ORGANIZATION. *James Martin.* How you can use the latest design and implementation techniques to improve your data system's real-time accessibility and improve flexibility—at the lowest possible cost. *Counts as 2 of your 3 books.* **$26.50**

00470. VAN NOSTRAND'S SCIENTIFIC ENCYCLOPEDIA. Fifth Edition. Hailed as the most comprehensive one-volume reference work of its kind, this edition has been thoroughly revised and expanded. Illustrated with 2,450 photographs, diagrams, graphs, and over 550 tables. *Counts as 3 of your 3 books.* **$67.50**

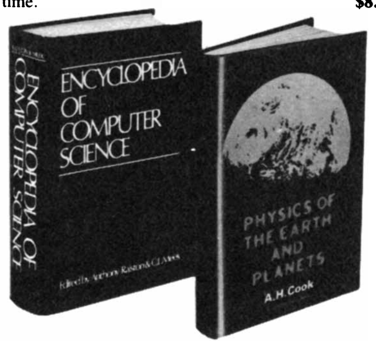
50576. GRAPHS, MODELS AND FINITE MATHEMATICS. *Joseph Malkevitch and Walter Meyer.* A problem-puzzle-solving work that is a delightful introduction to practical finite mathematics. Leads through graph theory, computers, functional models, probability, games and decisions—with examples, tables, illustrations and solutions to selected exercises. **$12.50**

33540. AN ANCIENT WORLD PRESERVED: Relics and Records of Prehistory in the Andes. *Revised and Updated by Frederic André Engel, Translated by Rachel Kendall Gordon.* Pieces together the complex relationship between men and the Andes and presents us with a vital picture of man at his best, mastering a hostile environment and erecting monuments that testify to his victory. **$12.95**

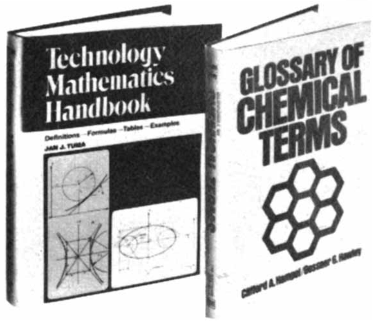
61640. THE MEANING OF FOSSILS. Episodes in the History of Paleontology. Second Edition. *Martin J. S. Rudwick.* Fine, intelligent examination of the cornerstones of paleontology, covering the men, the times and the ideas. **$15.00**

49267. FRONTIERS OF ASTROPHYSICS. *Edited by Eugene Avrett.* A contemporary source on astrophysical topics from recent solar research, to neutron stars, supernovae and intergalactic matter. **$20.00**

77500. THE SEARCH FOR THE GOLD OF TUTANKHAMEN. *Arnold C. Brackman.* The first complete story of the search, the discovery and the aftermath of one of the great archaeological finds of all time. **$8.95**



MEMBERSHIP BENEFITS • In addition to getting three books for $3.95 when you join, you keep saving substantially on the books you buy. • If you continue your membership past the trial period, you will be eligible for our Bonus Book Plan—an important way to save even more, at least 70% off publishers' prices. • Your Book Club **News**, describing the coming Main Selection and Alternate Selections, will be sent to you 15 times a year at three to four week intervals. • If you wish to purchase the Main Selection, do nothing and it will be sent to you automatically. • If you prefer one of the Alternates or no book at all, simply indicate your decision on the reply form always enclosed with the **News** and mail it so that we receive it by the date specified. • The **News** is mailed in time to allow you at least 10 days to decide if you want the coming Main Selection. If, because of late mail delivery of the **News**, you should ever receive a Main Selection without having had the 10 day consideration period, that Selection may be returned at Club expense.



82660. TECHNOLOGY MATHEMATICS HANDBOOK. *Jan J. Tuma.* Instant access to concise summaries of all major definitions, formulas, graphs, tables of elementary and intermediate mathematics—from algebra to determinants and vectors and much, much more. **$15.95**

50380. GLOSSARY OF CHEMICAL TERMS. *Clifford A. Hampel and Gessner G. Hawley.* Over 2,000 concise definitions: all major chemical groups, important functional terms, basic phenomena and processes, all the chemical elements, and more. **$14.95**

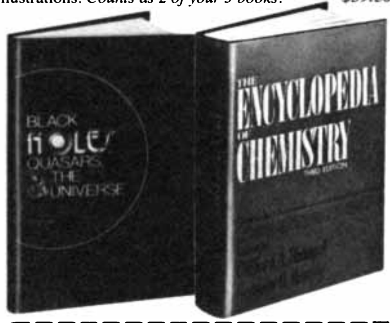
68735. PHYSICS OF THE EARTH AND PLANETS. *Alan H. Cook.* Brilliant investigation by one of the world's leading authorities on physical meteorology. *Counts as 2 of your 3 books.* **$24.75**

52320. HANDBOOK OF PHYSICAL CALCULATIONS. *Jan J. Tuma.* In one volume, a precise summary of the major definitions, formulas, tables and examples of technology physics. Summary of the major tools of physical calculation, with emphasis on practical applications. **$14.95**

36395. BLACK HOLES, QUASARS AND THE UNIVERSE. *Henry L. Shipman.* Also white dwarfs, neutron stars, supernovae, Seyfert Galaxies—and more. **$12.95**

44900. ENCYCLOPEDIA OF COMPUTER SCIENCE. *Edited by Anthony Ralston and Chester L. Meek.* Contains 470 articles, 1550 pages of the latest accurate information on hardware, software, systems, theory, economic, social, and professional issues, by 208 leading experts. *Counts as 3 of your 3 books.* **$60.00**

44861. THE ENCYCLOPEDIA OF CHEMISTRY. Third Edition. *Edited by Clifford A. Hampel and Gessner G. Hawley.* 600 leading chemists thoroughly explore over 800 theoretical and applied aspects of chemistry and chemical technology. Over 1200 pages, hundreds of invaluable charts, diagrams, and illustrations. *Counts as 2 of your 3 books.* **$39.50**



The Library of Science 2-5AK
Riverside, New Jersey 08075

Please accept my application for membership and send me the three volumes indicated, billing me only $3.95 for all three. I agree to purchase at least three additional Selections or Alternates during the first 12 months I am a member, under the membership plan described in this ad. Savings range up to 30% and occasionally even more. My membership is cancelable any time after I buy these three books. A shipping and handling charge is added to all shipments.

3 books for $3.95

Indicate by number the 3 books you want.

[ ] [ ] [ ]

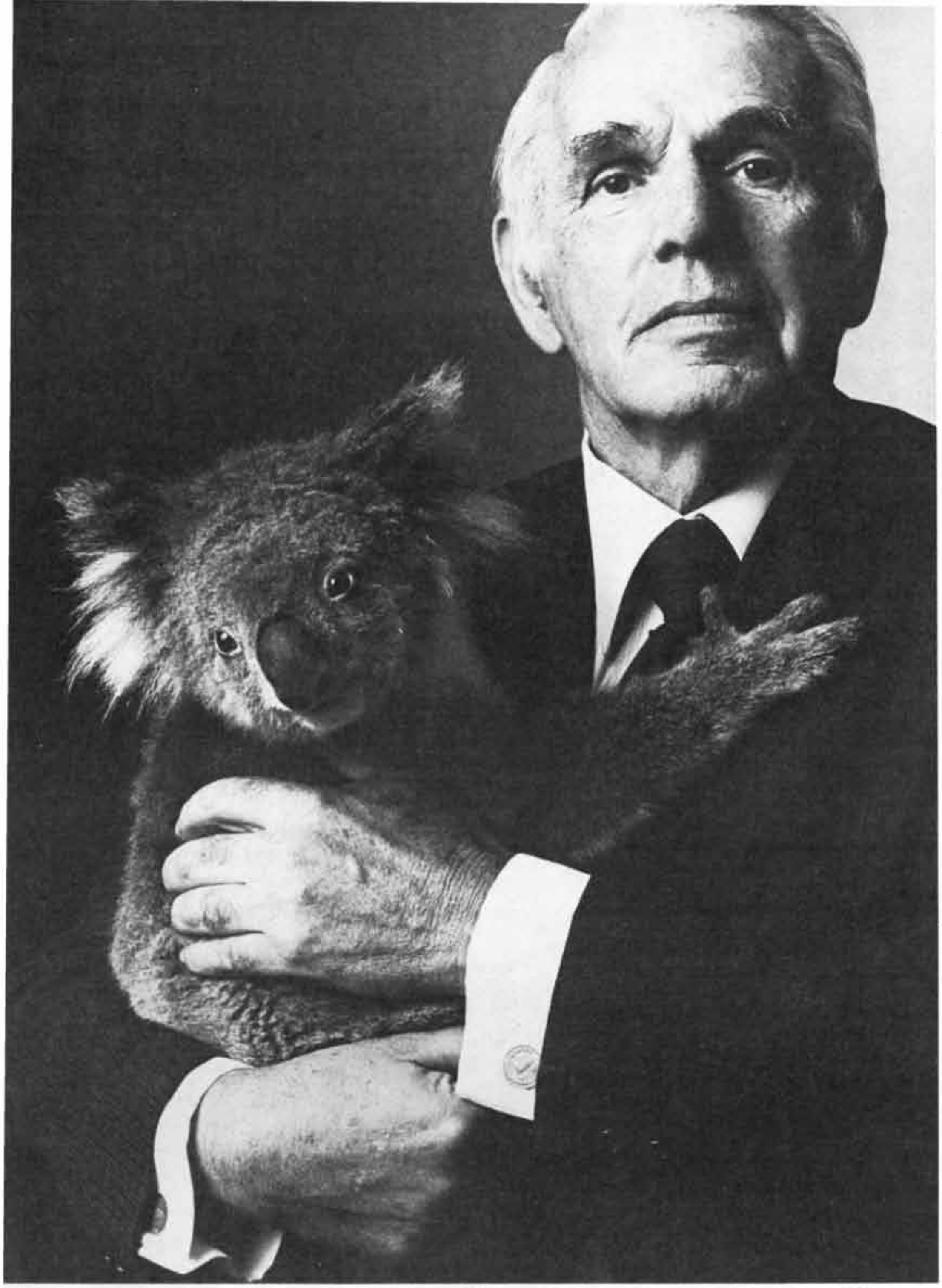
A few expensive books (noted in book descriptions) count as 2 or 3 choices.

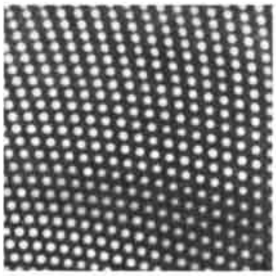
Name ______________________

Address ______________________

City ____________ State ______ Zip ______

(Offer good in Continental U.S. and Canada only. Prices slightly higher in Canada.) Book selections purchased for professional purposes may be a tax-deductible expense.





THE COVER

The photograph on the cover is an extreme close-up of the individual elements, or ommatidia, in the compound eye of an insect (see "The Compound Eye of Insects," by G. Adrian Horridge, page 108). The insect is a female horsefly, *Tabanus lineola*. The colors across the field are not the colors of the ommatidia themselves but are due to constructive and destructive optical interference.

THE ILLUSTRATIONS

Cover photograph by Fritz Goro

Page	Source
23	George V. Kelvin
24	Jack D. Griffith, Stanford University School of Medicine, and Andrew Staehelin, University of Colorado (*left*); Jack D. Griffith, Stanford University School of Medicine (*right*)
25	© David Scharf (*top*); Jack D. Griffith, Stanford University School of Medicine (*bottom*)
26	Maria Schnoss, Stanford University School of Medicine
27	Jack D. Griffith, Stanford University School of Medicine
28–33	George V. Kelvin
35–36	Jet Propulsion Laboratory, California Institute of Technology
37–39	Jerome Kuhl
40	Jet Propulsion Laboratory, California Institute of Technology
41–42	Jerome Kuhl
43	Jet Propulsion Laboratory, California Institute of Technology
45	Guy A. Cabral, Baylor College of Medicine (*top and bottom left*); Clifton R. Gravelle, Center for Disease Control (*bottom right*)
46	Gordon R. Dreesman, Baylor College of Medicine
47–49	Ilil Arbel
50	Ferenc Gyorkey, Veterans Administration Hospital, Houston
51–62	Ilil Arbel
66	Vernon French (*top*), Susan V. Bryant (*bottom*)
67–75	Bunji Tagawa
76	Peter J. Bryant
81	Bunji Tagawa
83	D. D. Double and A. Hellawell, University of Oxford
84	D. D. Double and A. Hellawell, University of Oxford (*top*); Ilil Arbel (*bottom*)
85	D. D. Double and A. Hellawell, University of Oxford
86	Ilil Arbel
88–90	D. D. Double and A. Hellawell, University of Oxford
93	Ralph Morse
94–105	Alan D. Iselin
108–109	Tom Prentiss
110	G. Adrian Horridge, University of Cambridge
111–113	Tom Prentiss
114	Type Systems Inc.
115–117	Tom Prentiss
118–120	Type Systems Inc.
123	Gabor Kiss
124	Burndy Library
125–128	Gabor Kiss
129	Dan Todd (*top*), Gabor Kiss (*bottom*)
130	Gabor Kiss
133	Ilil Arbel
134	Henry Regnery Company
135–137	Ilil Arbel
138	Fritz Goro
139–144	Michael Goodman
148	*The Magic Mirror of M. C. Escher.* © 1976, Random House

LETTERS

Sirs:

The interesting article "Amorphous-Semiconductor Devices," by David Adler [SCIENTIFIC AMERICAN, May], refers to work in this field that led to successful devices, including work on the Xerox Corporation's electrophotography photoconductors, as getting under way about 1958.

Several years prior to that date, the commonest television-camera tube, the Vidicon, had already entered commercial use, based on a retinal sensing layer purposely designed of amorphous photoconducting antimony trisulfide. This amorphous Sb_2S_3 device is still in production a quarter century later.

STANLEY V. FORGUE

Ocala, Fla.

Sirs:

The review of two recent books on slow viruses [SCIENTIFIC AMERICAN, May] was generally enlightening. One point made in the review, however, is rather debatable. The reviewer wrote: "Among the Fore people of New Guinea the women and children of certain groups suffer from kuru, a similar disease that is associated with the extraordinary and rather newly adopted ritual

Scientific American, July, 1977; Vol. 237, No. 1. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the author's risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 35 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name ____________________

New Address ____________________

____

____

Old Address ____________________

____

____

of rubbing on the mourning female kin the brains of a man who has died." Work by D. Carleton Gajdusek and others suggests that kuru is associated with a virus transmitted by cannibalism.

JOHN A. AGNEW

Syracuse University
Syracuse, N.Y.

Sirs:

In H. Moysés Nussenzveig's article "The Theory of the Rainbow" [SCIENTIFIC AMERICAN, April] he states that "the size dependence of the supernumeraries explains why they are easier to see near the top of the bow: raindrops tend to grow larger as they fall." If this were correct, then supernumeraries would gradually get closer to the main bow toward the base of the bow. However, the first supernumeraries maintain a roughly constant spacing before fading near the base of the bow. What, then, does account for the uneven distribution of supernumeraries around the bow?

The explanation lies in the aerodynamic distortion from the spherical of large falling drops. Drops with radii larger than about .25 to .5 millimeter become increasingly flattened on the bottom. These large drops will therefore not contribute to the horizontal (top) portions of the bow because a vertical section through the drop is not circular. They will contribute to the vertical (lower) portions of the bow because a horizontal section through the drop remains circular.

ALISTAIR B. FRASER

Pennsylvania State University
University Park, Pa.

Sirs:

Concerning "Phobos and Deimos," by Joseph Veverka [SCIENTIFIC AMERICAN, February], Jonathan Swift did more than merely echo Kepler's mathematical guess about the existence of two Martian satellites: he also stated that their orbital periods were 10 hours and 21.5 hours and that the satellites were three and five Martian diameters distant from the center of Mars.

Could Swift have had some source other than Kepler and his own vivid imagination? The odds against someone prior to Asaph Hall having a good enough telescope to observe the two moons are formidable but perhaps acceptable when compared with the odds against Swift guessing fairly accurately both the rotational period and the distance from Mars of the two satellites.

BENJAMIN R. MOSER

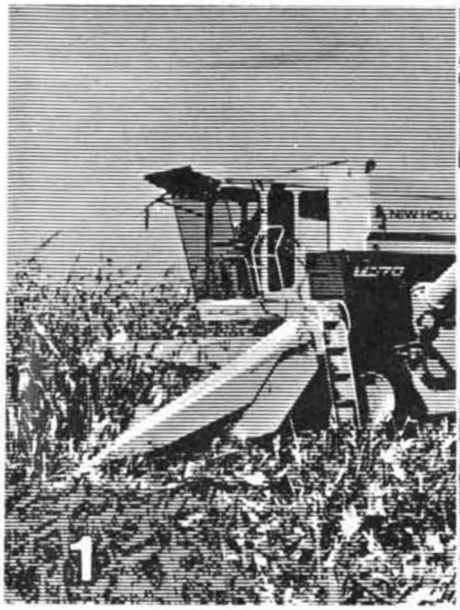
Stanley, Va.

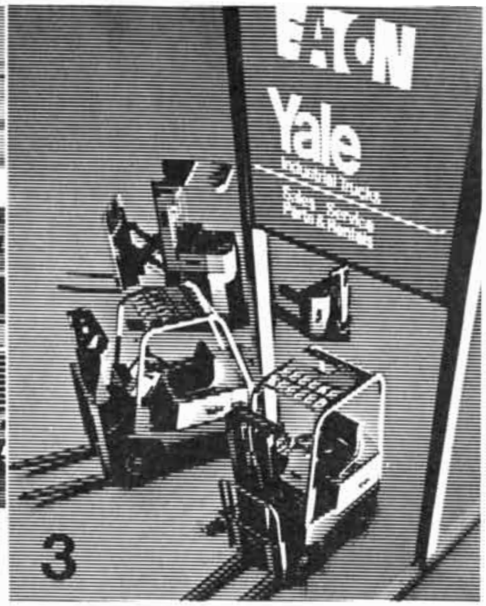
50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1927: "Whether the Bellanca monoplane will actually succeed in making a non-stop flight from New York to Paris or whether the French pilots Nungesser and Coli, making their attempt at the time these lines are being written, will be the first to achieve the crossing is uncertain. What is so admirable in the preparations of Bellanca and his associates is the careful way in which everything mechanical is being put into first-class order: the exhaustive flight tests to check fuel consumption at various revolutions of the engines and speeds of the plane; the marvelous endurance flight of 51 hours 11 minutes, the best possible test of the ability of the plane and pilots to maintain an unbroken voyage of nearly two days' duration across the ocean, and above all the care in watching weights. It was perhaps generosity in allowing weights to pile up that wrecked the hopes of both Sikorsky and Commander Davis. Clarence D. Chamberlain and Bert Acosta, companions in the splendid endurance flight, had become fast friends. When it came to selecting the pilot for the trans-Atlantic trip, Acosta relinquished his claim to the position because his sturdier build meant 60 pounds more weight than that of the slim Chamberlain."

"At the International Conference on Bituminous Coal, Dr. Friedrich Bergius of Heidelberg gave details of the manufacture on a commercial scale of light and heavy fuel oils, lubricating oil, benzene and phenol compounds and ammonia from waste coal dust or low-grade coal. That the process has passed beyond the experimental stage and is thought likely to become an important factor in the world-wide struggle for new sources of motor fuel is proved by the fact that it has been taken up by strong organizations in Germany, England and other countries. The essential principle of the process consists in combining hydrogen gas with coal by means of high heat and pressure. The coal is first ground into small pieces less than a tenth of an inch in diameter and then mixed with heavy oil to a thick pasty mass. This is placed in a light steel retort and heated to about 800 degrees Fahrenheit under a pressure of about 3,000 pounds per square inch. Most of the carbon unites with the hydrogen, giving a complex mixture of gaseous, liquid and solid compounds similar to those coming from natural wells. A difficulty of





Eaton Update:

1 Growing hydraulic power

Rugged hydraulic power by Eaton is improving the reliability and handling ease of many new agricultural machines. Our hydraulic motors and hydrostatic transmissions deliver power more smoothly, and stand up to dust, water and corrosive fertilizers better than any other form of power transmission.

Our Char-Lynn® hydraulic power steering, for instance, freed the designers of a new generation of combines from the restrictions of mechanical linkage, while giving the driver fingertip control of a seven-ton machine.

Eaton has been a leader in hydraulic power for thirty years. We saw the need, helped the agricultural revolution happen, and are growing with it.

2 Payoff at the truck stop

Meeting the urgent demand for improved fuel economy, Eaton's newly developed drive-train packages team our specially geared Fuller® transmissions with new low-ratio Eaton® axles. The result: definite fuel savings up to 10% in many cases.

These up-to-date economy combinations illustrate our commitment to maintain Eaton's important position in the truck market, which dates back 66 years. A commitment underscored by the construction of five new manufacturing facilities since 1970 and major capital expansion at seven others to serve our axle, brake and transmission customers.

3 Cutting the cost of moving things around

Eaton's response to the growing need for improved efficiency in materials handling has been the introduction of 23 new Yale® lift truck models in the last two years. This has made our line of industrial trucks the most up-to-date, as well as the broadest in the industry.

For more than 50 years Eaton has been the leader in what is now the fastest growing segment of the market—electric lift trucks. We stay ahead by anticipating the changing needs of various industries, where materials handling can account for as much as 40% of manufacturing or processing costs.

Eaton is a family of technologically related businesses with a balanced combination of manufacturing and engineering skills. We're always looking for new ways to use these skills in markets where needs are growing. This approach to the management of change has been achieving record sales and earnings. For the complete story, write to: Eaton Corporation, 100 Erieview Plaza, Cleveland, Ohio 44114.

EATON

the process, formerly regarded as being insuperable, is the high cost of hydrogen. But Bergius gets a sufficient quantity of hydrogen out of the gaseous products of the reaction itself."

"President Coolidge has made good use of radio and has saved much time and effort by addressing the people through the microphone instead of taking long and tiresome train trips in order to speak to them. The Coolidge inauguration on March 4, 1925, will go down in history as the first ceremony of its kind to be broadcast. On that occasion 27 stations from coast to coast were connected to the battery of microphones in front of the Capitol. This record tie-up of transmitters was surpassed on February 22 of this year, when the President addressed a joint session of Congress assembled to pay tribute to George Washington, through a network of 42 broadcasters scattered across the nation from Portland, Me., to San Francisco, Calif., reaching an audience estimated to be 20,000,000."

SCIENTIFIC AMERICAN

JULY, 1877: "Sir Thomas Watson has published a paper on zymotic diseases, in which he contends that the development of the entire group, including small pox, chicken pox, typhus fever, typhoid or enteric fever, scarlet fever, the plague, measles, whooping cough and mumps, is due solely to contagion. He would adopt, therefore, for the abolition of these diseases a process analogous to that which proved so successful in staying the cattle plague of 1865 in Great Britain. He would have the state exercise such powers as will ensure, first, the immediate isolation of a person affected; second, the thorough disinfection of his body, clothes, furniture and place of isolation, and, third, vigilant and effectual measures to prevent the importation of disease from abroad."

"The world cannot get along without rubber, which has now become one of the most necessary materials in a variety of trades. It has been the improvident practice to cut down trees 150 or 200 feet high to secure one hundredweight of rubber, and thus the forests of rubber trees, particularly in Brazil, are being destroyed and will ultimately belong to the past. Without waiting for such an event the British government has been shipping Brazilian rubber plants to Ceylon, Singapore, Burmah and other places. The plants thus distributed will do a great deal of good in preventing the otherwise impending calamity of a scarcity of rubber."

"That the earth was at one time incapable of sustaining life, and that at some


We help process one-fourth of all the oil and gas the world uses every day.
And that's being accomplished by just one member of Combustion Engineering: C-E Natco.
C-E Natco helps de-salt, de-foam, de-water and do other processing needed to prepare oil and gas for the pipeline.
And, besides processing it, we also help drill for it, extract it, pump it, clean it, deliver it and refine it.
Gray Tool Company, a new member of the C-E family, pioneered the first wellhead assembly able to operate at pressures of 30,000 pounds per square inch—extending the search for oil and gas into earth layers too difficult to control before.
C-E Crest has designed systems that produce 10-million barrels of oil a day.
C-E Lummus engineered over 30% of Europe's refining capacity. And is responsible for the design and construction of deck facilities for off-shore platforms in the North Sea that can withstand 130-mph winds and 100-foot-high waves.
We do more than help process oil and gas.
For over 60 years, we've been supplying equipment, technology and research to help produce energy, conserve it, squeeze more use out of it, even convert it to new forms.
For more information about our activities around the world, write Combustion Engineering, Inc., Dept. 7008-33, 902 Long Ridge Road, Stamford, Connecticut 06902.
COMBUSTION ENGINEERING
The Energy Systems Company

time in the course of events life began to be, no one doubts for a moment. It is also pretty generally admitted among scientific men that the beginning of life was in all probability a natural event; and that the earlier forms of life did not embrace the more complex types now existing, but were of simpler structure, perhaps not unlike the lowly organisms now studied under the microscope. Here the question arises: Was the beginning of life a phenomenon single and unique, and are the bacteria of today the unaltered descendants of the earliest forms of life? Or may life have begun, and may it still begin, at any time by the concurrence of suitable conditions?"

"Aerial telegraphy without wires or poles is spoken of as among the possibilities of the future. Professor Loomis claims to have communicated with an assistant 12 miles away by means of an aerial current. The current was reached by flying kites to a certain height at each point, the string used being a copper wire. When both kites had attained the same altitude, messages were sent and received by an instrument at the ground end of the wire, the only connection being the atmospheric current between the kites."

"Bones, though at first the main source, now furnish but little of the phosphoric acid of fertilizers, so great has been the demand for phosphatic manures. To supply the deficiency, the vast beds of mineral phosphates, so called, prove to be deposits of great aggregate value. The chief of them are located in Central Russia, Western Germany, Southern France, Canada, South Carolina, and in the Navassa, Sombrero, Jarvis and Baker Islands. The Navassa and South Carolina deposits are now the main sources of supply."

"The largest meeting of bicycle riders that has ever been held assembled on May 25th at Hampton Court in England. The drivers extended for a mile and a quarter on the Kingston Road, and numbered between 1,500 to 2,000. It is about 10 years since the first wooden bicycle was made in England, and only seven years since the iron machines with rubber tires, spider wheels and an improved felly were first made. In a country with good roads such as our own, the bicycle-rider traveling very swiftly and with perfect silence, might be very useful in war as a scout and dispatch bearer. Postmen have already used the machine, and it might also be useful to the police. It is the silence of the bicycle that constitutes its chief danger to pedestrians and horsemen in time of peace, and the expediency of obliging each bicycle to bear a bell, like the small bells that tinkle in the harness of horses in many foreign countries, has been seriously considered."

uh...2nd.

If you're in the market for a small imported car, you're probably looking at cars like the VW Rabbit, the Toyota Corolla, the Datsun B-210 and the Subaru DL.

But how about the Buick Opel?

That's what we thought. And that's the trouble. We know the Opel is a little dynamo of a car. But apparently very few of you even know it exists. Much less what it can do.

Thus, The Buick Opel 5-Car Showdown. In which we pitted our tough, talented Opel against four of its better known competitors in daring, fender-to-fender comparisons. Comparisons that would answer a lot of the questions that are probably on your mind. Like which car accelerates quickest? Which cars get the best EPA mileage ratings? Which corners flattest? Which is quietest and roomiest inside? Stuff like that.

To make sure the results would be fair and unbiased, we enlisted the aid of an independent engineering firm and an editor from *Car and Driver* magazine to oversee the procedures and review the data. All test cars were equipped as much alike as possible: all had 4-speed manual transmissions, available air conditioning, radial tires and a healthy complement of options. And the tests were conducted at the General Motors Desert Proving Grounds in Mesa, Arizona.

You may have seen our ads in your local newspaper, where we've been reporting the Showdown results on a regular basis. But in case you haven't or have missed some, here's what we've been up to.

SHOWDOWN #1: HUMAN ENGINEERING

Rabbit jumps ahead, takes first place.

	Interior Noise	Interior Room *	Total Showdown Points
	(Showdown Points)		
1. VW Rabbit (1600cc) Hatchback 2-dr. Sedan	5	5	**10**
2. Toyota Corolla (1600cc) Deluxe 2-dr. Sedan	5	4	**9**
3. Buick Opel (1800cc) Deluxe 2-dr. Coupe	5	3	**8**
4. Subaru DL (1600cc) Deluxe 2-dr. Sedan	5	2	**7**
5. Datsun B-210 (1350cc) Deluxe 2-dr. Sedan	5	1	**6**

*Using EPA procedures we calculated front and rear seat volume estimates. (Head x shoulder x leg room). A total passenger volume estimate resulted.

Here we focused on an area of traditional weakness in small, imported cars: creature comfort. In other words, how quiet, how roomy, and how generally good our test cars are at making life pleasant for their occupants.

The Interior Noise Test turned out to be too close to call, so we awarded each car five first-place points. But Rabbit's good showing in the Interior Room contest gave it first overall in the initial showdown. Opel, as you can see, finished third here.

So while we weren't exactly thrilled with Opel's third place finish, we did manage to beat two of our better-known competitors.

SHOWDOWN #2: UTILITY

Opel bounces back to grab victory.

	Trunk Capacity *	Pulling Power **	Parking Lot Maneuverability ***	Total Showdown Points
	(Showdown Points)			
1. Buick Opel	3	5	5	**13**
2. VW Rabbit	5	3	3	**11**
3. Toyota Corolla	2	4	4	**10**
4. Subaru DL	4	2	2	**8**
5. Datsun B-210	1	1	1	**3**

*Results based upon mfr. est's provided by the EPA.
**Measured pulling capabilities for (1) 30 mph (3rd gear) and (2) 55 mph (4th gear)
***Results based upon: (1) Measuring steering effort with the car standing still. (2) Measuring steering effort with the car moving at 10 mph. (3) Mfr. published turning diameters (Subaru measured)

The function of this Showdown was to test functionality: which car has the biggest trunk, which car is easiest to park, which has the best pulling power. The results were most gratifying.

The Rabbit, with its hatchback, easily took trunk capacity honors. But our intrepid little Opel outmaneuvered all the others in parking ease (where we measured two kinds of steering effort and calculated turning diameter) and received the highest grades in what the engineers call "gradability" (or what you'd probably call "pulling power").

And that pulled Opel into a first-place tie with Rabbit after two showdowns. Most gratifying, indeed. Especially for a little car you might not even have been considering.

SHOWDOWN #3: OPERATING EFFICIENCY

Toyota and Datsun tie for first; Opel—Gulp!—Last.

This was not our favorite event. Opel, with the biggest engine (1800cc) of all five cars, showed 27 miles per gallon* (based on EPA combined estimates: 55% city/45% highway). But it wasn't good enough.

Opel's bigger gas tank helped it do well when we measured Range (or how far a car can go on a full tank of gas), but not as well as Subaru and Toyota, who tied for first. And when we compared recommended maintenance schedules for 37,500 miles of normal driving, Opel (8 recommended visits) finished behind Rabbit (6 visits), and Datsun and Toyota (7 each).

All of which adds up to a first-place tie between Toyota and Datsun, with Opel—ahem—bringing up the rear, and dropping to third overall after three showdowns. As we said, this was not our favorite event.

	Combined Mileage Est. *	Combined Range Est. **	Recommended Maintenance Stops ***	Total Showdown Points
	(Showdown Points)			
1. Toyota Corolla (EPA est. 39 hwy/28 city/32 comb.) (13.2 gal. tank)	4	5	4	13
1. Datsun B-210 (EPA est. 41 hwy/29 city/34 comb.) (11.6 gal. tank)	5	4	4	13
2. Subaru DL (EPA est. 41 hwy/28 city/32 comb.) (13.2 gal. tank)	4	5	2	11
3. VW Rabbit (EPA est. 37 hwy/24 city/28 comb.) (11.0 gal. tank)	3	2	5	10
4. Buick Opel (EPA est. 36 hwy/23 city/27 comb.) (13.7 gal. tank)	2	3	3	8

*Based upon EPA combined estimates (55% city/45% hwy). All vehicles equipped with standard 4-speed manual transmissions. Actual mileage may vary with driving habits, the condition of the car and its equipment.
**Based upon multiplying EPA combined mileage est. by gas tank capacity as published by each manufacturer. Actual range may vary.
***Based upon the number of recommended maintenance stops as outlined by the mfr. The number and type of recommended inspections, adjustments and replacements would vary by visit. The less frequent the number of dealer visits, the higher the score.

SHOWDOWN #4: ACCELERATION, PULLING POWER, CORNERING FLATNESS, STEERING QUICKNESS

The fourth event was close; Rabbit and Opel outrun the pack.

This, however, could very well be our favorite event. Because the first thing most people want to know about any car is, "What'll she do?"

How fast does it accelerate? How's it corner? How sensitive is the steering? And—for Opel—the answer to all of these questions is, "Quite well, thank you."

So well, in fact, that our surprising little Opel was barely nosed out by a car much ballyhooed for its prowess in these areas, the VW Rabbit.

Our Acceleration Test clocked three capabilities (0-55 mph, 20 to 35 mph., and 30 to 70 mph.), and Rabbit beat the clock every time. But Opel—yes, Opel—pulled down a solid second place.

In the Gradability test, (the ability to climb a hill) Opel finished second. But Opel showed the way around turns, exhibiting (along with Datsun) the least body lean in the Cornering Flatness Test (assuming, of course, that flatter is preferred—if not, switch the order in the chart); and tied with Toyota, Rabbit and Datsun for first place when we measured Steering Sensitivity or quickness.

	Acceleration *	Pulling Power **	Cornering Flatness ***	Steering Quickness ****	Total Showdown Points
	(Showdown Points)				
1. VW Rabbit (1600cc) (1945# Avg. Curb Wt.)	5	5	4	5	19
2. Buick Opel (1800cc) (2193# Avg. Curb Wt.)	4	4	5	5	18
3. Toyota Corolla (1600cc) (2190# Avg. Curb Wt.)	3	3	3	5	14
4. Subaru DL (1600cc) (2031# Avg. Curb Wt.)	2	3	4	4	13
4. Datsun B-210 (1350cc) (2047# Avg. Curb Wt.)	1	2	5	5	13

*Based upon: (1) 0-55 mph (2) 20-35 mph (3) 30-70 mph
**Based upon pulling capabilities for: (1) 30 mph (3rd gear) and (2) 55 mph (4th gear)
***Based upon opinion that less lean is preferable to more lean
****Based upon: (1) measuring lateral acceleration and (2) calculating the resultant circle diameter (per 100 degree steering wheel angle @ 30 mph)

AND THE WINNER IS...

Let's put it this way: we didn't win. Well, alright, if you insist, we'll put it this way: Rabbit won.

But not by much. As you can see in the final point totals, the VW only beat our Opel by three points (with Calif. emission equipped cars: four points). And Opel did finish ahead of three highly regarded opponents in the areas we tested.

So if the point of this whole 5-Car Showdown was to convince you that our Opel could hold its own against better-established competitors . . . that Opel should definitely be considered when you wander out to shop . . . well, we think we've made it.

Results will vary in California.

THE BUICK OPEL 5-CAR SHOWDOWN.

THE AUTHORS

CLIFFORD GROBSTEIN ("The Recombinant-DNA Debate") is professor of biology and vice-chancellor of university relations at the University of California at San Diego. He received his Ph.D. in zoology from the University of California at Los Angeles and then worked as an instructor of zoology at Oregon State College. In 1946, after spending three years as an aviation physiologist in the Air Force, he joined the National Cancer Institute, where he worked as a research biologist for 10 years. For nine years thereafter he was professor of biology at Stanford University. He joined the faculty at San Diego in 1965, and from 1967 through 1973 he was dean of the School of Medicine and vice-chancellor of health sciences.

CONWAY B. LEOVY ("The Atmosphere of Mars") is professor of atmospheric sciences and geophysics and adjunct professor of astronomy at the University of Washington. Intrigued by the weather since boyhood, he first did his undergraduate work in physics at the University of Southern California and then joined the Air Force as an Air Weather Service forecaster and weather reconnaissance observer in the North Pacific and Central Pacific. After leaving the Air Force he went to the Massachusetts Institute of Technology, where he received his Ph.D. in meteorology in 1963. For the next five years he was a research meteorologist at the Rand Corporation, leaving in 1968 to go to the University of Washington. He has been interested in the atmospheres of the planets for more than a decade; in 1968 he began working with Yale Mintz of the University of California at Los Angeles on models of the dynamics of the atmosphere of Mars. Since then he has been involved in the interpretation of the data on Martian atmospheric phenomena gathered by *Mariner 6, Mariner 7* and *Mariner 9* and in the meteorology experiments aboard the Viking landers.

JOSEPH L. MELNICK, GORDON R. DREESMAN and F. BLAINE HOLLINGER ("Viral Hepatitis") are members of the faculty of the Baylor College of Medicine. Melnick is Distinguished Service Professor of Virology and Epidemiology. He first began doing research on viral hepatitis during World War II, after obtaining his Ph.D. from Yale University. He remained at the Yale School of Medicine for 18 years before moving to Baylor. This past year Melnick served as chairman of the World Health Organization's Committee on Viral Hepatitis. Dreesman is associate professor of virology. After obtaining his B.A. at Central College in Iowa and his M.A. at the University of Kansas, he earned the first Ph.D. given by the department of microbiology at the University of Hawaii at Manoa. His research interests have centered on viral immunology. Hollinger received his M.D. from the University of Kansas in 1962 and is a specialist in internal medicine. After working for two years at the Center for Disease Control of the U.S. Public Health Service in Atlanta, Ga., and then completing a two-year special postdoctoral fellowship in basic virology funded by the National Institutes of Health, he joined the faculty at Baylor in 1970.

PETER J. BRYANT, SUSAN V. BRYANT and VERNON FRENCH ("Biological Regeneration and Pattern Formation") are all zoologists. The Bryants, husband and wife, are associate professors in the department of developmental and cell biology at the University of California at Irvine. Both were born in England and studied zoology at King's College of the University of London. Peter Bryant received his Ph.D. in genetics from the University of Sussex in 1967; Susan Bryant received her Ph.D. in zoology at St. Mary's Hospital Medical School the same year. Both did postdoctoral work at the Developmental Biology Center of Case Western Reserve University before joining the faculty at Irvine. French is lecturer in the department of zoology at the University of Edinburgh. He received both his undergraduate and graduate education at the University of Sussex. After doing postdoctoral work at the University of Science and Medicine in Grenoble and spending six months in the department of zoology at the University of Leicester he worked at the National Institute for Medical Research at Mill Hill in London. He went to Edinburgh in January of this year.

D. D. DOUBLE and A. HELLAWELL ("The Solidification of Cement") are respectively research fellow in the department of metallurgy and science of materials at the University of Oxford and professor of materials at the University of Wisconsin at Milwaukee. Double received his undergraduate and graduate education at Oxford, where his Ph.D. dissertation was on the solidification of metals. Since then he has been studying the hydration of portland cement. Hellawell received his undergraduate degree in chemistry from Oxford in 1953 and his Ph.D. three years later. He then became a lecturer in the department of metallurgy, where he remained until this year.

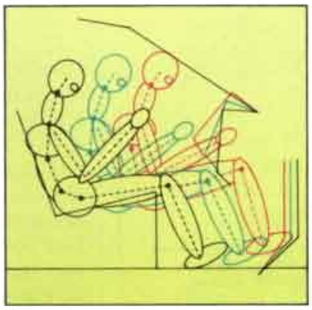
JAMES L. DYE ("Anions of the Alkali Metals") is professor of chemistry

We're expanding the body of knowledge about the human body.

A 73-year-old stuntman, Henri LaMothe, dives 35 feet into 13 inches of water. He does it regularly. Without harm. Now that says something about body toughness.

Once, in a less severe dive, our instruments went with him. And we found that his chest safely withstands "g" forces beyond those commonly thought to be injurious.

Learning about the impact response of the body is just one phase of our extensive biomedical research program at the General Motors Research Laboratories. The goal in this phase: Develop a better understanding of man's reaction and tolerance to collision forces.



To this end we have done mechanical response and trauma studies on the head, chest, and leg; on tissue; and on the body as a whole. We've also developed mathematical models of the man-car system for simulating crashes in a computer and predicting car occupant kinematics.

Another phase of our work concerns the health effects that may be related to GM products. It includes respiratory response to air pollutants, the dynamics of air flow in the lungs, and toxicity investigations.

The program in biomechanics started here some 15 years ago. It spawned the first energy-absorbing steering column put into production... a head-injury-severity index used by the United States Department of Transportation and football helmet designers ...a sophisticated anthropomorphic test dummy.

Our biomedical research continues, through meaningful data, to promote GM's most important commitment to society: a safer, practical car and a healthier environment.

General Motors Research Laboratories
Warren, Michigan 48090

at Michigan State University, where he has been on the faculty since 1953. After being graduated from Gustavus Adolphus College he received his Ph.D. in physical chemistry from Iowa State University in 1953. In 1961 he was a National Science Foundation Science Faculty Fellow doing research on fast chemical reactions in collaboration with Manfred Eigen at the Max Planck Institute for Biophysical Chemistry in Göttingen. He has studied solutions of alkali metals in ammonia and amines since 1954, currently with research support from the Energy Research and Development Administration (ERDA). In addition, over the past five years Dye has begun research on the optical spectra of the intermediate compounds formed during reactions catalyzed by enzymes.

G. ADRIAN HORRIDGE ("The Compound Eye of Insects") is a professor in the department of zoology at the University of Cambridge. Born in England, he received his undergraduate and graduate education at Cambridge, studying subjects ranging from jellyfish nerve networks to the coral reefs of the Red Sea to Persian carpets. His first job was at the Royal Aircraft Establishment at Farnborough, where he worked on the design of engineering structures made with composite materials. From there he went to St. Andrew's University, where he later became director of the Gatty Marine Laboratory. In 1969 Horridge went to the Australian National University in Canberra, where he founded the department of neurobiology. He returned to Cambridge in February of this year. His principal avocation is the study of Indonesian boatbuilding at the shipyards where these traditional sailing vessels are still being built. "This interest has taken me to remote parts of Java, Madura, Bali, Timor, the Moluccas and Celebes and to equally inaccessible libraries and collections," he writes.

IAN STEWART ("Gauss") is at the Mathematics Institute of the University of Warwick. After being graduated from the University of Cambridge in 1966, he received his Ph.D. in mathematics from Warwick. In 1974 he was a Humboldt Foundation Fellow at the University of Tübingen, and in 1976 he spent six months at the University of Auckland. Beginning this fall he will be at the University of Connecticut at Storrs. Stewart's main field of research is infinite-dimensional Lie algebras. He writes: "My interest in the history of mathematics stems from a desire to understand how the subject of mathematics fits together. Many of the cross-connections become clear only if the history is borne in mind." He has published eight books on mathematics (including two for children) and is now finishing a ninth.

Established 1845 SCIENTIFIC AMERICAN July 1977 Volume 237 Number 1

The Recombinant-DNA Debate

The four-year-old controversy over the potential biohazards presented by the gene-splicing method and the effectiveness of plans for their containment is viewed in a broader context

by Clifford Grobstein

The guidelines for research involving recombinant-DNA molecules issued a year ago by the National Institutes of Health were the culmination of an extraordinary effort at self-regulation on the part of the scientific community. Yet the policy debate over recombinant-DNA research was clearly not laid to rest by the appearance of the NIH guidelines. Instead the debate has escalated in recent months both in intensity and in the range of public involvement. A watershed of sorts was reached in March at a public forum held by the National Academy of Sciences in Washington. The forum was in part a repeat performance by scientists arguing fixed positions that were established early in the debate. There were, however, new participants on the scene, and they presented a varied and rapidly shifting agenda. They made it clear that research with recombinant DNA had become a political issue. As one speaker remarked, the Academy forum may have been the last major public discussion of recombinant DNA arranged by the scientists involved in the research. Nonscientists at the forum, by word and deed, reiterated the theme that science has become too consequential either to be left to the self-regulation of scientists or to be allowed to wear a veil of political chastity.

Science of course is crucially consequential to society, precisely because it is an intensifying source of both benefits and risks. Research with recombinant DNA may provide major new social benefits of uncertain magnitude: more effective and cheaper pharmaceutical products; better understanding of the causes of cancer; more abundant food crops; even new approaches to the energy problem. These and other possible outcomes are envisioned in "best-case scenarios" for the future application of recombinant-DNA technology. "Worst-case scenarios" can also be conceived: worldwide epidemics caused by newly created pathogens; the triggering of catastrophic ecological imbalances; new tools for militarists and terrorists; the power to dominate and control the human spirit.

Both the best-case and worst-case scenarios are largely speculative; the gap between them symbolizes the large degree of uncertainty that surrounds this major step forward in molecular genetics. The material basis of biological heredity has been broken into in the past two decades, and it seems as though each of the fragments has acquired a life of its own. In this resulting period of instability fear threatens to override wonder as the implications of the research diffuse more widely. The fear is not so much of any clear and present danger as it is of imagined future hazards. The classic response to such fears is rigid containment: the Great Wall, the Maginot Line, the cold war. All are manifestations of the effort to provide absolute security against unpredictable risks, and yet each generates its own risk. The escalation of the recombinant-DNA debate has a component of this kind of behavior, but there is a more rational component as well.

The first round of the fateful debate began in 1974, when investigators at the leading edge of work in this field declared a voluntary moratorium on several types of experiment judged to be conceivably risky. A set of techniques had been developed that made it possible to cut the long, threadlike molecules of DNA into pieces with the aid of certain enzymes, to recombine the resulting segments of DNA with the DNA of a suitable vector, or carrier, and to reinsert the recombinant into an appropriate host cell to propagate and possibly to function.

The significance of the new developments is rooted in the central biological role of DNA as the transmitter of genetic information between generations. The transmission of the encoded genetic message depends on the ability of a cell to generate exact replicas of the parental DNA and to allocate the replicas among the offspring. In addition the success of genetic transmission depends on the ability of the offspring to "express" the

GENETIC CODE of an extremely small bacterial virus, the bacteriophage designated ϕX174, is given by the sequence of letters on the opposite page. The letters stand for the four nucleotides cytosine, guanine, adenine and thymine, which are linked end to end to make up each strand of the normally double-strand DNA molecule. The genetic message embodied in each strand of DNA is represented by the particular sequence of nucleotides, any one of which may follow any other. In the ϕX174 virus the DNA molecule, which has only a single circular strand for part of its life cycle, consists of approximately 5,375 nucleotides; the nucleotides are grouped into nine known genes, which are responsible in turn for coding the amino acid sequences of nine different proteins. For example, the dark-color segment of the molecule, called gene *J*, codes for a small protein that is part of the virus; this segment also happens to be the shortest gene in the ϕX174 genome. The complete nucleotide sequence for the DNA in ϕX174 was worked out recently by Frederick Sanger and his colleagues at the British Medical Research Council Laboratory of Molecular Biology in Cambridge. About 2,000 pages of this type would be required to show the nucleotide sequence for the DNA in the chromosome of a typical single-cell bacterium; roughly a million pages would be needed to similarly display the genetic code embodied in DNA molecules that make up chromosomes of a mammalian cell.

GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAACGAGGAATTACTACTGCTTGTTTA
TCAACTACCGCTTTCCAGCGTTTCATTCTCGAAGAGCTCGACGCGTTCCTATCCAGCTTAAAAGAGTAAAAGGCGGTCGTCAGGTGAAGCTAAATTAAGC
AACGATTCTGTCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTAGATATGAGTCACATTTTGTT
TGAACTGAGTACTAAAGAATGGATAATCACCAACTTGTCGTAGCCTGAGTCTATCATTAGGTGCGAGAAAATTTTACAGTTGTTCTCTTAGAGATGGTAC
TACTGAACAATCCGTACGTTTCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTTCGATTTTCTG
TCGCTCCCATAGGATGTTTCAGGTCGCATGGTATTTGCGTTCGGAGTTGCGTCGCTGCTCGTGCTCTCGCCAGTCATCGTTAGGTTTGAAACAATGAGCA
TTCCTGCTCCTGTTGAGTTTATTGCTGCCGTCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
CAGTCATTCTTGCAGTCACAAAGGACGCGCATGTGCGTTCCATTTGCGCTTGTTAAGTCGCCGAAATTGGCCTGCGAGCTGCGGTAATTATTACAAAAGG
GCAGAAGAAAACGTGCGTCAAAAATTACGTGCGGAAGGAGTGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTT
TTAAATAGGAGTTCATTCCCCGGCTTCGGGGACGTTAATTTTAACAACTGGTGGATGTATGGTTTCTGCTCGCGGAAATGCGAACGGAAATCATGGAGCG
ATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCATCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTA
TACAGATGTCATCTCAGTTATCGTTCCGGTGCTGCGTTACCTCTTTCTGCCTCTCGCGGTTGCCGCAGGTAGAGCTTCCTCAGCGGTCGCTATTGGCCTC
TTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACCAA
GATATCTATAGTTTATTGGGACTTTGTTTACGAATCCCTAAAATAACCATAGTCCCAATTAGCACGGTTCTTTTCGCCGTACCAGTTATATTGGTCATCA
CGTATTTTAAAGCGCCGTGG—ATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGCTCGTTATGGTTTCCGTTGCTGCCATCT
TCGTCGAACGTCTGGGTATTACAGTTATCTACACCATCTTCAGCAGTAAACCGCTCTTTCGAGTCAGAGTCCTCCTTCGCCTCGTCAGGTTTACAAAAAC
TATGCTAATTTGCATACTGACCAAGAACGTGATTACTTCATGCAGCGTTACCATGA-GTTATTTCTTCATTTGGAGGTAAAACCTCATATGACGCTGACA
ACTTGTGCTGGTCTTTTGACCGGATTGCTGCAAACCAGTCAAGGTAGTTGTAGTATCGGTCTACGGGTCTCTAATCTCGCGTACTGTTCATTTCCTGCCA
ACAGACCTATAAACATTCTGTGCCGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTGGTTCGTTTTCCGCCTACTGCGACTAAAGAG
GGAAGTATCTTTAAAGTGCGCCGCCGTTCAACGGTATGTTTTGTCCCAGCGGTCGTTATAGCCATATTCAGTTTCGTGGAAATCGCAATTCCATGACTTA
ATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGTTTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCT
TTGACTTGCTGACTTTGTGACCAGTATTAGTACCAACGCTTATTCATGCGCAAGAACGTTTAGTGGTCTTCCGCCAAGGACTTACTTACCCTTCGGAAGT
GTTGCAGTGGATAGTCTTACCTCATGTGACGTTTATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTA
TTTCAGACTTTGTACTAATTTGAGGATTCGTCTTTTGGATGGCGCGAAGCGAACCAGTTGGGGAGTCGCCGTTTTTAATTTTAAAAATGGCGAAGCCAAT
TATTTCTCGCCACAATTCAAACTTTTTTTCTGATAAGCTGGTTCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACA
GGACTAATCGCCGCAACTGTCTACATAGGTAGACTTACGTTACTTCTTTTGGTGGTAATGGTCGTAATTGGCAGTTTGATAGTTTTATATTGCAACTGCT
TTGTTTCAGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGCCTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGAC
TAACGGCCCGCATGCCCCTTCCTGCAGTTATCAGTGTGTCAGGAACTGCCATATTATTGGTGGTAGTACCGCTGGTAGGTTTCCTATTTGTAGTATCCGT
AACGTCTACGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGTTTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCC
AACTGGCGGAGGTTTGTTAAATCTGTACCGCGGTGGTCGTTCTCGTCTTCGTTATGGCGGTCGTTATCGTGGTTTGTATTTAGTGGAGTGAATTCACCGA
AAAGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTA
CTTCACGGTCGGACGTTGCATGGAAGTTCTTCAGGAAATGGTCGAAATCGTTATCGTGTGCTTTGTTTTGATCCCCGCCGGAGTAGTCCCAATCCTTGTA
TGCCGTTTCTGATAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTATCTTGCTGCTGCATTTCCT
GGTCAACGTAAAATCATTCGAGAAAAACTAAGAGTTTAGGCCGCAGTTGGTATGGTCGTCTCCTTCGTAGTCGTGGTCGTGCGAGGGTTCGTAATTCGAG
ACAATCAGAAAGAGATTGCCGAGATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGACCAGGTATA
CCTTTGGACGACAACGAACCTTTCTAACCACAAAGGTATTATCTGCGTTGCGTCGTCATCTGAGGAAGACACTTATTCGTTCGTAGAGTAAAACACGT
GAGATTATGCGCCAAATGCTTACTCAAGCTCAAACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTG
GGTCTTCGTCGTAGTCACTGCTGTAATCTTTATAGGAAACGTCATCGCGGTTATACTCTTCTCGGTATGGCGACTAAGACGCAAACGACTACTTGATTCA
TGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGATACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCT
TAAGTCTTCCCATTATTCTTGCTTGGTATTTTTTCGGAGGTTCTAAACCTCCGTACTTTTGTATGTTAACCCTCCCACAGTTAGGACTGCCAATAAAGGA
GTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTATTGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTT
CGGCAGTTGTATGTATAGTGGTAATAGCTTGAGTTGCGGGACGTATGCTTTTCTGTCTTAGAGAAGGTTCTCGAACTACGCCAATAGGTAGACGAATACC
CATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGAATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATA
CCCGCAAGTCGTCGGTCGAACGTTTTGACGCATTGGCAGAAGAGCAAGAGATTTTTGGTAAAAAGCAGGGGAAGCCCCGCCACCAGATATCACAATAATT
TCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATTGCTGGAGGCCTCCACTAAGATATCGCGTAGA
GATTAGCCAGCAGTCGGTTGCACTCTCACAGTTTTTGCTATTTGGTTGGTAGTCGTACTCGGACAGCGTAACGTAAGTAGTTTGCGACTTATCGTTTCGG
AGGCGTTTTATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCGCAAGGCTAATGATTCACACGC
CGATGGACATCCTTCACAGGCGTATTTCACGTGGCGTACCTTTACTTCTGCCGGTAATCGACATGGTATGAGTCCGTGTGTTTTTATGACTATCGTCAGC
GTTGACCCTAATTTTGGTCGTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCATCGCAGTTCGCT
TGTATCTTTGGTTGTCGGTATATTGACCATCGAAATTCGCCGAGTGGAAATCGTAGTTGTCCGGTGTTGGTTGGTCTTGCACTTTTTTCGCAGGACGCACA
GGCTAAATACGTTAACAAAAAGTCAGATATGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACTT
CGAACCATTCAACCTAATTCGTGAGGCACCTGTCTAAACAGTAACACTCGTAAAAGTAGGGCTTCAACGCCGAGTAAGACTAAGACTTGTCGAAGAACCC
TGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGCAGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGC
ACGTCCAACCTATGCGGTTAGTAAAAATAGCTTCGCGCGTATTTAAACTCGTCTAAACAGCAGTGTCCAACGCGG

encoded information properly by referring to it to control essential life processes. The mechanism of genetic expression in higher organisms is at present only dimly understood, and the discovery of the new recombinant-DNA techniques seemed immediately to open a broad new avenue to increased knowledge in this field.

The detailed mechanisms of genetic replication and expression are enormously complex. The essence of the matter, however, is found in the famous "double helix" structure of DNA. Both of the two long, interwound and complementary strands of the DNA molecule are made up of four kinds of nucleotides, cytosine, guanine, adenine and thymine (abbreviated *C, G, A* and *T*), which are linked end to end like a train of boxcars. The genetic message of each strand is embodied in the particular sequence of nucleotides, any one of which may follow any other. For example, the sequence *CATTACTAG* contains five identifiable English words: *CAT, AT, TACT, ACT* and *TAG.* The genetic message, however, is "written" in triplets: *CAT, TAC* and *TAG.* In general each triplet "codon" determines, through a series of intermediate steps, the position of a specific amino acid in a protein molecule.

Proteins, like nucleic acids, can be visualized as long trains of boxcars coupled end to end; here, however, the subunits are amino acids rather than nucleotides. The sequence of nucleotides in a given DNA molecule determines the sequence of amino acids in a particular protein, with each triple-nucleotide codon placing one of 20 possible amino acids at each successive position in the protein chain. The sequence of amino acids in turn specifically establishes both the structure and the function of the protein. Thus the nucleotide sequence of DNA precisely specifies the protein-building properties of the organism. Moreover, virtually every property of the organism, from enzymatic action to eye color, depends on protein structure in one way or another.

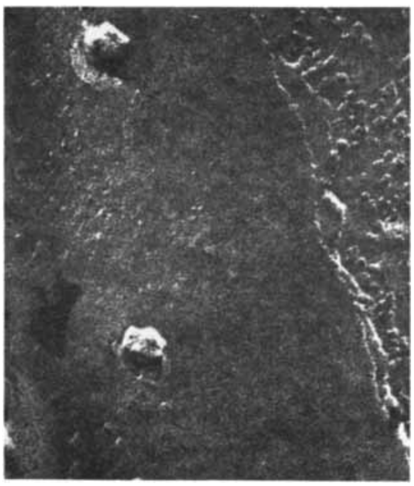
The transmission of the essential genetic information between generations depends on the precise replication of the nucleotide sequences of DNA. The mechanism for replication stems from the complementary relation between the two strands of the DNA molecule. A sequence on one strand (for example *CATTACTAG*) lies immediately opposite a complementary sequence (*GTAATGATC*) on the other strand. The strands are complementary because *C* and *G* are always opposite on the intercoiled strands, as are *A* and *T.* Complementarity depends on the special chemical affinity, or binding, between *C* and *G* on the one hand and *A* and *T* on the other. The sum of these bonds, repeating along the length of the strands, is what holds the strands together in the double helix. Under appropriate conditions affinity is reduced and the two strands can unwind and separate. The single strands can again pair and rewind when conditions for high affinity are restored.

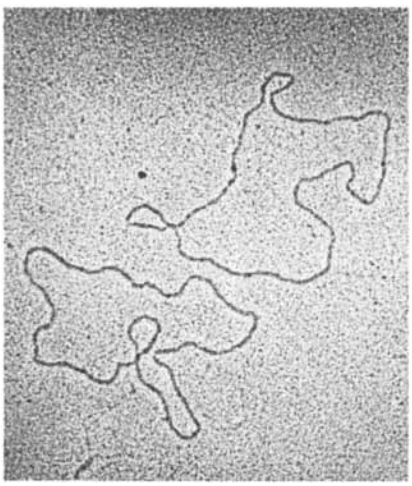
Double-strand DNA replicates by means of an extension of these properties. The unwinding and separation of the strands begins at a localized site along the DNA molecule. In the presence of suitable enzymes and free nucleotides a new chain is formed next to the exposed portion of each unpaired older chain. Each nucleotide lines up next to its opposite number (*C* next to *G, A* next to *T*). The complementary sequence thus established is then linked end to end by an enzyme that closes the nucleotide couplings. When the replication process has traveled along the entire length of the original double helix, two new helixes identical with the first one have been formed. The replication of DNA is the most fundamental chemical reaction in the living world. It fully accounts for the classical first principle of heredity: like begets like.

If DNA replication always worked without error, life would be far more homogeneous than it is. Here, however, a second classical principle of heredity intervenes: the principle of mutational variation, or the appearance in the offspring of new hereditary characteristics not present in the progenitors. Mutations arise through error, at least partly in the replication process. For example, the substitution of one nucleotide by another changes the triplet codon and puts a different amino acid in the corresponding position in the resulting protein. Single-nucleotide errors lead to single-amino-acid errors. Thus, a single-nucleotide error is responsible for the human disease sickle-cell anemia. Most mutations are not such simple, single-nucleotide exchanges; nevertheless, they correlate directly with altered, transposed or deleted nucleotide sequences in DNA. When these changes appear in a gene (that is, a segment of DNA that codes the amino acid sequence of a particular protein), a change in the protein and hence in the hereditary properties it controls is the result.

Therein lies the crux of recombinant-DNA technology. It makes possible for the first time the direct manipulation of nucleotide sequences. Changes in nucleotide sequence that are produced by "natural" errors are random, even when their overall frequency is artificially increased. In natural populations Darwinian selection "chooses" among the random errors, increasing the representation in breeding populations of those errors that lead to more offspring in particular environments. Artificial selection, practiced by human beings for millenniums, favors errors that meet human needs (agricultural breeding) or whims (exotic-pet breeding). The success of both natural and artificial selection, however, is dependent on the random occurrence of desirable mutations. There was no way to direct genetic change itself until recombinant-DNA techniques came along. The new techniques enable one to deliberately introduce known and successful nucleotide sequences from one strain or species into another, thereby conferring a desired property.

The recombinant-DNA approach involves experimental ingenuity and detailed knowledge of the DNA molecule. It begins with an attack on DNA by the proteins called restriction enzymes, which are isolated from bacteria. The enzyme attack breaks the double chain



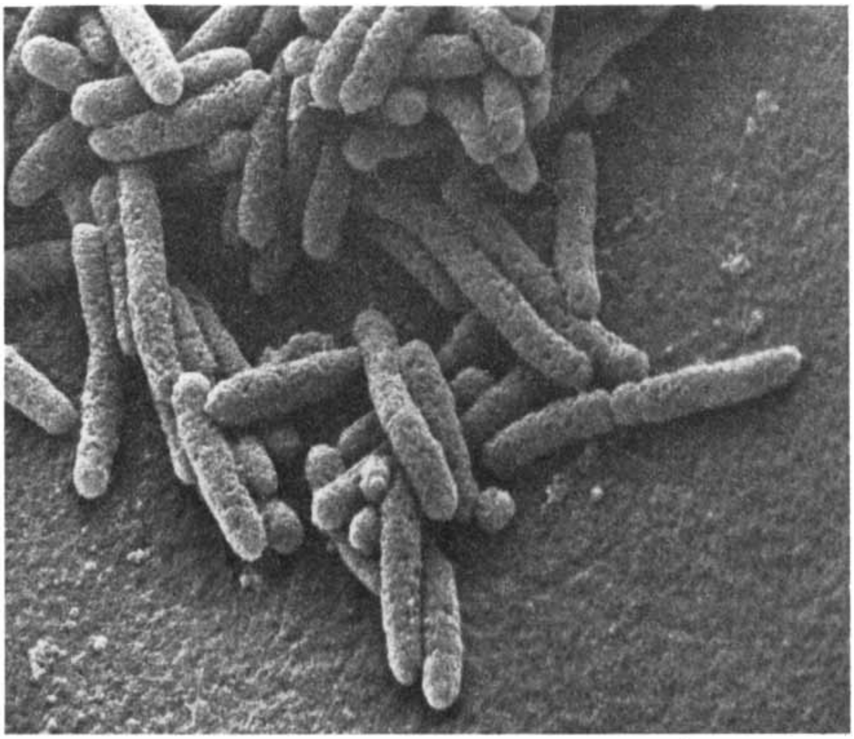


BACTERIOPHAGE ΦX174 AND ITS DNA are portrayed in this pair of electron micrographs. The virus infects the common intestinal bacterium *Escherichia coli.* In the micrograph at left, made by Jack D. Griffith of the Stanford University School of Medicine and Andrew Staehelin of the University of Colorado, two φX174 particles are seen attached to surface of an *E. coli* cell. In micrograph at right, made by Griffith, the DNA molecules of two φX174 viruses are seen in their double-strand form; each molecule is about 18,000 angstroms long.

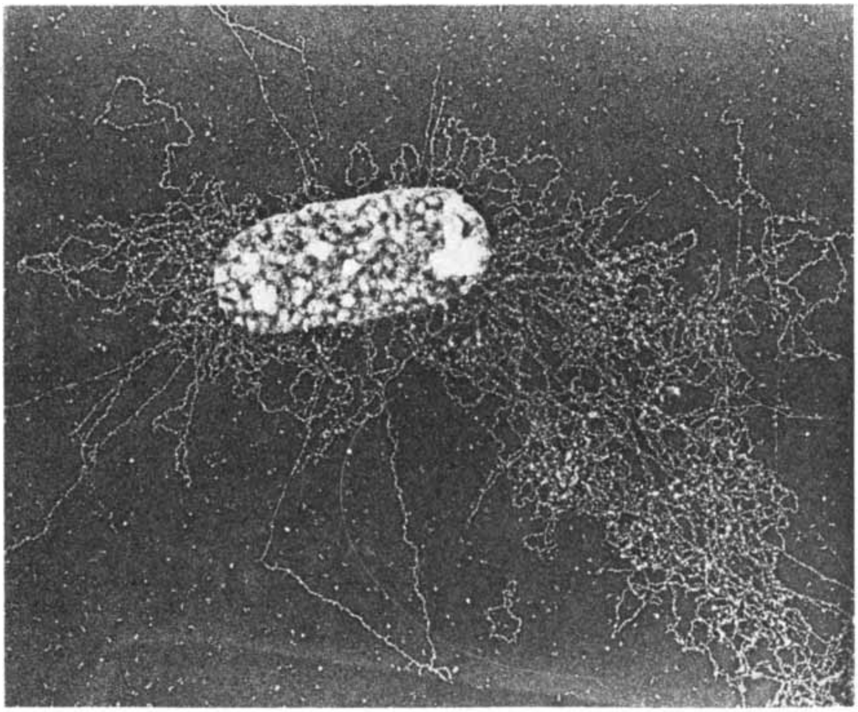
of DNA at particular sequences, say at the sequence *CATTAC*, which is opposite the complementary sequence *GTAATG*. The break does not always occur at the same point on the two strands. It may, for example, be between the two *T*'s in the first strand but just to the right of the *ATG* in the second strand. On separation one piece therefore ends in *TAC*, whereas the other ends in *ATG*. Since the single-strand ends are complementary, they will under suitable conditions stick side by side, and they can then be coupled together end to end. If the same restriction enzyme is used on the DNA from two different sources, both of which have the appropriate target sequence, then sequences with the same "sticky" ends will result. By taking advantage of this stickiness two sequences from any source can be recombined into a single DNA molecule.

The only further step necessary is to put the recombinant DNA into a suitable host organism. The recombinant must have the ability to penetrate the host and become part of its genetic system. An effective way to accomplish this has been developed for the common intestinal bacterium *Escherichia coli*. In addition to its single large circular chromosome the *E. coli* bacterium may have one or more independently replicating, smaller loops of DNA known as plasmids. The plasmids can be isolated from the bacteria, broken open by restriction enzymes and used as one component of a recombinant. After linking up the plasmid DNA with the "foreign" DNA the circular form of the plasmid can be restored and the structure returned to a whole cell. There it can resume replication, duplicating not only its own native sequence but also the foreign one. A strain of bacteria is thus obtained that will yield an indefinite number of copies of the inserted nucleotide sequence from the foreign source.

Standing alone, none of this appears to be particularly momentous or threatening; it is only a new and intriguing kind of chemistry applied to living organisms. Given the complexity of living organisms and the still more complex world of social phenomena, however, this new chemistry quickly builds into varied new potentials, both speculative and real. Suppose, for example, one were to isolate the nucleotide sequence necessary to produce a potent toxin and to transfer it to *E. coli*, usually a harmless inhabitant of every human intestinal tract. Would a dangerous new pathogen be created? Would the transformed *E. coli* release a toxin in the human gut? Might such a new pathogen escape from control and induce epidemics? Questions of this kind have answers, but they take time to find. To gain some time for reflection investigators in 1974 called for a partial and temporary moratorium on those experi-



PILE OF E. COLI CELLS appears in this scanning electron micrograph made by David Scharf. Some of the cells have been caught in the act of asexual reproduction (cell division); a few appear to be transferring their DNA by means of the threadlike connection characteristic of the process known as conjugation. *E. coli* bacteria are considered by most investigators to be most suitable host cells for recombinant-DNA experiments. Magnification is 11,000 diameters.



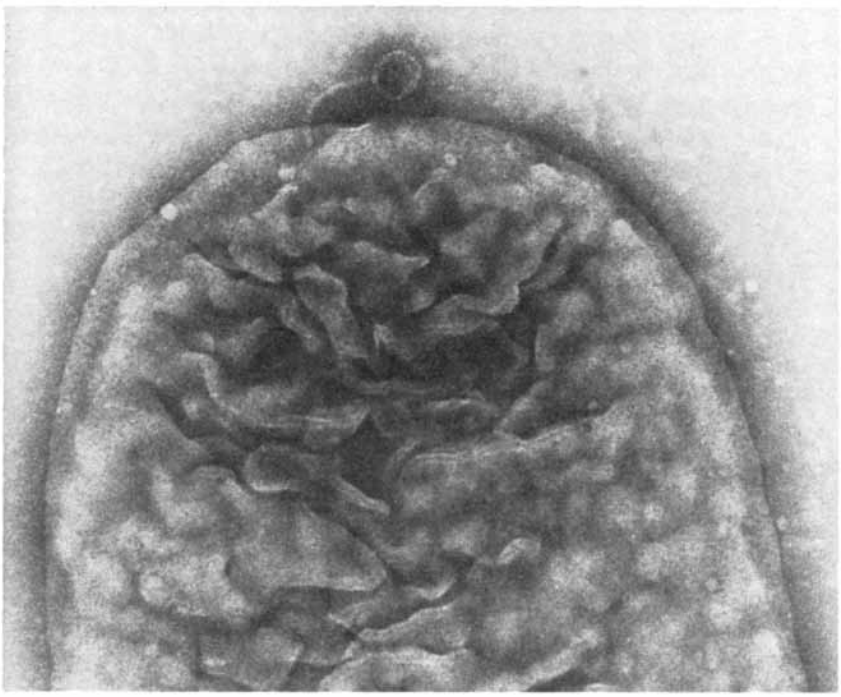
E. COLI SPEWS OUT DNA through its chemically disrupted cell wall in this electron micrograph by Griffith. Most of the DNA is in the form of a single large molecule of double-strand DNA, which constitutes the chromosome of this simple prokaryotic organism. In addition the *E. coli* bacterium may have one or more of the independently replicating loops of DNA known as plasmids; one of these smaller extrachromosomal DNA molecules can be seen near the bottom. Plasmids derived from *E. coli* cells play an important role in recombinant-DNA research, since they form one class of vectors, or carriers, into which segments of "foreign" DNA can be spliced prior to their being reinserted into an appropriate host cell to propagate, thereby duplicating not only their own native nucleotide sequence but also the foreign sequence.

ments thought to be potentially the riskiest. The separation of the certainly safe experiments from the less certainly safe ones became the chief function of the guidelines released by the NIH in June, 1976. The guidelines, which replaced the temporary moratorium, were derived from worst-case analyses of various kinds of experiments; the object was to evaluate the possible range of hazards and to prescribe appropriate matching safeguards in order to minimize the unknown risks. The guidelines assigned heavy responsibility to individual investigators, and they buttressed this responsibility with special monitoring committees in the sponsoring institutions and in the funding agency.

If such regulations have been adopted, why is debate continuing? Briefly, it is because the matching of estimated risk and prescribed containment adopted by the guidelines is regarded by critics as being inadequate in dealing with potential biohazards and incomplete in failing to address other important issues. The most vocal critics have presented their own worst-case analyses in the scientific and general press. These accounts have led to widespread alarm and to public-policy deliberations at the level of local communities, states and the Federal Government. The expressed concerns of the critics have generated a revised agenda for what is now emerging as a broadened second round of policymaking.

Potential biohazards and estimated degrees of risk continue to dominate the debate. The NIH guidelines balance the estimated risk of a given experiment and recommend specific measures for containing the risks. (Risk, it must be remembered, means possible danger, not demonstrated danger.) Those experiments judged to present an excessive risk are entirely proscribed. At the other end of the spectrum experiments judged to present an insignificant risk require only the safeguards of good laboratory practice. Between these extremes the guidelines establish various levels of estimated risk and prescribe combinations of suitably increasing physical and biological containment. The release into the environment of any recombinant organisms is forbidden.

Unfortunately, given the growing but still limited state of knowledge, wide disagreement is possible, both as to estimated degrees of risk and as to the efficacy of the proposed containment. Some critics project fragmentary information into the inevitable spread of dangerous, newly created organisms, threatening both the public health and the environment. Some defenders project the same fragmentary information to the conclusion that the NIH guidelines are already overly cautious. They believe the actual hazard under existing precautions will turn out to be no greater than that routinely faced in the use of automobiles, jet aircraft and other accepted technologies. The wide range of estimates is possible because of the multiplicity of conceivable experiments and because experience and critical data are inadequate for certainty on many points. One fact that is certain is that no known untoward event has yet resulted from recombinant-DNA research.

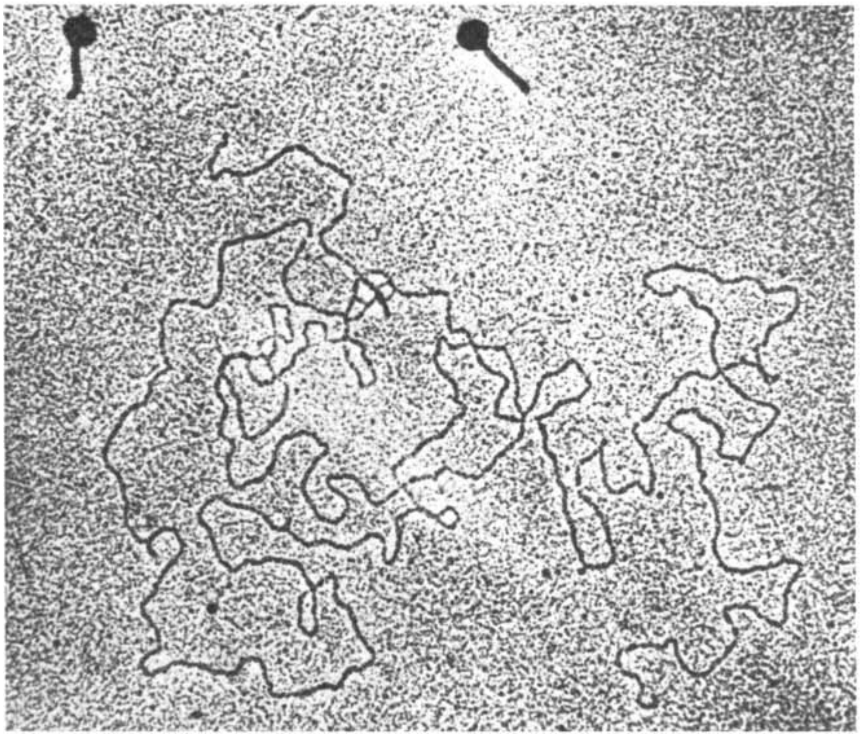


BACTERIAL VIRUS IS ATTACHED to the wall of an *E. coli* cell in this electron micrograph made by Maria Schnoss of the Stanford School of Medicine. This particular virus, named bacteriophage lambda, normally infects the bacterium by injecting its DNA into the host cell through a long taillike appendage. The magnification is approximately 140,000 diameters.

What emerges on the new policy agenda, then, is the need for effective policy-oriented research to reduce the current uncertainty as to the risk of particular kinds of experiments. For example, there is dispute over the use of *E. coli* as a host for recombinant DNA. One side argues that scientists must be mad to pick a normal human inhabitant (and a sometime human pathogen) to serve as a host for recombinant DNA. This view, in extreme form, demands the suspension of all recombinant-DNA research until an organism safer than *E. coli* can be found. The other side argues (1) that the vast amount of information available on *E. coli* makes it invaluable, (2) that the *K*-12 strain of *E. coli* actually used in laboratory research has been so modified genetically in adapting to laboratory conditions that it survives only with difficulty in the human intestine and (3) that new strains of *K*-12 have been developed with additional genetic deficiencies that will make survival outside of laboratory conditions essentially impossible. The use of such genetically deficient strains is what is meant by the term "biological containment." The concept is supported by proponents of the research as an efficacious new approach to safety and derided by critics as likely to be circumvented by natural recombination.

Such differences of opinion are normally reduced by scientists to experimental questions. For example, the suitability of the *K*-12 strain of *E. coli* as an experimental organism can be judged only from the effect of recombinant genes on the ecological relations of *E. coli* within the human intestine, including the degree of success of recombinant strains in competing with other strains of *E. coli* and with other organisms. Information on these matters is growing. Such questions, however, are not normally subjects of profound scientific interest. They have recently become matters of priority only because they may provide information that would be useful in arriving at a policy decision. Research on policy-oriented questions has never had a very high status among scientists engaged in basic research or even among those engaged in applied research. Therefore policy-oriented research must be encouraged through special funding mechanisms and through suitable new institutional arrangements. A regulatory agency for recombinant-DNA research and other conceivably hazardous kinds of research is urgently needed outside the NIH, and it should include a research component. The Center for Disease Control and its National Institute for Occupational Safety and Health come to mind as possible models for such a dual-purpose agency.

Also related to the question of biohazards is a controversy over the desirability of centralizing recombinant-DNA research facilities. Some of those who fear severe dangers from recombinant organisms have urged that the potentially more hazardous research be concentrated in remote places with extremely stringent containment procedures. Those who minimize the hazard are opposed to the concentration concept because it would tend to separate the research from the intellectual mainstream and would be unnecessarily expensive in facilities. The argument has been particularly strenuous with respect to experiments requiring *P*3 facilities, which are defined as those necessary to contain "moderate risk" experiments. *P*4 facilities for "high risk" experimentation are expected to be fewer in number because of their high cost; generally speaking they are likely also to be comparatively isolated. The current NIH guidelines provide little direction in these matters. A decision on a firmer policy belongs on the discussion agenda. Particularly urgent is careful consideration of such intermediate possibilities as the use of centralized, high-risk facilities for making particular recombinations for the first time. These activities, together with preliminary testing of new recombinants for possible hazards, might also be carried out by the proposed new regulatory agency.



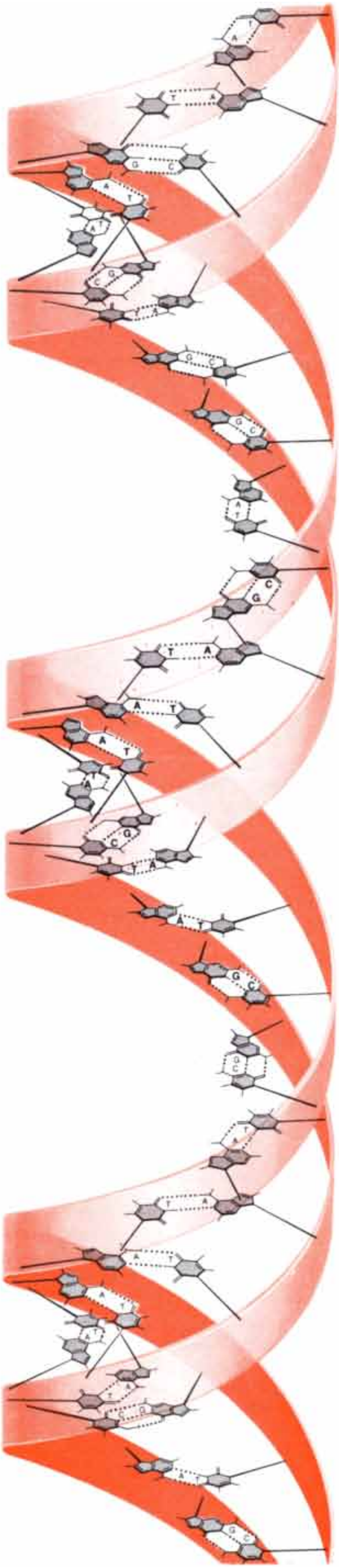
BACTERIOPHAGE LAMBDA AND ITS DNA are both represented in this electron micrograph provided by Griffith. Two complete lambda viruses are at the top; the long double-strand DNA molecule of a disrupted lambda is below them. DNA from bacteriophage lambda can also serve as a vector for recombinant-DNA experiments involving *E. coli* host cells.

A special case that emphasizes the advantages of initial testing in a central facility is provided by what are called "shotgun" experiments. These experiments, which offer special advantages to the investigator, may also present special hazards. Shotgun experiments involve exposing the total DNA of a given organism to restriction enzymes in order to obtain many DNA fragments. The fragments are then each recombined with DNA from a suitable vector and the recombinants are randomly reinserted into *E. coli* host cells. The next step is to spread the *E. coli* cells on a nutrient substrate so that each recipient cell, containing a particular inserted foreign sequence, grows into a colony. If the experiment is successful, the yield is a "library" of all the nucleotide sequences of a particular organism, each sequence growing in a separate strain and accessible to manipulation and cross-combination at will.

This experimental approach is laborious but far less so than anything else available for the exploration of the complex genetic systems of higher organisms. There is, however, a risk of unknown magnitude that portions of the DNA with unknown or repressed functions might duplicate and create unanticipated hazards. The result might be particularly unfortunate if the original DNA preparation were to contain genetic material from parasites or from viruses associated with the species under study. Under the NIH guidelines, therefore, shotgun experiments are regarded as being more dangerous than those involving purified and characterized DNA. Experiments in this category are treated as being increasingly more dangerous as the test organism under study is biologically more like the human organism. Thus experiments with primate DNA are considered to be more dangerous than experiments with mouse DNA. This approach appears to represent a reasonable precaution with respect to human health hazards, but it is less reasonable with respect to potential ecological effects. For example, shotgun recombinants involving DNA from plant sources could conceivably lead to ecologically dangerous effects if they were to escape into the environment. Shotgun procedures might therefore be best conducted first in special centralized facilities that could also act as storage and distribution centers for the recombinant products once they had been tested for safety.

These examples suggest several advantages for the creation of a Center for Genetic Resources. The center might not only carry out DNA recombinations suspected to be hazardous but also function to preserve genetic information contained in threatened natural species and in special strains of cells or organisms developed for research and other purposes. Stored genetic information can be expected to be increasingly important in the future. For example, new genetic infusions into domesticated stocks of plants and animals from their wild progenitors have long been used to strengthen the response of the domesticated stocks to changing conditions of husbandry. The sources of wild progenitors are threatened by the reduction of wild habitats all over the world.

The possibility of a biohazard need not arise only as a by-product of basic research. The practical applications of recombinant-DNA techniques, together with the applied research and development leading to them, are at least equally likely sources. For example, recombinant techniques may enormously expand the use of bacteria (and other microorganisms) for the production of certain proteins and other pharmacological products. Microorganisms have long played an essential role in the food, beverage, pharmaceutical and chemical industries, and more precise genetic control of their characteristics has already yielded large benefits. The recombinant-DNA techniques not only offer advances on current practice but also suggest a new realm of "bacterifacture" in which the rapid, controlled growth of microorganisms is coupled to the pro-



duction of specific products normally made only by higher organisms. Included among the possibilities are the production of insulin, blood-clotting factors and immunological agents. The probability of those possibilities ever being realized is no more easily assessed than the risks, but success in realizing them clearly could provide substantial economic and social benefits. Accordingly entrepreneurial interests have been aroused.

The NIH guidelines are silent on the matter of commercial applications other than stipulating that large-scale experiments (beyond production batches of 10 liters) with recombinants "known to make harmful products" be prohibited unless specially sanctioned. The guidelines also require detailed reporting of proposed recombinant-DNA experiments, a provision that runs counter to the protection of proprietary interest. There have been discussions of these matters between the NIH and representatives of industry. In addition industry spokesmen have testified at Congressional hearings. It is known that some industrial research already is under way and that representatives of industry generally endorse the precautionary approach of the NIH guidelines, but they are resistant to limitations on proprietary rights and on the size of batch production. Moreover, patent policy has come up as an issue and there has been some uncertainty in the Department of Commerce as to how it should be handled. Indeed, the possible commercial applications of recombinant-DNA techniques have yet to be publicly evaluated as a serious policy question, and they must be high on the agenda of the next round of discussions.

The problems of commercial applications lead from immediate issues to broader ones and to a larger time frame. Recombinant-DNA techniques have revived the debate over "genetic engineering" and have once again raised questions about the applications of fundamental biomedical research to technology, to the quality of life and to the future of society. Recombinant DNA has now joined nuclear fission, overpopulation, famine and resource shortages in the doomsday scenarios of "creative pessimism." These issues are even more difficult to deal with objectively than those related to potential biohazards, but they are plainly apparent in the general public discussion and in the public statements of respected scientists.

For example, Robert L. Sinsheimer of the California Institute of Technology has persistently raised issues that are in part practical and in part philosophical. Along with George Wald of Harvard University and Erwin H. Chargaff of the Columbia University College of Physicians and Surgeons, he suggests that the entire recombinant-DNA approach to gaining an understanding of the complexities of higher genetic systems is misbegotten. The argument is not that the approach may not work but that its alleged huge risks are unnecessary because less risky, although slower, means are available. Sinsheimer emphasizes the fundamental difference between simple prokaryotic organisms such as bacteria and complex eukaryotic organisms, including human beings. Prokaryotes, typically one-cell organisms, have a single, comparatively simple chromosome floating freely within the cell body, whereas eukaryotic cells have a nucleus that is bounded by a membrane and contains a number of far more complex chromosomes. The paleontological record suggests that prokaryotes existed on the earth for a billion or more years before the more complex eukaryotes arrived on the scene. Sinsheimer proposes that throughout the evolution of the eukaryotes there has been a genetic barrier between them and the prokaryotes, behind which eukaryotes have developed their more complex mechanisms of genetic control. To transfer these mechanisms, which are possibly the key to the evolutionary success and enormous diversity of eukaryotes, to prokaryotes may introduce, he says, incalculable evolutionary dangers. The prokaryotes may be made far more effective, both as competitors and as parasites, negating an ancient evolutionary strategy.

Sinsheimer's argument has won only a few vocal adherents among biologists, and he himself concedes that it is speculative. Nevertheless, his argument

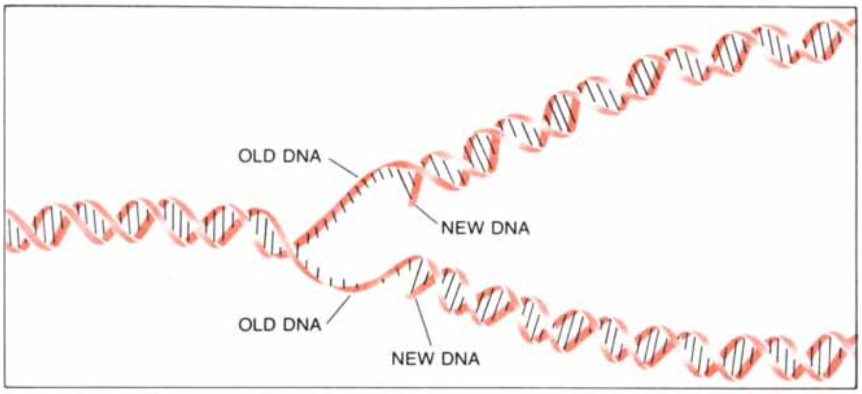
DOUBLE-HELIX STRUCTURE OF DNA is evident in this simplified diagram of a short segment of the deoxyribonucleic acid (DNA) molecule. The sugar and phosphate groups that are linked end to end to form the outer structural "backbones" of the double-strand molecule are represented schematically here by the two helical colored bands. The inner portion of each polynucleotide chain, drawn in somewhat greater detail, consists of a variable sequence of four kinds of bases: two purines (adenine and guanine, or *A* and *G*) and two pyrimidines (thymine and cytosine, or *T* and *C*). The two chains, which run in opposite directions, are held together by hydrogen bonds (*dotted black lines*) between pairs of bases. Adenine is always paired with thymine, and guanine is always paired with cytosine. The planes of the bases are perpendicular to the common axis of the two helixes. The diameter of the double helix is 20 angstroms. Adjacent bases are separated by 3.4 angstroms along the axis and are displaced successively around the axis by an angle of 36 degrees. The structure therefore repeats after 10 bases on each chain (360 degrees), or at intervals of 34 angstroms. The genetic information is stored in the sequence of bases along each chain. In this case the sequence *CATTACTAG* on one strand is identified in boldface type opposite complementary sequence *GTAATGATC* on other strand.

has attracted significant public attention, and it is widely cited to support opposition to continued recombinant-DNA research. Bernard D. Davis, a Harvard Medical School microbiologist, has provided a rebuttal, particularly with respect to the concept of a genetic barrier between prokaryotes and eukaryotes. He believes there has been an ample and continuous opportunity for the exchange of DNA between the two groups. He points out that bacteria can take up naked DNA from their immediate environment and that *E. coli* would be exposed to such DNA arising from dead human cells in the human intestine. Microorganisms might similarly take up DNA in the process of decomposing dead animals. Therefore, Davis argues, most recombinants probably have already been tried in the natural evolutionary arena and have been found wanting. Reasoning on analogy with extensive information on pathogenic bacteria, Davis concludes that under the existing NIH guidelines the probability for survival in nature of laboratory-produced prokaryote-eukaryote recombinants is vanishingly small.

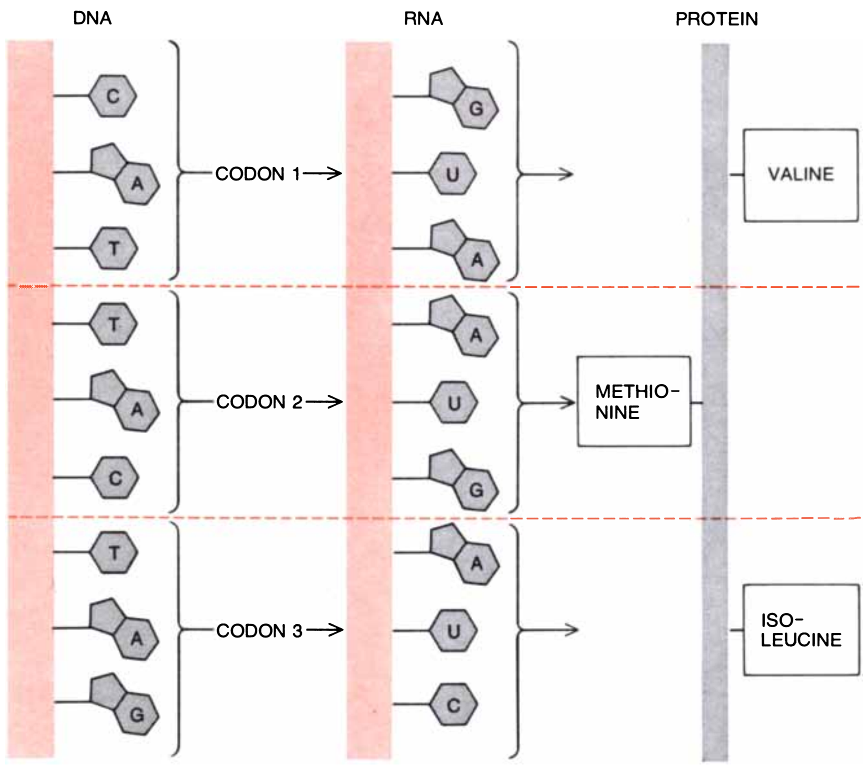
This clash of opinion on a major biological issue illustrates the difficulty of assessment of even comparatively value-free questions when critical information is fragmentary. The controversy over the risk-benefit ratio becomes even more intense when issues involve substantial value judgments as well. Here again a concern of skeptics and opponents of recombinant techniques is sharply articulated by Sinsheimer. He asks: "Do we want to assume the basic responsibility for life on this planet? To develop new living forms for our own purposes? Shall we take into our own hands our own future evolution?" Since the questions include such concepts as responsibility, purpose and control of the future, they clearly involve considerations beyond science alone.

The human species has, of course, been altering life on this planet from the beginnings of human culture. When hunting and gathering gave rise to animal husbandry and agriculture, human choice and purpose began to influence the evolution of selected species. Unconscious human selection was replaced by deliberate plant and animal breeding, and the further development of human culture is now clearly altering the entire ecosystem. Moreover, the biocultural progression of the human species, based partly on human purpose, is undoubtedly altering the human gene pool and will slowly modify the species in unpredictable ways. Nevertheless, the advent of recombinant-DNA techniques has obviously enhanced the prospects for genetic engineering and has restressed the need to assess its implications. Can it be assumed that success in introducing recombinant DNA's into *E. coli* means that there will be similar success in introducing them into the human species? If it can, what is the probable time frame for applying the technique to the human species? Is it accurate and responsible to suggest that we have almost in hand control of "our own future evolution"? These certainly are questions for scientific assessment, and they should have a prominent place on the new policy agenda.

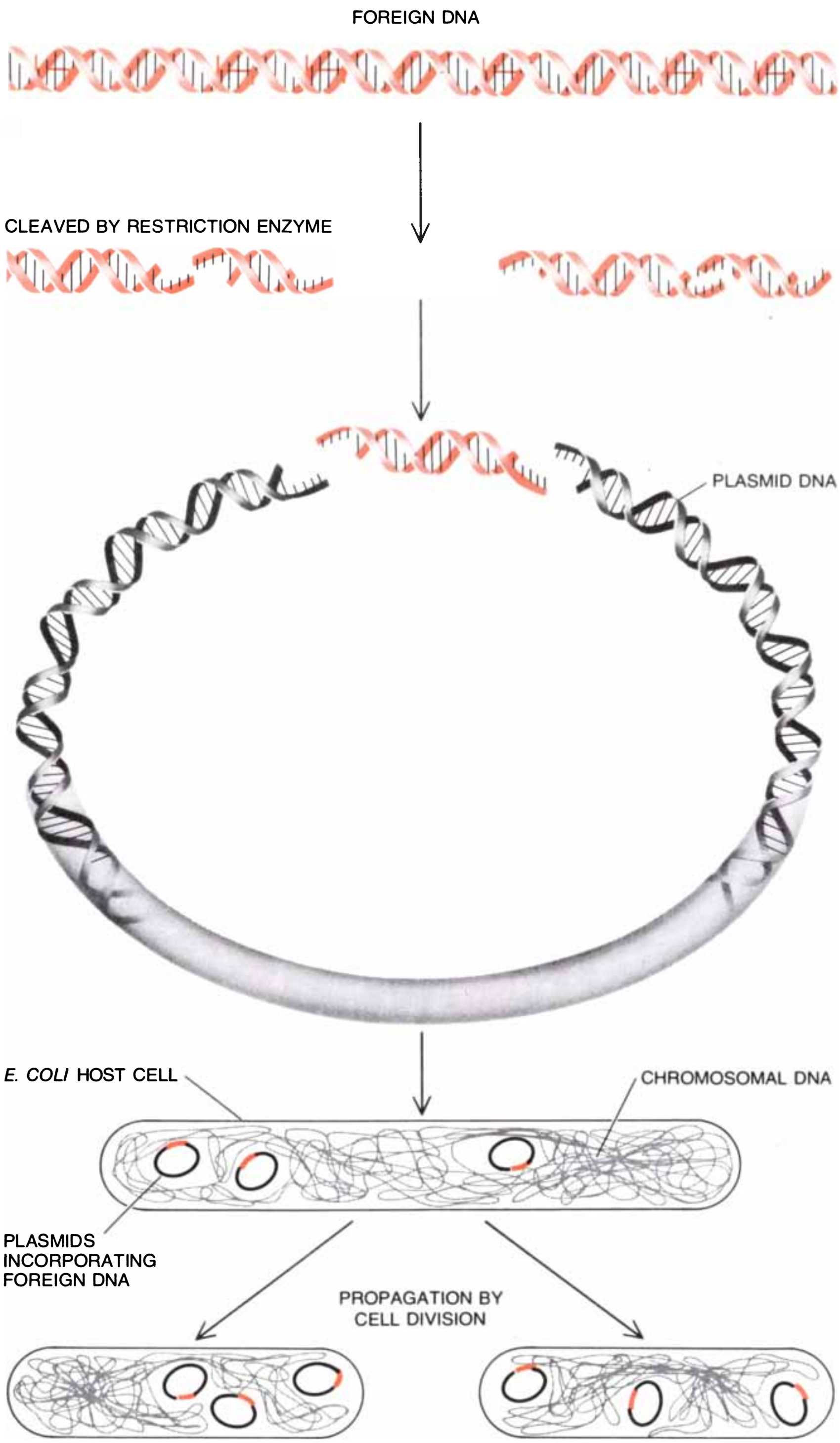
Sinsheimer has gone into still another controversial area, not only for the scientific community but also for the entire



REPLICATION OF DNA depends on the complementary relation between the nucleotide sequences on the two strands of the DNA molecule. Under appropriate chemical conditions the hydrogen bonds between the bases are weakened and the two strands can unwind and separate. In the presence of suitable enzymes and free nucleotides a new chain can be formed next to the exposed portion of each unpaired older chain. The complementary sequence that is formed by each nucleotide lining up next to its opposite is then linked end to end by an enzyme that "zips up" the nucleotide couplings. In this way two new helixes identical with the first can be formed.



ROLE OF DNA IN PROTEIN SYNTHESIS is suggested by this highly schematic diagram. The genetic message contained in the nucleotide sequence *CATTACTAG*, for example, is "written" in the form of the triplet "codons" *CAT*, *TAC* and *TAG*. Each codon determines, through a series of intermediate steps involving a molecule of ribonucleic acid (RNA), the position of a specific amino acid in a protein molecule. Thus the sequence of nucleotides in a given DNA molecule specifies the corresponding sequence of amino acids in a particular protein, with each triple-nucleotide codon placing one of 20 possible amino acids at each successive position in the protein chain. Since the sequence of amino acids in turn establishes both the structure and the function of the protein, the nucleotide sequence of DNA determines virtually every property of organism. Letter *U* stands for the pyrimidine uracil, a constituent of RNA.



RECOMBINANT-DNA TECHNIQUE makes it possible for the first time to deliberately introduce nucleotide sequences from the DNA of one strain or species of organism into the DNA of another. The DNA of the "foreign" organism is first treated with restriction enzymes, which cleave the double-strand molecule at particular nucleotide sequences (typically thousands of base pairs apart) on a random basis. The same enzyme is then used to cleave the DNA of a suitable vector, in this case a plasmid isolated from *E. coli* bacteria. Since the break caused by the enzyme does not occur at the same point on both strands, the chemical treatment results in a mixture of DNA segments that have complementary single-strand ends. Under suitable conditions the "sticky" ends of two different sequences can be coupled to form a single DNA molecule. For example, after recombining the foreign DNA with the plasmid DNA the circular form of the plasmid can be restored and the structure can be inserted into a suitable host cell (in this case *E. coli*), where the plasmid can resume replication, thereby propagating an indefinite number of "cloned" copies of the inserted nucleotide sequence from the foreign source.

society. Arguing that time may be needed to "pace" new genetic knowledge to human capacities for putting nature to intelligent use, he wonders whether "there are certain matters best left unknown, at least for a time." This is high heresy in the scientific community, whose fundamental premise is that the growth of knowledge is the driver and not the captive of other values. The rejection of the concept of "forbidden knowledge" was part of the heroic period at the beginning of modern science, when it included willingness to face the Inquisition and the stake. Having been seared by the nuclear flame and now confronting the more subtle implications of the innermost language of life, 20th-century scientists fear not the stake but the judgment of history. Chargaff, a pioneer in the investigations that led to the decipherment of the genetic language, says: "My generation, or perhaps the one preceding mine, has been the first to engage, under the leadership of the exact sciences, in a destructive colonial warfare against nature. The future will curse us for it."

Sinsheimer and Chargaff, along with a number of philosophers, historians and sociologists of science, are clearly suggesting that the possible consequences of knowing must be consciously included in decisions about the directions of the search for knowledge itself. No issue cuts more deeply to the core of modern science. The self-doubt expressed by some scientists reflects a general questioning in the U.S. of the net benefits of science and technology. Cost-benefit analysis is a current preoccupation, and it is being increasingly applied to the generation of knowledge itself. It is hard enough to assess what we may gain or lose from particular new knowledge; it is even harder to assess the costs of not having it. This problem is epitomized by the recombinant-DNA controversy. The rise of molecular genetics in the U.S. is the direct product of a series of decisions made after World War II that provided funds for biomedical research. The objective was the conquest of the "killer" diseases: cancer, heart disease and stroke. Those diseases are still much with us, although they are better understood and cared for. Meanwhile, out of Federally supported research also came the impetus that led to the discovery of the double helix, the genetic code, the structure of proteins and recombinant DNA. In a classic "double take" the public is now asking whether it has been buying health and well-being or chimeric monsters. Is molecular genetics and all biomedical technology a sorcerer's apprentice? Are we increasing rather than lessening our burden of pain and anxiety?

The last question leads to yet another issue. Biohazard and ecohazard may arise inadvertently, but "sociohazard" may be the product of deliberate malevolence. The U.S. is a signatory to an international legal convention that has renounced biological warfare, including research to produce the necessary agents. Not all countries have taken this step, and public renunciation without adequate inspection cannot ensure that covert activities do not exist. Opponents of recombinant-DNA research see its techniques as being ideally suited to serve malevolent purposes, either as agents of organized warfare or of sabotage and terrorism. The techniques do not require large installations or highly sophisticated instrumentation. Contrary views have not denied this but have noted that recombinant-DNA techniques would not be the first technology to have potential malevolent applications. Explosives have such applications, but society does not completely ban them; it takes prudent precautions against their misuse.

Nevertheless, the issue of the possible misuse of recombinant-DNA technology deserves a place on the policy agenda, because it emphasizes the need for international discussion of the implications and management of recombinant-DNA research and recombinant-DNA applications. It can be argued that the U.S. is not ready for such discussion until its own policies are in better order. It is not too early, however, to begin the internal consideration of how best to approach the international arena.

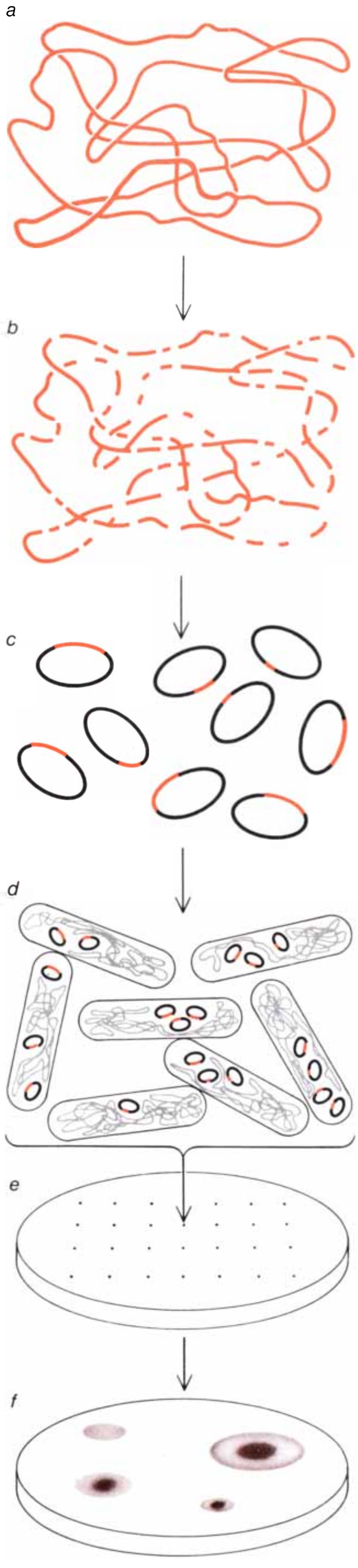
These are the chief issues that have emerged from the policy debate so far. It is a not inconsiderable list. The debate has not been raging on every street corner, but it became strenuous enough in Cambridge last summer to have repercussions across the continent. For example, an evaluation presented by a panel of nonscientists to the Cambridge City Council was not too different in content from one produced by a task force of the Quality of Life Board of the City of San Diego, where I live. Both groups accepted within their community the continuance of recombinant-DNA research requiring *P*3 facilities but sought somewhat greater assurances of safety than those provided by the NIH guidelines. Meanwhile the Attorney General of the State of New York held a public hearing, and a joint hearing was conducted by two committees of the Assembly of the State of California Legislature. Legislation regulating DNA research was later introduced in the California Assembly, and it is still under consideration. Congress has also held several hearings and various items of regulatory legislation have been introduced in both the Senate and the House of Representatives. These local, state and Federal initiatives emphasize the necessity to get on with the policy agenda.

The agenda should be viewed in at least two time frames: immediate and longer range. A consensus has been growing that there is an immediate need to give the quasi regulation represented by the NIH guidelines a statutory base. In particular, regulation must be extended to activities not supported by Federal agencies, especially in the industrial sector. However this is to be done, it is important to maintain flexibility, since the problems to be dealt with will change as greater knowledge and experience are acquired.

Moreover, given the complexity of the longer-term issues, immediate legislation probably should be provisional and limited. A mechanism should be included, however, that actively leads toward a more definitive future policy. This requires provision for a new, comprehensive assessment of all the issues raised by recombinant-DNA research, including the probable effectiveness of the regulatory devices put in place under the NIH guidelines.

The need for such a new national assessment is demonstrated by the nature of the critical challenge to the product of the earlier assessment. First, it has been alleged that the 1975 Asilomar conference establishing the pattern for the NIH guidelines was dominated by scientists involved in the research, and therefore it could not yield a broad enough perspective. Second, it is argued that the earlier assessment was devoted primarily to the question of potential biohazards and did not address in any depth other gravely important questions. The passage of time has added several more points: that circumstances already have changed as research has progressed, that experience has grown and that a wider range of opinion has come to bear on the issue. Whatever format is adopted for the reappraisal of the recombinant-DNA issue, the public must be assured that the process is a comprehensive and objective one.

Whoever undertakes this new national review should first carefully examine the current situation, including the actual effectiveness of the regulatory mecha-



"SHOTGUN" EXPERIMENT is a type of recombinant-DNA experiment in which the total DNA of an organism (*a*) is exposed to restriction enzymes in order to yield many fragments (*b*), which are then recombined with the DNA from a suitable vector (*c*) and randomly reinserted with the vector into the host cells (*d*). The *E. coli* hosts are next spread on a nutrient substrate (*e*) so that each recipient cell, containing a particular inserted foreign nucleotide sequence, can grow into a colony (*f*). The result, if the experiment is successful, is a "library" of all the nucleotide sequences of the organism. Under the guidelines issued by the National Institutes of Health last year shotgun experiments are regarded as being potentially more hazardous than those involving purified and characterized DNA, since it is not known whether portions of the DNA with unknown or repressed functions might cause unexpected problems.

nisms provided by the NIH guidelines. Particular attention needs to be paid to the local institutional biohazards committees mandated by the NIH guidelines. Beyond the responsibility assigned to the principal investigator these committees are the only source of local surveillance and standard-setting. Their composition and charge are unique, yet their authority and procedures are stipulated only generally in the NIH guidelines. They may well need the stimulus and support of external interests to carry out their important task. Moreover, no provision has been made for budgeting what may turn out to be their considerable cost for technical surveillance, personnel training and medical monitoring. Like all insurance, security against biohazard must be bought. The cost should be borne as an additional expense of the research, not as a competitor for existing funds.

Similarly, the actual performance of the NIH study sections, which are mandated by the guidelines to be independent evaluators of biohazards and containment, needs to be examined. Study sections are already heavily overloaded with the job of evaluating scientific quality. Yet these part-time peer groups are asked to assume another difficult function. If the responsibility is to be taken seriously, it too will entail additional costs.

Of special importance for early attention is an effective monitoring system for following the actual directions of recombinant-DNA research. The techniques involved are so rich in possibilities, whether for fundamental research or applications, for benefit or risk, that "early warning" is essential. Systematic following of the directions of investigators' interests, from applications for support through informal communication to formal publication, is essential to the early detection and assessment of either risks or opportunities. Needless to say, monitoring is particularly difficult in industrial research. It might therefore be desirable to limit or postpone certain development efforts pending closer study and greater knowledge of the underlying problems.

Equally urgent is a determined effort toward a more effective assessment of risks and their limitation. The specific assignment of responsibility for this kind of policy-oriented research should be an early recommendation of the body undertaking the reassessment. Given the differing perspectives required by regulation and the NIH mission to promote health-related research, the regulatory function probably belongs elsewhere in the long run. On the other hand, given the need for careful study of the implications of relying on existing agencies or of establishing a new one, the temporary continued assignment of this responsibility to the NIH may be desirable. This interim solution, if it is adopted, must be accompanied by additional funding to carry it out effectively.

Considerations of biohazards and physical and biological containment have necessarily had a high priority in this early phase of recombinant-DNA research. Many informed observers believe, however, that these concerns will decline in importance as research continues and experience grows. Therefore although the current furor makes a rational approach to the biohazards question an essential part of any successful recombinant-DNA policy, this approach does not exhaust the longer-term requirements and may even distort them. More crucial in the long run may be several other issues that have been raised directly or indirectly.

For example, in investing in fundamental genetic research that can profit from recombinant-DNA techniques, what relative priorities should be assigned to potential applications? In the past the national strategy in biomedical research has been to invest directly in basic research, without declared objectives, while also investing in specific objectives, allowing some of the latter support to "trickle down" to basic research. Thus an investigator of the interaction of viruses and cells, say, might be alternatively or simultaneously supported by funds for fundamental investigation and by funds intended for promoting the development of an effective therapy for cancer. What should be the priorities among possible practical applications of molecular genetics? Competing lines of inquiry include the microbiological synthesis of drugs, specific human gene therapies, the improved efficiency of photosynthesis, nitrogen fixation by food crops, enhanced agricultural production and so on. There are quite different potential risks and benefits in each of these directions, and all are unlikely to be maximally supported at once. In the new areas that are opening up is a new research strategy called for? If it is, by what procedures should it be formulated and how should it be implemented?

It is widely recognized that there is a logical continuum running from basic research through applied research and development to technological application. It is also recognized that movement along this continuum is neither smooth nor fully predictable and that varying motivations and institutional arrangements operate along its length. Recombinant-DNA techniques are the product of fundamental investigation, supported almost entirely by the partnership of the Federal Government and the universities. For the moment, at least, the techniques are likely to remain useful primarily in that area. The techniques may also be useful for various industrial purposes, however. Given the nature of the original investment as well as the complex issues raised, should technological uses, at least for a time, be kept under Federal control? Should some of the return from successful applications be employed to recycle the original investment of Federal resources? Should this promising new technology be a prototype for establishing a revolving capital fund to support a more stably financed basic-research effort?

The possibilities of genetic engineering and evolutionary control illustrate the fundamental dilemmas raised by the new capabilities conferred by scientific knowledge. Society has entered an age of intervention, in which the automatic operation of natural processes is increasingly, through informed intervention, brought consciously into the orbit of human purpose. Many events that humanity formerly could regard only as a boon or a scourge—an act of God or of nature—are now the partial product of human decision and intervention. If human beings do not have the capability today to invent new organisms or to initiate life itself, they may soon have that capability. If they cannot today consciously and fully control the behavior of large ecosystems, that power is not far beyond what has already been achieved. The humility of individuals understandably shrinks from awesome powers that were earlier assigned to divine will. It was not, however, the humility of individuals that conferred these emerging capabilities or is called on to control them today. It was the social interaction of individuals, operating through social institutions, that brought us to the present fateful decision making. Imperfect though they are, our social institutions built the platform for the age of intervention.

The policy challenge we face, refracted in the exquisite structure and potential of the double helix, is whether we can create institutions able to transform the fruits of an age of reason into the achievements of an age of intervention. There are voices today urging us not only to eschew conscious intervention but also to distrust and limit the uses and consequences of reason itself. Perhaps it needs to be restated that it was, after all, natural selection that evoked the double helix and all it conveys. Included among the products are human knowledge and judgment, to which has now passed the duty of designing social processes and structures that can cope with the manipulability of the double helix itself.

The concept and control of the double helix signal a new frontier of biocultural progression. A stereoscopic vision that includes both "creative pessimism" and "creative optimism" is now required. Neither alone can do justice to the profound revelations human beings have recently experienced. A single eye is particularly limited in yielding depth and perspective. For the age of intervention at least two are needed.

PHYSICAL CONTAINMENT	BIOLOGICAL CONTAINMENT (FOR *E. COLI* HOST SYSTEMS ONLY)		
	*EK*1	*EK*2	*EK*3
*P*1	DNA from nonpathogenic prokaryotes that naturally exchange genes with *E. coli* Plasmid or bacteriophage DNA from host cells that naturally exchange genes with *E. coli*. (If plasmid or bacteriophage genome contains harmful genes or if DNA segment is less than 99 percent pure and characterized, higher levels of containment are required.)		
*P*2	DNA from embryonic or germ-line cells of cold-blooded vertebrates DNA from other cold-blooded animals and lower eukaryotes (except insects maintained in the laboratory for fewer than 10 generations) DNA from plants (except plants containing known pathogens or producing known toxins) DNA from low-risk pathogenic prokaryotes that naturally exchange genes with *E. coli* Organelle DNA from nonprimate eukaryotes. (For organelle DNA that is less than 99 percent pure higher levels of containment are required.)	DNA from nonembryonic cold-blooded vertebrates DNA from moderate-risk pathogenic prokaryotes that naturally exchange genes with *E. coli* DNA from nonpathogenic prokaryotes that do not naturally exchange genes with *E. coli* DNA from plant viruses Organelle DNA from primates. (For organelle DNA that is less than 99 percent pure higher levels of containment are required.) Plasmid or bacteriophage DNA from host cells that do not naturally exchange genes with *E. coli*. (If there is a risk that recombinant will increase pathogenicity or ecological potential of host, higher levels of containment are required.)	
*P*3	DNA from nonpathogenic prokaryotes that do not naturally exchange genes with *E. coli* DNA from plant viruses Plasmid or bacteriophage DNA from host cells that do not naturally exchange genes with *E. coli*. (If there is a risk that recombinant will increase pathogenicity or ecological potential of host, higher levels of containment are required.)	DNA from embryonic primate-tissue or germ-line cells DNA from other mammalian cells DNA from birds DNA from embryonic, nonembryonic or germ-line vertebrate cells (if vertebrate produces a toxin) DNA from moderate-risk pathogenic prokaryotes that do not naturally exchange genes with *E. coli* DNA from animal viruses (if cloned DNA does not contain harmful genes)	DNA from nonembryonic primate tissue DNA from animal viruses (if cloned DNA contains harmful genes)
*P*4		DNA from nonembryonic primate tissue DNA from animal viruses (if cloned DNA contains harmful genes)	

"SHOTGUN" EXPERIMENTS USING *E. COLI K*-12 OR ITS DERIVATIVES AS THE HOST CELL AND PLASMIDS, BACTERIOPHAGES OR OTHER VIRUSES AS THE CLONING VECTORS

EXPERIMENTS IN WHICH PURE, CHARACTERIZED "FOREIGN" GENES CARRIED BY PLASMIDS, BACTERIOPHAGES OR OTHER VIRUSES ARE CLONED IN *E. COLI K*-12 OR ITS DERIVATIVES

SOME EXAMPLES of the physical and biological containment requirements set forth in the NIH guidelines for research involving recombinant-DNA molecules, issued in June, 1976, are given in this table. The guidelines, which replaced the partial moratorium that limited such research for the preceding two years, are based on "worst case" estimates of the potential risks associated with various classes of recombinant-DNA experiments. Certain experiments are banned, such as those involving DNA from known high-risk pathogens; other experiments, such as those involving DNA from organisms that are known to exchange genes with *E. coli* in nature, require only the safeguards of good laboratory practice (physical-containment level *P*1) and the use of the standard *K*-12 laboratory strain of *E. coli* (biological-containment level *EK*1). Between these extremes the NIH guidelines prescribe appropriate combinations of increasing physical and biological containment for increasing levels of estimated risk. (In this table containment increases from upper left to lower right.) Thus physical-containment levels *P*2, *P*3 and *P*4 correspond respectively to minimum isolation, moderate isolation and maximum isolation. Biological-containment level *EK*2 refers to the use of new "crippled" strains of *K*-12 incorporating various genetic defects designed to make the cells' survival outside of laboratory conditions essentially impossible. Level *EK*3 is reserved for an *EK*2-level host-vector system that has successfully passed additional field-testing. Because of the very limited availability of *P*4 facilities and because no bacterial host-vector system has yet been certified by the NIH as satisfying the *EK*3 criteria, the recombinant-DNA experiments now in progress in the U.S. with *E. coli* host systems are with a few exceptions limited to those in the unshaded boxes. Experiments with animal-virus host systems (currently only the polyoma and SV40 viruses) require either the *P*3 or the *P*4 level of physical containment. Experiments with plant-virus host systems have special physical-containment requirements that are analogous to the *P*1-to-*P*4 system.

The Atmosphere of Mars

It is less than a hundredth as dense as that of the earth, but it is still the principal agency altering the surface of the planet. Its winds and clouds resemble their terrestrial counterparts

by Conway B. Leovy

The brilliantly successful series of space missions to Mars, beginning in 1964 and culminating last year in the safe touchdown of the two Viking landers, have made the planet nearly as familiar as the moon. Unlike the moon, whose story appears essentially to have ended one or two billion years ago, Mars is still evolving and changing. On Mars, as on the earth, the most pervasive agent of change is the planet's atmosphere, itself the product of the sorting of the planet's initial constituents that began soon after it condensed from the primordial cloud of dust and gas that gave rise to the solar system 4.6 billion years ago.

Although some information about the atmosphere of Mars had been gleaned from telescopic observations prior to the space flights, the information was both unverifiable and subject to misinterpretation. With the much more accurate information provided by spacecraft that have flown past Mars, gone into orbit around it and finally landed on it, we now have sufficient detailed knowledge of a second planetary atmosphere to test our understanding of atmospheric evolution and atmospheric processes previously based on the only example available: that of the earth. Except for its strikingly different composition the atmosphere of Mars behaves much like a rarefied version of our own. It transports water, generates clouds and exhibits daily and seasonal wind patterns. In response to seasonal changes in solar input localized dust storms occur and sometimes grow in strength until they envelop the entire planet. Such global-scale dust storms are a uniquely Martian phenomenon.

As a result of atmospheric weathering the primitive crystalline rock on Mars has been broken down into fine particles, oxidized and combined chemically with water to produce the characteristic reddish minerals so vividly apparent in the pictures made by the Viking landers. The mechanisms of weathering on Mars are clearly different from the dominant weathering processes on the earth, which depend on liquid water. Everywhere on Mars there is evidence of wind erosion and sedimentation. For example, the two Viking landing sites appear to have been severely scoured by winds. Deep layers of wind-blown sediment have accumulated in the Martian polar regions, and dune fields larger than any on the earth have been photographed in the region surrounding the north pole.

Because of the low atmospheric pressure on Mars and the prevailing low temperature the only stable forms of water found there are water vapor and ice. In spite of the present absence of liquid water, photographs made from vehicles in orbit show vividly that the surface of the planet has experienced fluid erosion on a grand scale. The fluid was very likely water. Running water may once have existed on the surface as a result of the melting or vaporization of subsurface masses of ice. It is also possible that at one time the climate was milder and the atmosphere was sufficiently dense and moist to have produced rainfall.

The most fundamental attributes of a planetary atmosphere are its composition and its mass, or what is equivalent, its surface pressure. As recently as 1963, the year before *Mariner 4* was launched on its successful voyage to Mars, it was believed from telescopic evidence that the atmospheric surface pressure on Mars was about 85 millibars, a value somewhat less than 10 percent of the earth's average surface pressure of 1,013 millibars. It was also believed the most abundant gas in the Martian atmosphere was nitrogen. Carbon dioxide and water had been identified spectroscopically, but both were thought to be minor components because neither one was abundant enough to contribute more than a small fraction to the assumed total of 85 millibars. Although nitrogen cannot be observed spectroscopically, it seemed the most logical principal component in view of its abundance in the earth's atmosphere.

The pressure estimate of 85 millibars was based on the intensity and polarization of reflected sunlight, using certain assumptions to separate the effects of light reflected by the planet's surface from the effects of light scattered or reflected by the atmosphere. We now know that the fraction of the reflection attributed to the gaseous component of the atmosphere was much too high because of a failure to allow for the large amount of dust and haze that was subsequently shown to be present. Although the gravity at the surface of Mars is only 38 percent of that at the surface of the earth, there is no reason in principle why Mars could not hold an atmosphere with a surface pressure nearly as high as the earth's. Hydrogen and helium are the only gases whose normal kinetic energy in the form of thermal motion is more than sufficient to enable them to escape the gravitational grip of either planet.

While *Mariner 4* was en route to its destination an analysis of better spectrographic data showed that the surface pressure on Mars had to be much lower than 85 millibars. Just how low was finally established when *Mariner 4* flew past the planet in July, 1965. From the viewpoint of terrestrial observers *Mariner 4* disappeared behind Mars and re-

EARLY-MORNING ICE FOG whitens the floors of the deep depressions in the western part of the region on Mars known as Coprates Chasma. Wispy ice clouds also drift across the surrounding mesas and valleys. The daily recurrence of such fogs suggests that water vapor condenses as ice on the ground in the night and is evaporated by the morning sun. The vapor then refreezes as it rises through the cold atmosphere. The picture is a composite of three black-and-white images made through violet, green and orange filters from the *Viking 1* orbiter in July, 1976. This picture and the one on page 36 were prepared by the Image Processing Laboratory of the Jet Propulsion Laboratory of the California Institute of Technology. These and other Viking pictures were provided by Geoffrey A. Briggs of the Viking orbiter imaging team.

CRESCENT MARS was photographed by the *Viking 2* spacecraft in August, 1976, from a distance of 450,000 kilometers. The large circular feature to the south, the Argyre Basin, is covered with a thin frost extending from the winter polar cap. The linear features farther north are portions of the canyon system Valles Marineris, which stretches for 4,000 kilometers parallel to the equator. Farther west, near the morning terminator, the two dark circles are the enormous volcanic mountains Ascraeus Mons and Pavonis Mons. Extending westward from Ascraeus Mons are two dense cloud streamers that finally disappear into the terminator shadow 1,000 kilometers away. Cloud shadow, visible even at this range, indicates that streamers are about as high as the volcano, which is more than 20 kilometers above mean surface level, or more than twice the height of Everest. Cloud reveals a strong wind is blowing from the east.

appeared 54 minutes later. As the spacecraft's radio signals grazed the rim of the planet they were refracted by the atmosphere. The amount of refraction indicated a surface pressure of between five and seven millibars. The new spectrographic data had shown enough carbon dioxide to account for a pressure of that magnitude, so that carbon dioxide was clearly the major constituent of the Martian atmosphere. Nitrogen was relegated to the status of a minor constituent, if indeed it was present at all.

Nitrogen and the noble gases (helium, argon, neon, krypton and xenon) were measured by both Viking landers during their descent through the Martian atmosphere and subsequently on the surface of the planet. The analyses were performed by mass spectrometers, devices that first ionize a gas sample in a vacuum and then sort the resulting ions according to their mass. Two different spectrometers were used: one that operated in the near-vacuum of the Martian upper atmosphere as the lander was traversing the region between 200 and 120 kilometers above the surface and another, primarily intended for soil analysis, in which the vacuum was created artificially. Although no measurements were made between 120 kilometers and the surface, the composition of the atmosphere in the unsampled region can be inferred with considerable confidence.

The surface measurements show that carbon dioxide molecules make up about 96 percent of the total. The next most abundant gas is nitrogen, at 2.5 percent, followed by the most abundant isotope of argon, argon 40, at 1.5 percent. In agreement with earlier measurements made from the earth, oxygen molecules are present to the extent of only .1 percent. Krypton and xenon are also present in small but measurable amounts, probably along with traces of neon and helium.

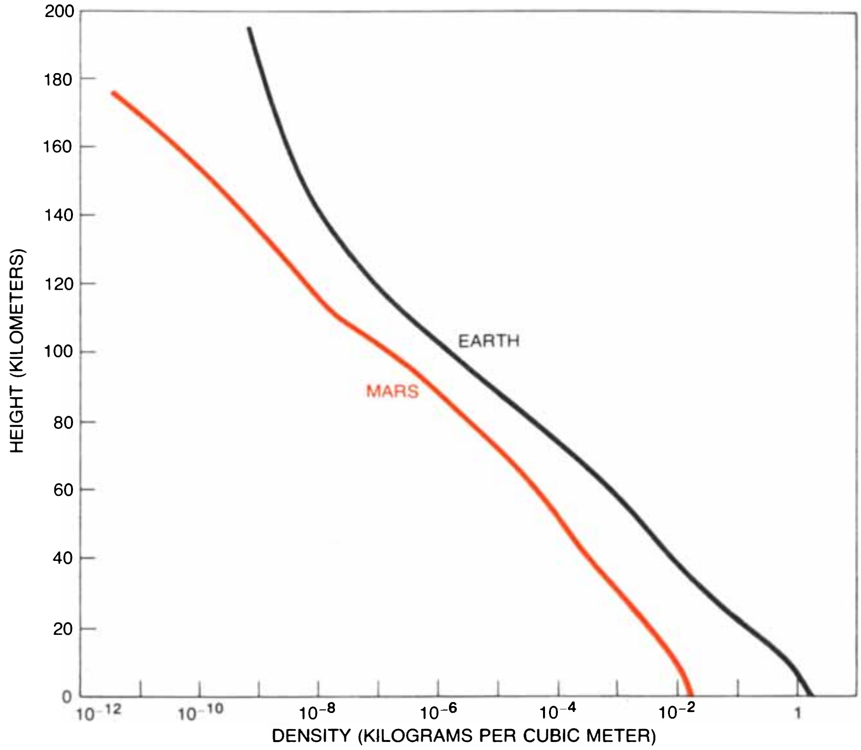
On both Mars and the earth the atmosphere consists of gases released from the interior of the planet through volcanism and less violent forms of venting. Only a portion of the vented gases, however, remains in the two atmospheres as we observe them today. On the earth most of the released volatile substances are now in other reservoirs. The reservoir for most of the vented water vapor is, of course, the ocean. Most of the outgassed carbon dioxide is now tied up in calcium carbonate rocks, such as limestones, which formed on the ocean bottom. Other reservoirs of carbon, originally outgassed as carbon dioxide, are represented by deposits of coal, petroleum and oil shale and by carbon dissolved in the ocean and dispersed throughout the biosphere. The amount of carbon stored as carbon dioxide in the earth's atmosphere is insignificant.

There may also be large nonatmospheric reservoirs of vented volatile substances on Mars. Water could be tied up in caches of polar ice or widely dispersed in permafrost. Some water is known to be bound chemically in the soil. Large quantities of carbon dioxide may be adsorbed in the soil, and it is possible that some may be frozen in the small permanent south polar cap. Because the amounts of volatiles stored in such potential reservoirs are difficult to estimate, additional clues are needed to determine what the total volume of outgassed water and carbon dioxide has been. The abundance of nitrogen and noble gases in the Martian atmosphere can provide such clues.

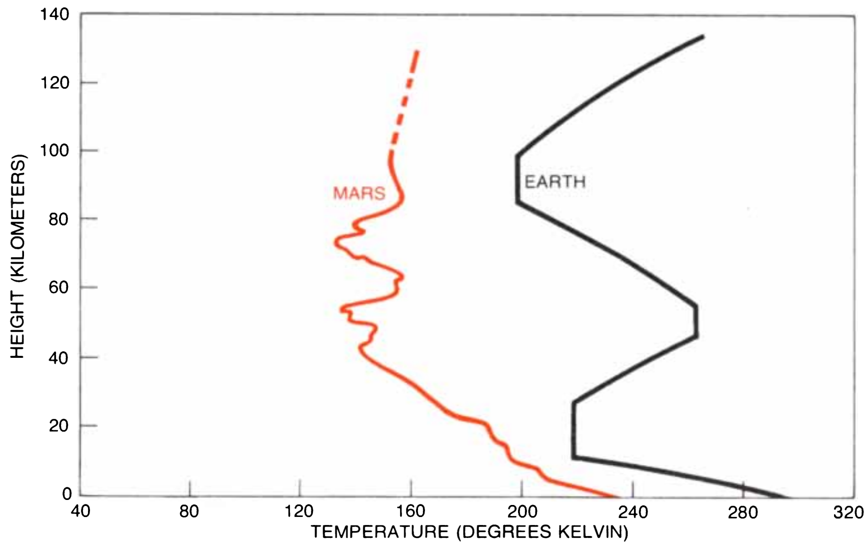
One line of argument runs as follows. Assume that the earth and Mars were originally similar in composition, at least with regard to carbon, hydrogen and the noble gases. The noble gases do not freeze out or react chemically, and with the exception of helium they do not readily escape from the top of the atmosphere. Thus the ratio of the noble gases in the atmosphere to the total amount of carbon dioxide and water outgassed over the history of the solar system should be roughly the same for Mars and the earth. This is the "earth analogue" model.

The best noble gases to use for this comparison are krypton and the two rare isotopes of argon, argon 38 and argon 36. The more abundant isotope, argon 40, is a radioactive-decay product of potassium 40 and is therefore a better indicator of the availability of potassium than of total outgassing. Viking measurements on Mars have shown that the relative abundances of the two argon isotopes and their ratios to krypton are similar to the values measured on the earth, which is consistent with the earth-analogue model. It is risky to extrapolate the model to carbon dioxide, but if one does, one finds that the total mass of carbon dioxide that has been outgassed on Mars is only about 10 times the amount now in the atmosphere. This is a surprisingly modest quantity, and if it is even approximately correct, it indicates that the total amount of gas vented on Mars has been less than that vented on the earth by a factor of many hundreds.

A completely different way to estimate the total volume of outgassing on Mars is to examine the isotopic abundances of nitrogen, oxygen and carbon in the atmosphere to see how much the heavier isotopes of those elements have been enriched with respect to the lighter ones, which have a greater probability of escaping from the top of the atmosphere. Although neither oxygen nor nitrogen possesses enough kinetic energy in the form of thermal motion to escape from the atmosphere, both molecules can escape slowly as a consequence of being ionized by ultraviolet radiation in the upper atmosphere. When the ionized



DENSITY OF THE MARTIAN ATMOSPHERE is only about a hundredth the density of the earth's atmosphere at ground level. Up to a height of about 100 kilometers the density of the earth's atmosphere decreases more rapidly than the density of Mars's. Above 100 kilometers the situation reverses. The density profile of the Martian atmosphere below 100 kilometers is derived from an analysis of the performance of the *Viking 2* lander as it traversed the atmosphere on its way to the surface. Analysis was done by Alvin Seiff and Donn B. Kirk of Ames Research Center of National Aeronautics and Space Administration. Above 100 kilometers atmospheric density is inferred from an analysis of data supplied by entry mass spectrometer of *Viking 2* lander. The analysis was done by Alfred O. C. Nier of University of Minnesota.



TEMPERATURE OF THE MARTIAN ATMOSPHERE departs markedly from the pattern in the earth's atmosphere. The one common characteristic of the two atmospheres is a steady drop in temperature up to a height of about 10 kilometers. Below that height both atmospheres are warmed by absorbing energy from the ground, which is directly heated by the sun. The temperature peak in the earth's atmosphere near 50 kilometers is caused by the absorption of ultraviolet radiation by ozone. There is no analogous ozone layer in the Martian atmosphere to absorb energy. The earth's atmosphere above 100 kilometers is also heated by ultraviolet radiation. Comparable heating is again absent in the atmosphere of Mars. Irregularities in the temperature profile of the Martian atmosphere between 50 and 100 kilometers are probably due to the expansion cooling of layers of rising gas and the compression heating of alternate layers of sinking gas. Such layers may be produced by a global system of tidal winds. Martian temperature near 130 kilometers was obtained from a Viking entry experiment analyzed by William B. Hanson of University of Texas. Remaining temperatures are from analyses by Seiff and Kirk.

molecules and electrons recombine, the energy released is more than enough to dissociate the molecules into their constituent atoms. The energy left over imparts enough kinetic energy to the atoms of oxygen and nitrogen so that both gases can slowly escape. The same process does not operate on the earth because of the earth's stronger gravity.

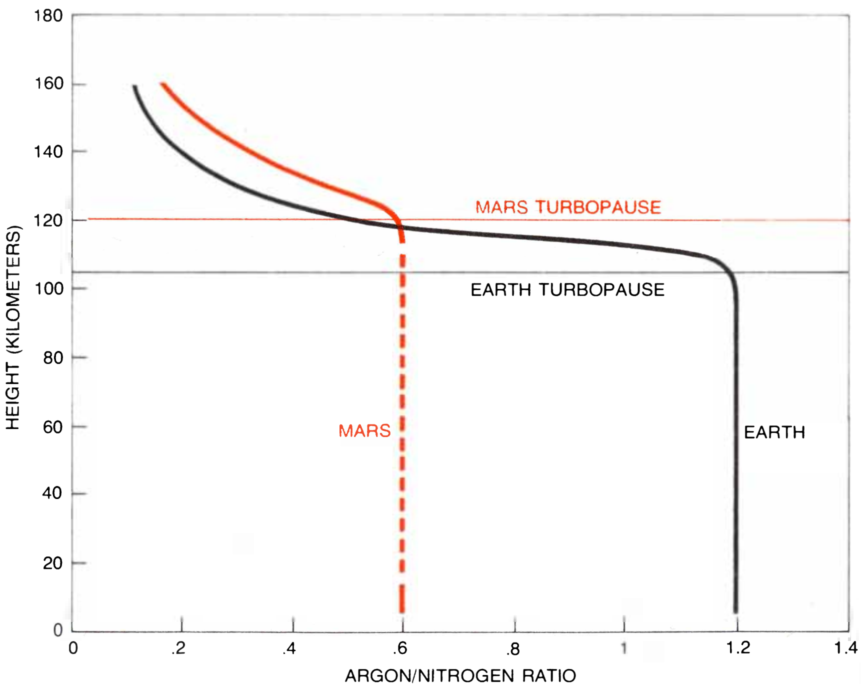
The Viking measurements show that the various isotope abundances remain constant up to 120 kilometers. Above that altitude the ratios begin to change because turbulence is insufficient to keep the gases well mixed. In the ultrarefied atmosphere above 120 kilometers the concentrations of the various gases are controlled by molecular diffusion, a process that enables each gas to distribute itself according to its molecular mass, the lighter gases rising to the top like cream and ultimately escaping into space. The zone that separates the well-mixed lower region of the atmosphere from the diffusion region is called the turbopause. On Mars the turbopause is at about 120 kilometers, where the density of the atmosphere is only about a hundredth the density at the terrestrial turbopause.

Outward escape from a planetary atmosphere is possible only for molecules above the turbopause, because it is only there they can escape before colliding with other molecules. Since lighter isotopes are relatively more abundant above the turbopause, they will escape more rapidly than their heavier counterparts. If continued long enough, this process will lead to selective enrichment of the heavy isotopes with respect to the light ones throughout the atmosphere. The Viking measurements show that the ratio of the rare heavy isotope of nitrogen, nitrogen 15, to the common isotope, nitrogen 14, is larger than the corresponding terrestrial ratio by about 60 percent. The degree of enrichment depends on the total amount of nitrogen vented, when it was vented and the height of the turbopause. The higher the turbopause, the less the enrichment of nitrogen 14 with respect to nitrogen 15 in the region from which escape can occur. Enrichment also depends on such external factors as the flux of solar ultraviolet energy, which facilitates the escape process by imparting the necessary energy of ionization. Models of the evolution of the Martian atmosphere based on the ratio of nitrogen isotopes suggest that the outgassing of nitrogen began early in the planet's history and that the total volume vented has been large, perhaps 100 times the amount now present in the atmosphere.

According to the Viking measurements, there has been no comparable enrichment in the heavier isotopes of carbon and oxygen. How can that be explained? One proposal is that the atmospheric carbon and oxygen, unlike the atmospheric nitrogen, are in contact with large surface or subsurface reservoirs of carbon- and oxygen-containing materials, so that there can be atomic exchanges between the atmosphere and the reservoir. The amounts of carbon and oxygen in the atmosphere may represent only the tip of the iceberg. Thus the absence of observable enrichment in the heavier isotopes of carbon and oxygen can be used as an argument for a large volume of outgassing having created a large reservoir of carbon dioxide. It seems safe to conclude that the outgassing of Mars, although it was far less efficient than the outgassing of the earth, produced substantially more carbon dioxide and water vapor than is now found in the atmosphere.

Unlike nitrogen and the noble gases, water vapor is a highly variable component of the Martian atmosphere. Earth-based measurements had shown a seasonal cycle and a suggestion of a daily one. Spectrometers aboard the Viking orbiters have now shown that the amount of water vapor varies strongly with latitude. Although the vertical distribution of water vapor is difficult to establish spectroscopically, the total mass of water per unit of area of a column extending from the ground to the top of the atmosphere is readily calculated. By converting the amount of water vapor in such a column to its liquid equivalent one obtains the number of "precipitable centimeters": the depth of the liquid if it were precipitated.

The values measured so far by the Viking orbiters range from less than 10^{-4} precipitable centimeter in high southern latitudes in midwinter to 10^{-2} precipitable centimeter in high northern latitudes in midsummer. The earth's atmosphere normally contains two or three precipitable centimeters, so that one's first impression is that the atmosphere of Mars is exceedingly dry. A fairer comparison, however, is one made with the total amount of water in the earth's atmosphere above the level at which its pressure matches the surface pressure on Mars: seven millibars. By this criterion the atmosphere of Mars is very wet; even the lowest value measured on Mars is several times larger than the water content of the earth's atmosphere. A more relevant measure is the relative humidity, or degree of saturation of the atmosphere. On the average the relative humidity of the Martian atmosphere is so high that the abundance of water vapor seems to be controlled largely by saturation. The atmosphere is about as wet as it can be for the prevailing atmospheric temperature.



RATIO OF ARGON TO NITROGEN in the atmospheres of Mars and the earth declines in parallel above about 120 kilometers but exhibits sharply different values below the respective turbopauses of the two planets. The turbopause marks the transition zone between the lower region of the atmosphere, where winds keep the component gases well mixed, and the low-density upper region, where components can become separated by molecular diffusion. Above the turbopause the argon/nitrogen ratio falls sharply as nitrogen, the lighter of the two gases, increases in relative abundance prior to its slow escape from Mars. Martian profile above 110 kilometers is derived from Viking entry experiments. Values near Martian surface are from mass spectrometer on *Viking 2* lander in an experiment by Klaus Biemann of Massachusetts Institute of Technology and Tobias C. Owen of State University of New York at Stony Brook.

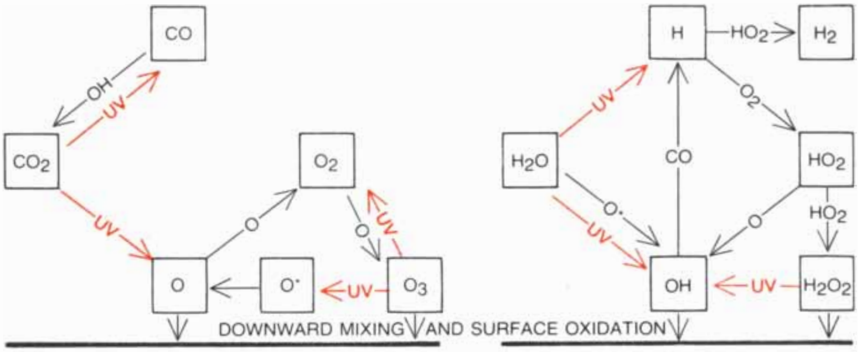
Factors other than saturation, however, are clearly needed to maintain the observed variation in water-vapor content with latitude. One important factor is the distribution of water-vapor sources and sinks. In late summer at each pole there is a small residual ice cap, and at the north pole, at least, the cap is known to be water ice. Sublimation of the ice under the summer sun would provide an abundant source of water vapor. Not surprisingly, the largest amounts of vapor observed by Viking are around the edges of the north polar cap. Curiously, however, no similar enhancement was detected around the residual midsummer south polar cap by the *Mariner 9* orbiter. If the Viking orbiters succeed in gathering data for a full annual cycle, the mystery may be cleared up

Water vapor is the key ingredient in the chemistry of the Martian atmosphere. Unlike the earth, which is shielded from solar ultraviolet radiation by a dense layer of ozone, Mars is exposed to this energetic radiation right down to the surface. As a result throughout the atmosphere carbon dioxide is dissociated into carbon monoxide (CO) and atomic oxygen (O), and water vapor is dissociated into atomic hydrogen (H) and the hydroxyl radical (OH). The H and OH are extremely reactive. For example, they can catalyze the recombination of CO and O back to carbon dioxide, which helps to explain why the atmosphere consists mainly of carbon dioxide. Atomic hydrogen and molecular oxygen combine through a series of reactions to produce hydrogen peroxide (H_2O_2), a powerful oxidizer, which may have an important influence on the chemistry of the soil.

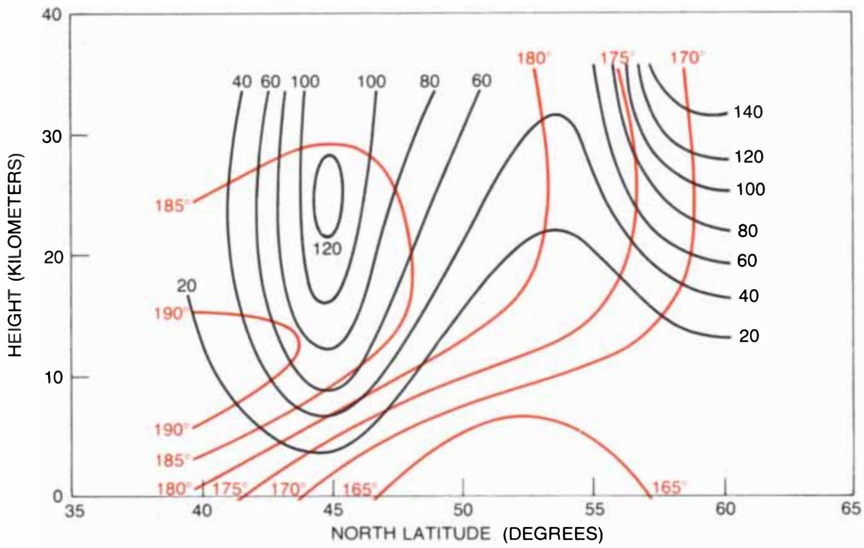
A small amount of molecular hydrogen (H_2) is produced by the dissociation of water vapor. Being quite stable, H_2 is well mixed below the turbopause. Above the turbopause it dissociates into atomic hydrogen, which slowly escapes. Extrapolated over the lifetime of Mars (4.6 billion years) the current estimated rate of escape could account for the loss of a layer of liquid water less than a meter deep. That is probably much less than the total amount of water vented from the planet.

Although the four hydrogen-containing species, H, H_2, OH and H_2O_2, are chemically important, they are below the threshold of detection by the Viking instruments. The atmospheric chemical events I have described are inferred from terrestrial experience. Ultraviolet spectrometers on the Mariner spacecraft, however, have detected the atomic hydrogen that is escaping from the upper atmosphere of Mars. These observations lend support to the presumed dissociation of molecular hydrogen above the turbopause.

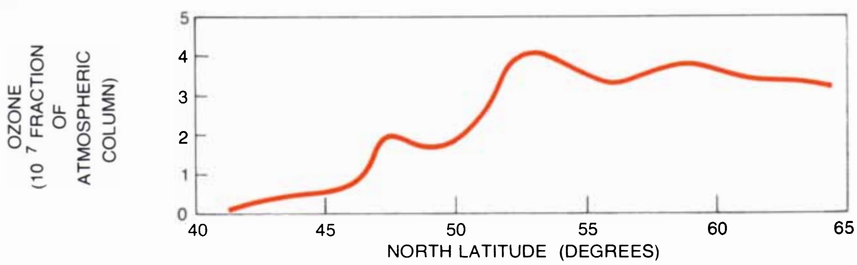
Stronger support for the inferred



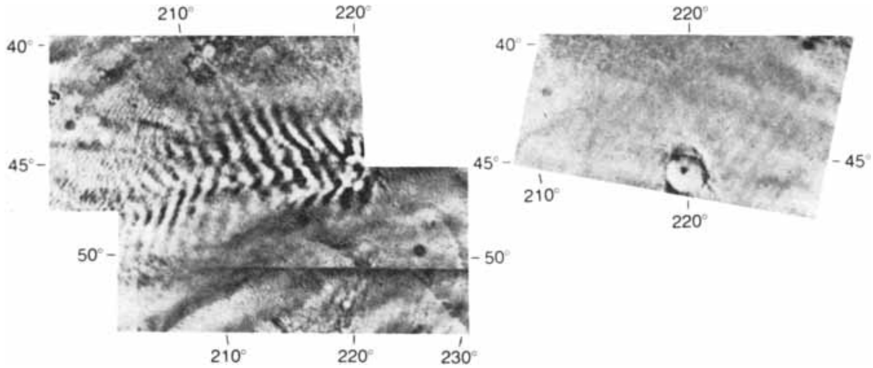
ATMOSPHERIC CHEMICAL REACTIONS are initiated when carbon dioxide (CO_2) and water vapor (H_2O) are dissociated by ultraviolet radiation (UV). In the oxygen cycle (*left*) CO_2 breaks up into carbon monoxide (CO) and atomic oxygen (O). The latter recombines to form O_2 and ozone (O_3). Both O and O_3 contribute to the oxidation of rocks and soil. In the hydrogen cycle (*right*) H_2O dissociates into atomic hydrogen (H) and the hydroxyl radical (OH). Reaction of H with O_2 produces the perhydroxyl radical (HO_2), which can react with itself to form hydrogen peroxide (H_2O_2). Both H_2O_2 and OH can contribute to surface oxidation. The two cycles interact mainly through the reactions of CO with OH, O_2 with H, and O with HO_2. Excited atomic oxygen (O^*), formed by ultraviolet dissociation of O_3, can also break H_2O up into OH radicals. Molecular hydrogen (H_2) is also formed in the hydrogen cycle. Both H_2 and O_2 are slowly removed from the atmosphere by upward mixing and eventual escape. Quantitative aspects of the cycles depicted here have been worked out by Michael B. McElroy of Harvard University and Thomas M. Donahue of University of Michigan and independently by Truman D. Parkinson and Donald M. Hunten of the Kitt Peak National Observatory.



TEMPERATURE OF WINTERTIME MARTIAN ATMOSPHERE (*color*), plotted in a meridional plane, has been derived from infrared measurements made by *Mariner 9*. The temperatures, given in degrees Kelvin, can be used to infer the distribution of the east-to-west component of the winds (*black*), given in meters per second. (One hundred meters per second equals 224 miles per hour.) Wind velocities shown are in addition to surface winds, which are typically weak in winter. *Mariner 9* infrared experiment from which temperatures were derived was by Rudolf A. Hanel, B. J. Conrath and co-workers at Goddard Space Flight Center of NASA.



WINTERTIME OZONE DISTRIBUTION is plotted for the same plane as that used for temperatures and wind velocities in middle of page. Vertical axis shows the number of ozone molecules as a fraction of all molecules in a vertical column of atmosphere. Ozone increases toward pole, where temperatures and water-vapor concentrations are low. Values were obtained by Charles A. Barth and collaborators at University of Colorado from *Mariner 9* ultraviolet data.



WAVE CLOUDS IN THE ATMOSPHERE OF MARS in midwinter are clearly visible in the mosaic of three pictures at the left, made by *Mariner 9* in 1972. The long train of clouds almost obscures the large crater over which the clouds have formed. The individual wave clouds are about 20 kilometers apart, or about half the diameter of the crater. The cloud-free crater can be seen at the lower edge of the single picture at the right, which was made just a day earlier. Latitudes and longitudes are given along the borders of the mosaics. In order to keep the apparent relief of the crater and the other surface features from reversing, the pictures are reproduced with north at the bottom, so that the sun illuminates the surface features from the top.

WAVE CLOUDS IN THE LEE OF A MARTIAN CRATER were photographed by *Mariner 9* in 1972. The broad, diffuse wave cloud that appears at the upper right is similar to the clouds visible above the 40-kilometer crater in the low-resolution mosaics at the top of the page. The clouds are probably water ice. The crater that gives rise to this cloud is only about seven kilometers in diameter. To the lower left and downwind from crater are cloud streamers whose sharp edges and small-scale structure suggest they are composed not of water but of solid carbon dioxide, which can freeze out of the atmosphere rapidly over polar regions in winter.

chemistry is provided by the observed distribution of ozone (O_3) in the atmosphere of Mars. Although the ozone is much less abundant than it is in the atmosphere of the earth, there are readily detectable concentrations of it over the Martian polar regions in winter. At lower and warmer latitudes it is usually undetectable. Ozone is produced by the combination of O and O_2, which have been generated by the decomposition of carbon dioxide. Ozone and its precursor, atomic oxygen, are destroyed by catalytic reactions that depend on H and OH. Thus the ozone concentration will be low wherever H and OH are in good supply, which they are wherever the source material, water vapor, is abundant. Since water vapor is virtually absent around the winter pole, ozone is plentiful there. The limitation the dissociation of water vapor places on ozone concentration plays a similar role in the earth's atmosphere. Above an altitude of about 40 kilometers the ozone concentration is limited by the same catalytic reactions with H and OH. Hence the Martian atmosphere resembles the earth's stratosphere in chemistry as well as in pressure.

Since the Martian atmosphere is close to saturation, one would expect ice clouds to be plentiful, as indeed they are, at least during the northern summer season the Viking vehicles have examined so far. As one might also expect, the clouds of water ice are thin, with diffuse edges. There is nothing resembling the dense, sharp-edged cumulus clouds we know on the earth. The Martian clouds are of four general types: convective clouds, wave clouds, orographic clouds and fogs.

Convective clouds are formed as gas near the surface is heated during the day, rises and cools by expansion. When the temperature falls to the saturation point, clouds form in distinct puffs of nearly uniform size and spacing. Such cloud patterns are frequent at about noon over equatorial upland regions. The spacing of the puffs and the cloud heights estimated from the locations of their shadows suggest that active convection is taking place in a layer from five to eight kilometers deep, which is somewhat greater than the depth of the diurnally heated layer over continental areas on the earth. The powerful daily stirring by convection ensures that all the atmospheric constituents of the layer can interact efficiently with the Martian surface.

The best pictures of wave clouds were made by *Mariner 9*, which examined the northern hemisphere in winter. Wave clouds form when a strong wind blows steadily across an obstacle such as a high ridge. Trains of atmospheric waves resembling water waves form in the lee of the obstacle. If moisture and temperature conditions are right, the crests of the waves are marked by clouds that appear to be stationary.

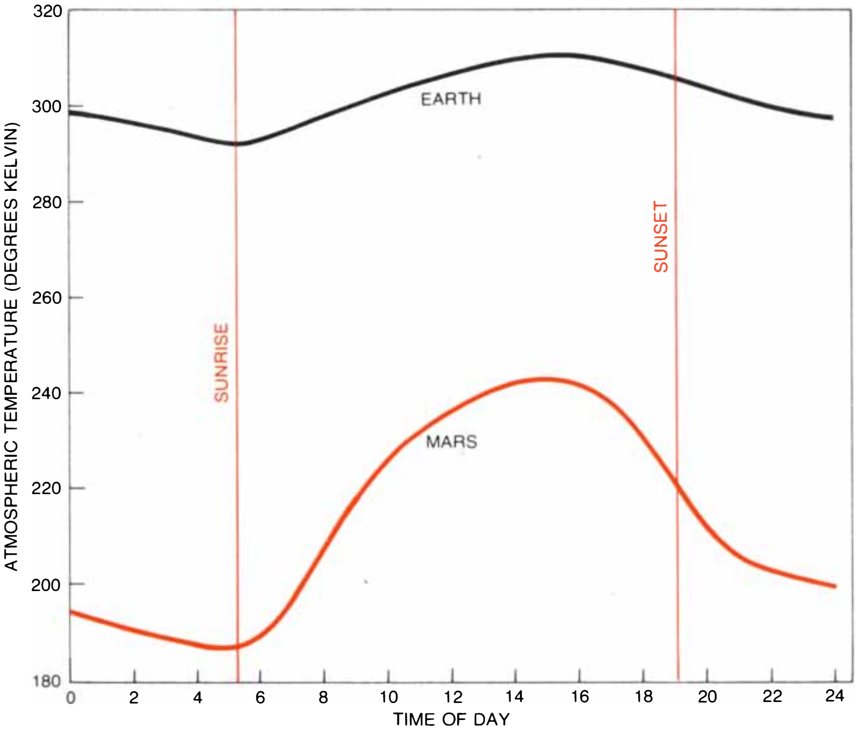
The wave clouds on Mars show that in winter the meteorology of the middle-latitude regions bears a striking resemblance to that of similar regions on the earth. The prevailing winds are westerly at all levels up to a considerable altitude, and the strongest winds, like terrestrial jet streams, are found several kilometers above the surface. By combining remotely sensed temperature information with wave-cloud photographs one can reconstruct a three-dimensional picture of the wind distribution on particular days. Typically the wind blows from the west with speeds of 10 to 20 meters per second near the surface to more than 100 meters per second (224 miles per hour) near the level of the jet stream, above 10 kilometers. These speeds are about twice as great as those that are found at the corresponding levels on the earth.

Orographic clouds form when atmospheric gas is forced to rise slowly and steadily as it moves up the slope of a large upland. Such clouds are common over the higher volcanic regions of Mars. In the northern summer season, when the atmosphere is moist, large areas are covered with rather uniform thin clouds, which can be observed from the earth. For days on end their pattern of growth and decay is repeated with remarkable precision, indicating the regularity of the day-to-day weather.

One of the most surprising atmospheric observations made by the Viking orbiters was early-morning ground fogs in several low-lying areas. The fogs are evidently created when ground frost is converted into vapor by the morning sun. The vapor condenses again as it rises through the chill atmosphere of the early morning. Since the fogs are seen day after day in the same locations, it seems likely that the atmospheric water vapor forms a new frost on the surface every night.

Water is not the only substance capable of forming clouds on Mars. In the winter polar regions and at high altitudes the temperature can fall low enough for carbon dioxide to condense, creating clouds of dry ice. Since the condensation and sublimation of carbon dioxide can be rapid, dry-ice clouds, unlike water-ice clouds, can have sharp edges. Several bright, sharp-edged clouds photographed by Mariner and Viking spacecraft are suspected of being dry-ice clouds. It is also possible that dry-ice snowstorms may occur in winter and help to create the seasonal dry-ice polar caps. Much of the dry ice in the caps, however, is probably deposited when carbon dioxide gas comes in direct contact with the frigid Martian soil and condenses.

About 20 percent of the carbon dioxide in the Martian atmosphere is cycled between the atmosphere and the winter polar cap each season, causing a corresponding variation in the atmospheric pressure. The pressure variation, which is felt throughout the atmosphere, not just in the polar-cap regions, has been measured by the barometers on both Viking landers. The magnitude of the pressure change agrees remarkably well with pre-Viking theoretical models of the carbon dioxide exchange between the polar caps and the atmosphere. The measurements show that any other large reservoirs of carbon dioxide, for example carbon dioxide adsorbed on particles of Martian soil, do not exchange seasonally with the atmosphere. There may, however, be a large reservoir of adsorbed carbon dioxide capable of exchanging with the atmosphere on a much greater time scale.



DAILY VARIATIONS IN ATMOSPHERIC TEMPERATURE at the *Viking 1* landing site (*color*) are qualitatively similar to those at China Lake, Calif., a desert site (*black*). In both cases the temperature touches a minimum around sunrise and reaches a peak about 10 hours later. The daily range, however, is about three times greater on Mars than it is on the earth. At Viking site range is 55 degrees, from about 187 to 242 degrees Kelvin (−86 to −31 degrees Celsius). At China Lake range is 18 degrees, from 292 to 310 degrees K. (19 to 37 degrees C.).

Winds on Mars have long been a subject of speculation. In recent years serious efforts have been made to develop theoretical models of them with techniques similar to those applied to the forecasting of terrestrial weather and climate. The techniques should be applicable since both planets have a similar sequence of seasons (because their polar axes are tipped almost identically) and a nearly identical daily rotation rate (24 hours 37 minutes for Mars). The rotation rate is important because it determines the magnitude of the deflection of moving masses of atmosphere. The deflection, or Coriolis force, is responsible for the counterclockwise rotation of winds around low-pressure areas in the northern hemisphere and the clockwise rotation around low-pressure areas in the southern hemisphere.

The picture that emerges from the theoretical investigations is a planetary wind pattern with two major regimes. One is a middle-latitude winter regime similar to such a regime on the earth, with prevailing westerlies, high-altitude jet streams and traveling disturbances, or storms. The other is an equatorial summer regime free of traveling disturbances, in fact free of all day-to-day variations except the gradual ones associated with the slow march of the seasons. Under the second regime the main influence on the wind flow is the daily variation in solar heating and its interaction with local topography. Thus the heated gas should move up regional slopes during the day and down again during the night. Since there are terrain slopes almost everywhere on Mars, slope winds, deflected by planetary rotation, should be widespread.

In addition the daily heating cycle should create atmospheric tides, or planetwide oscillations in the wind pattern. Since Mars is a desert planet, the tides should be stronger than those on the earth. Near the surface the tidal winds are weak: on the earth they attain at most one meter a second; during the northern hemisphere summer season on Mars they should reach perhaps five or 10 times that velocity. The velocities increase sharply, however, with altitude. On the earth they reach about 50 meters per second between altitudes of 80 and 100 kilometers. Although the corresponding velocities on Mars have not been measured, one can infer from the height of the turbopause that they are higher. The tidal winds are effective stirring agents in the upper atmosphere, and so they probably control the height of the turbopause on both planets. Since the Martian turbopause is at a lower density level than the turbopause on the earth, as a consequence of more efficient stirring by winds, it seems likely that the atmospheric tides on Mars are stronger than those on the earth.

Do Viking observations provide any other support for this conjecture? Perhaps the best evidence for the strength of the atmospheric tides on Mars is supplied by barometer readings at the two Viking landing sites. The instruments show large daily variations in atmospheric pressure, a clear indication of strong atmospheric tides. Supporting evidence was provided by measurements of vertical temperature profiles made by the entry vehicles. At higher levels, where the tidal winds should be strong, there should be large temperature excursions caused by alternating layers of atmospheric gas heated by compression and gas cooled by expansion. Theory predicts that the excursions will be about 20 kilometers apart and that their amplitude will increase with height.

The temperature profiles recorded during the entry of the Viking lander vehicles show temperature excursions with just these properties. Hence it does appear that on Mars, as on the earth, the height of the turbopause is set by the strength of the atmospheric tides. The turbopause in turn acts as a kind of valve regulating the escape of light gases. Because the Martian turbopause is relatively high the volume of light gases available for escape is correspondingly low, and it is partly for this reason that the current escape rate of hydrogen on Mars is so low.

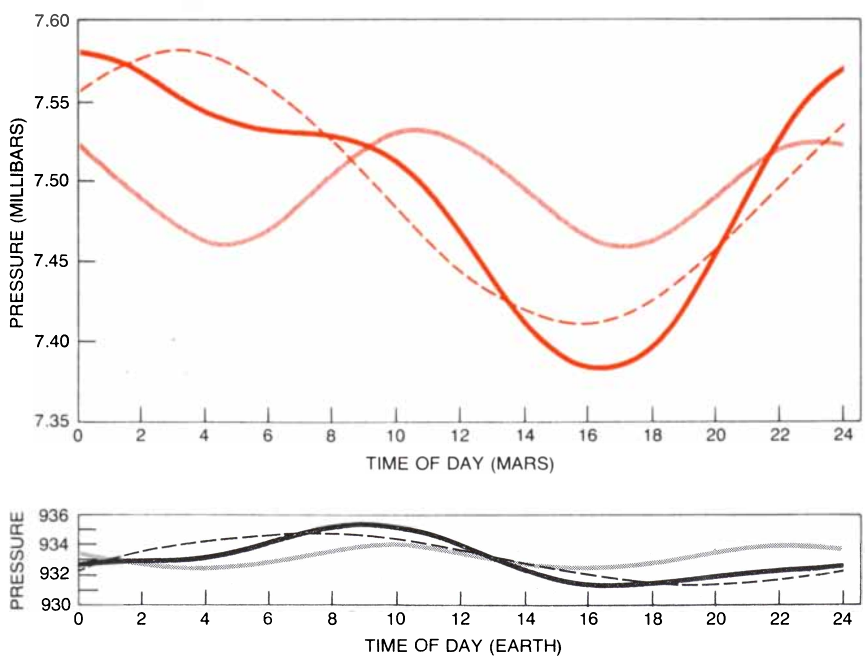
The great dust storms of Mars were observed long before the age of space exploration. Because of the low density of the Martian atmosphere much stronger winds are needed to raise dust storms on Mars than are needed on the earth. A wind velocity as low as six or seven meters per second (13 to 16 miles per hour) can begin to raise sand grains in many terrestrial deserts. Laboratory studies show that winds of at least 30 to 60 meters per second (65 to 135 miles per hour) are required on Mars. Since there is no rain on Mars to clear the air, small dust particles can remain suspended for weeks or months.

Winds strong enough to cause dust storms develop in winter in the snow-free belt adjacent to the edge of the Martian polar cap. Several dust storms were photographed in that region by both *Mariner 9* and the Viking spacecraft. Far more impressive are the global dust storms, one of which totally blanketed the planet when *Mariner 9* arrived and went into orbit in November, 1971. Prior to the Viking missions the global storms were believed to occur regularly, once each Martian year, at about the time the planet makes its closest approach to the sun. Very recently, however, the Viking spacecraft have detected dust storms on a nearly global scale at a much earlier Martian season. Evidently the very large storms are commoner than has been thought. Beginning with a small-scale storm (or several such storms) in the southern tropics, the global storms spread rapidly, so that within a few days they cover most of the planet.

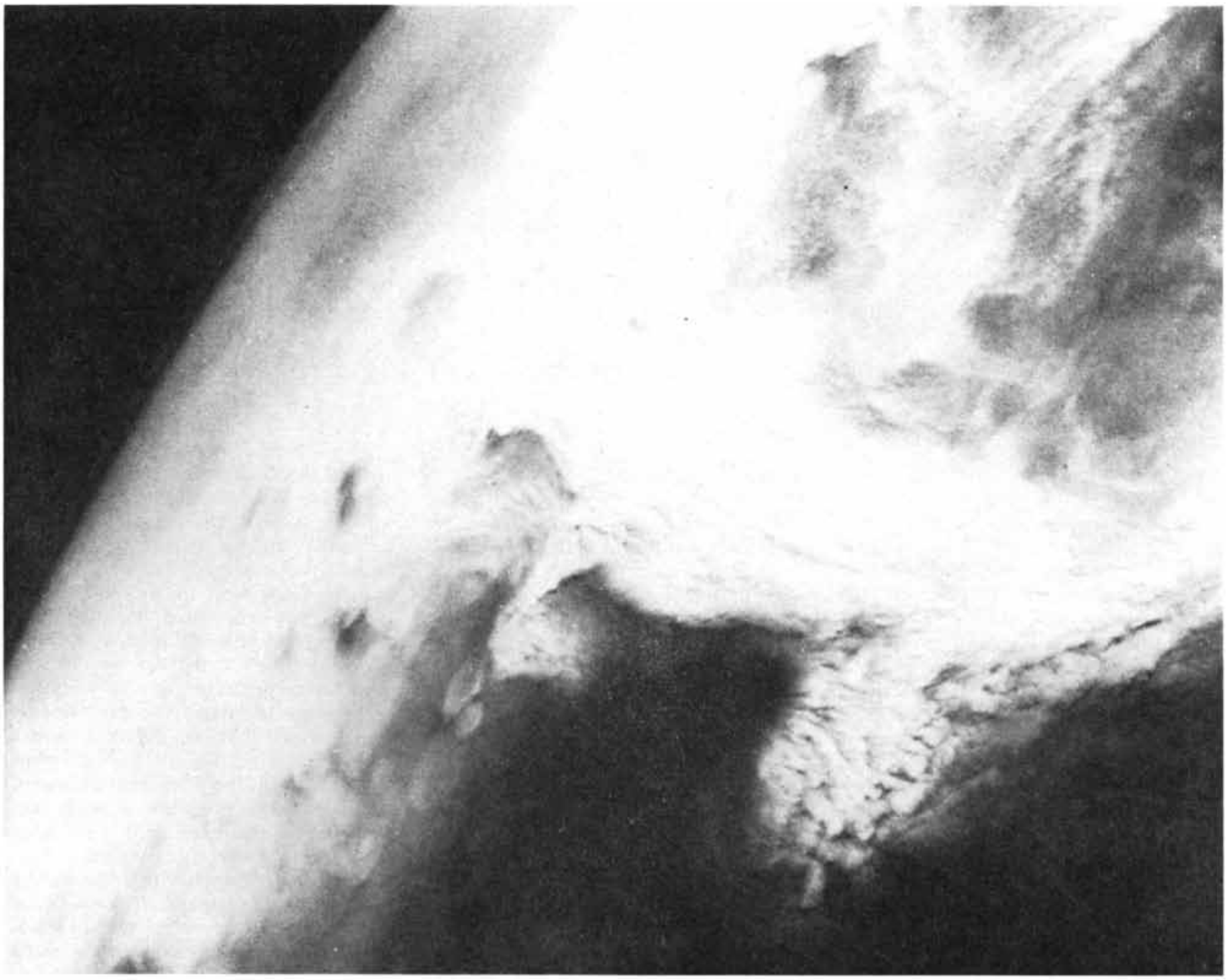
What can generate the strong winds of such huge storms? It must be significant that Mars's orbit is considerably more elliptical than the earth's. The amount of solar energy impinging on Mars is about 40 percent greater when the planet is closest to the sun than it is when the planet is farthest away. The corresponding figure for the earth is only 3 percent. This fact alone, however, cannot account for the wind velocities required.

A number of explanations have been advanced, but I shall discuss only the one I find most plausible. The strength of the global tidal winds is very sensitive to the heating of the atmosphere. That heating in turn depends not only on the amount of solar energy reaching the Martian surface but also on the dustiness of the Martian atmosphere: the dust particles absorb sunlight and can heat the surrounding gas directly and efficiently. If the atmosphere is sufficiently laden with dust, the global tidal winds can grow in intensity until they are nearly strong enough to raise dust by themselves. Presumably this self-regenerative process begins in localized areas, where the tidal winds and local topographic winds combine to produce velocities sufficiently high to put a critical amount of dust into the local atmosphere. As the dust is spread by the winds it augments the heating of the atmosphere over a large area, giving rise to tidal winds strong enough to raise dust over still wider areas.

It seems clear that suspended dust influences the behavior of the atmosphere on Mars beyond anything known on the earth. Even during the season of the recent Viking landings, a season in which the atmospheric dust load was fairly light, the sky had a pinkish-yellow cast



DAILY VARIATIONS IN ATMOSPHERIC PRESSURE, in relation to total pressure, are about five times greater at the *Viking 1* landing site than they are at China Lake on the earth. The solid curve in color at the top shows that in the course of a Martian day the atmospheric pressure varies by .2 millibar, which is about 2.5 percent of the total mean pressure of 7.5 millibars. The curve can be decomposed into a daily harmonic (*broken curve*) and a semidaily harmonic (*light-color curve*). The black curve below it is the daily pressure variation at China Lake, which is similarly decomposed into two harmonics. To match the daily variation on Mars a barometer on the earth would have to vary by some 23 millibars in the course of a day. Actual range at China Lake is only 4.5 millibars. Large daily fluctuations on Mars indicate that atmospheric tides and accompanying wind systems are much more intense there than on the earth.



VAST MARTIAN DUST STORM was photographed by the *Viking 2* orbiter in February of this year from a point 33,000 kilometers above the surface of the planet and at 51 degrees south latitude with a camera pointed to the west-northwest. The center of the image is at 42.5 degrees south latitude and 108 degrees west longitude, which places the storm in the region near Claritas Fossae and Thaumasia Fossae between 20 and 45 degrees south latitude. Within a few days the storm had spread over most of the southern hemisphere and then lasted for several weeks. When the storm began, it was early spring in the Martian southern hemisphere. Before these observations were made by the Viking spacecraft it was believed such near-global dust storms were limited to the season near the southern summer solstice.

in pictures taken from the landers—evidence of a substantial amount of light-scattering dust in the atmosphere. It also seems necessary to invoke the solar heating of a dust-laden atmosphere to account for daily variations in surface pressure as large as those measured at the landing sites. Thus there seems to be an efficient positive-feedback system whose components are winds, dust and atmospheric heating.

Although uncertainties remain, the present picture of processes operating in the atmosphere of Mars is remarkably detailed in spite of the brief history of the closeup exploration of the planet. The Mariner and Viking pictures and measurements have disclosed many striking similarities between the Martian atmosphere and the one we know on the earth. Chemistry, dynamical processes and atmospheric evolution are linked closely on both planets. The effort to understand these interacting processes on another planet is helping to add perspective and breadth to studies of our own planet.

Can we yet say anything about the Martian climate in the past? In particular, do the great channels and other apparent manifestations of erosion by water indicate that Mars once had an extended period of warm and wet weather? I believe the weight of the evidence makes it unlikely. It appears that the total amount of outgassing on Mars has been far less than that on the earth. Even if the total vented mass were 100 times the present atmospheric mass or more, the availability of water vapor in the atmosphere may not have been much greater then than it is now. The limitation on the amount of water vapor in the Martian atmosphere is the low temperature. Even an atmosphere approaching the density of the earth's composed chiefly of carbon dioxide and nitrogen could not have been much warmer than the present one, and so it could not have held much more water vapor.

Conceivably the sun itself was hotter when water eroded the features we now see on Mars, but it seems more likely that these features resulted from cataclysmic events such as meteorite impacts or volcanic explosions. The Martian crust has probably contained large amounts of water, either frozen or chemically bound, for a long time, so that any sudden heating could have released local floods. Such a cataclysm could also have produced brief, intense local storms in which the sudden vaporization of water was quickly followed by torrential rains. There may have been similar cataclysmic floods in very early stages of the earth's history, before there were oceans. It appears that more definite answers to the questions of past Martian climates and the origin of the channels will have to await the next stage in the exploration of Mars.

Viral Hepatitis

Much has been learned in recent years about the epidemiology and immunology of this debilitating disease. Soon a vaccine should be available that will offer protection against one form, hepatitis B

by Joseph L. Melnick, Gordon R. Dreesman and F. Blaine Hollinger

The primary requirement for a vaccine against a viral disease has always been a large supply of the virus, and in this century the crucial task for investigators trying to develop a vaccine has been to discover a good system, such as eggs or cell cultures, for propagating the virus. It is therefore remarkable that vaccines are now being tested that may bring under control a serious disease, hepatitis *B*, even though the virus that causes it has never been grown in the laboratory. The vaccines are based on components of what appears to be an envelope, or external shell, of the virus. Neither the precise function of these components nor even when and how they become associated with the virus is known, but they do appear in large numbers in the blood of hepatitis *B* patients and people who are unwitting carriers of the disease. By stimulating the production of highly specific antibodies they also confer immunity against the infectious virus itself. In the general account of viral hepatitis that follows we shall concentrate on tracing the pathways that have brought us close to having a vaccine against hepatitis *B*.

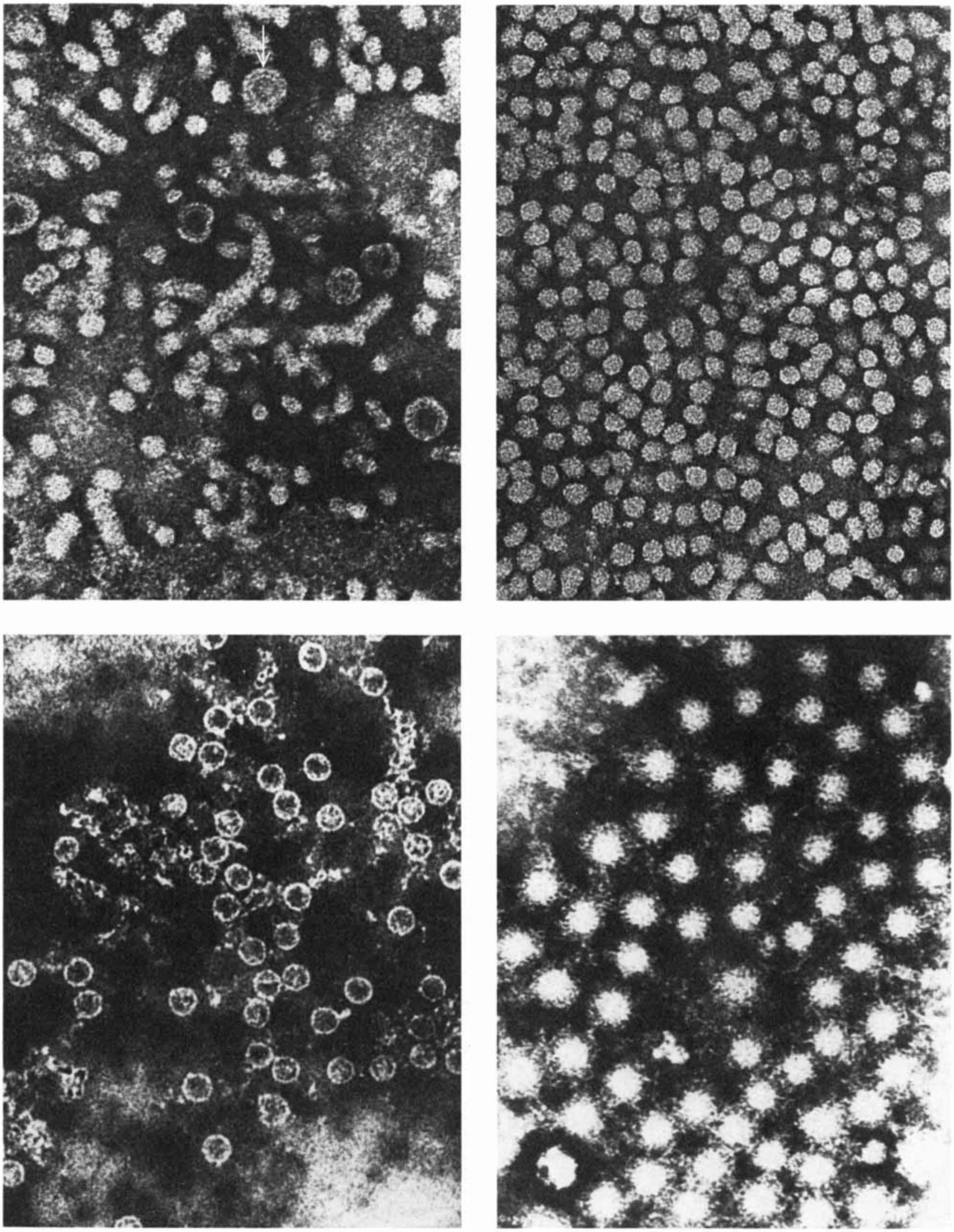
Viral hepatitis is a term that is generally reserved for acute infections of the liver caused by one of at least three different viruses. The most notable sign of these diseases is jaundice: a yellow discoloration of the skin and the white of the eye caused by the deposition of bile pigments that a damaged liver has failed to remove from the blood. One major form is type-*A* viral hepatitis, or short-incubation hepatitis, which used to be called infectious hepatitis (or epidemic jaundice or acute catarrhal jaundice). Its virus is transmitted primarily by the entry of material from the feces of an infected person into the alimentary tract of a susceptible person, but it can be transmitted in other ways. The other major form is type-*B* viral hepatitis, or long-incubation hepatitis; it used to be known as serum hepatitis (or homologous serum jaundice, transfusion jaundice or "needle stick" hepatitis). As the older names imply, hepatitis *B* is transmitted by inoculation through the skin of infected blood or blood components, but recent studies make it clear that the virus can be transmitted by other routes as well. In addition to these two well-characterized infections a new form of hepatitis associated with blood transfusions has been detected, with clinical symptoms and an incubation time like those of hepatitis *B*. It appears to be caused by an agent or agents unrelated to the hepatitis *A* or *B* viruses, and for convenience it is referred to as "hepatitis *C*" even though several different agents may actually be involved.

Last year nearly 60,000 cases of acute hepatitis were reported in the U.S.; about 60 percent were diagnosed as hepatitis *A* and 25 percent as hepatitis *B*, and in 15 percent of the cases the type was unspecified. Many cases are so mild that they go undetected, however, and physicians often fail to report the disease even when it is detected; the actual number of cases has therefore been estimated to be as much as 10 times larger than the number reported, making hepatitis a very significant disease in terms of human disability and public cost. In the case of hepatitis *B* some recovered patients become carriers of the virus, and some of them develop a chronic form of the disease. There are about a million hepatitis *B* carriers in the U.S.

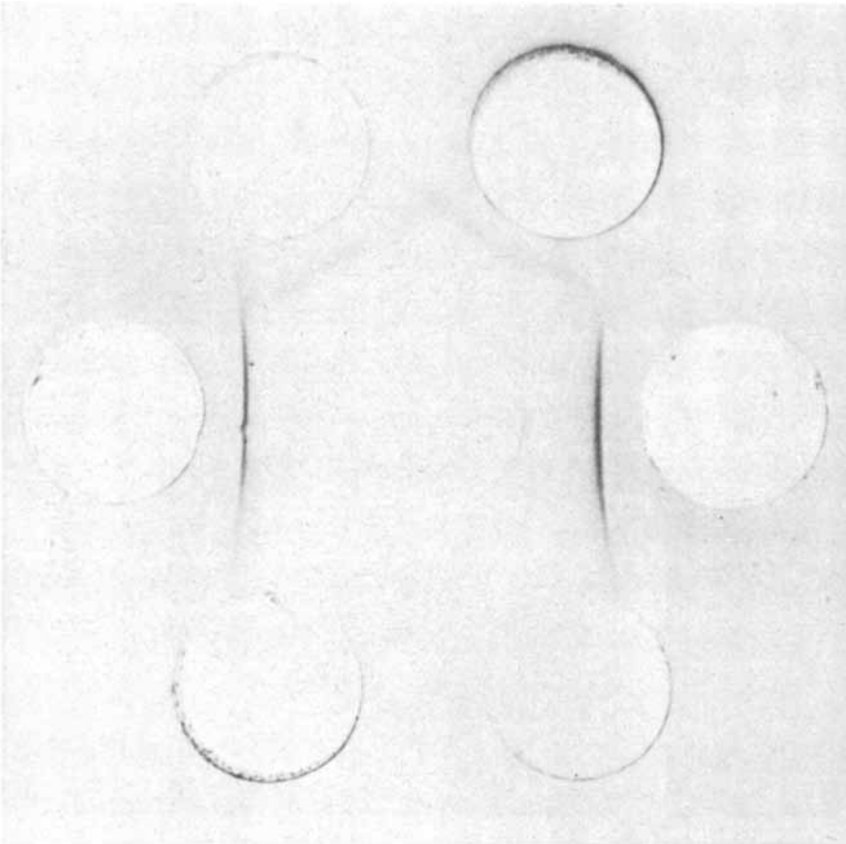
Until the early 1960's the two distinct viruses responsible for the diseases we now call hepatitis *A* and hepatitis *B* had not been characterized; it was simply known that there seemed to be two somewhat different forms of the disease, one with a longer incubation period than the other. Then Saul Krugman of the New York University School of Medicine and his colleagues gathered and preserved serum from patients with each form of the disease. They found that children exposed to the serum from short-incubation patients would regularly develop clinical hepatitis about 30 days later, whereas children given serum from long-incubation patients developed their symptoms about 60 days later; serum from these new patients in turn regularly produced the specific form of the disease in children to whom the serum was administered. Infection with one type of the disease conferred immunity against that type but not against the other type. Krugman and Joan P. Giles went on to show that if serum from a hepatitis *B* patient was inactivated by boiling and then inoculated, children given the heat-inactivated serum were protected against infection by hepatitis *B* virus. The experiments demonstrated that two different viruses were responsible for the two different forms of hepatitis. They also indicated that it should be possible to develop vaccines against the diseases, if the viruses could be isolated or propagated.

The chain of events that led to the discovery of the hepatitis *B* virus was a tortuous one, however. It began in 1963, when Baruch S. Blumberg of the Institute for Cancer Research in Philadelphia was examining thousands of blood samples from diverse populations in a study not of hepatitis but of genetic variation of serum proteins. He discovered that a serum sample from an Australian aborigine contained an antigen that reacted with an antibody in the serum of an American hemophilia patient. The patient, who had received multiple blood transfusions, had presumably produced the antibody in response to a previous exposure to the "Australia antigen." What was that antigen? Subsequent studies showed that it was fairly rare in North American and western European populations but much commoner in some African and Asian populations. In time Blumberg and his colleagues found evidence that linked the antigen to what was then still known as serum hepatitis. By 1968 other investigators, notably Alfred M. Prince of the New York Blood Center and K. Okochi of Tokyo University Hospital, had confirmed that the Australia antigen was found specifically in the serum of hepatitis *B* patients: the Australia antigen was a hepatitis *B* antigen.

This important antigen-antibody reaction was detected and studied primarily by a gel-diffusion technique in which two serums are placed in wells in an agar



HEPATITIS B VIRUS and its subunits are enlarged 200,000 diameters in three electron micrographs made by Guy A. Cabral of the authors' laboratory. Three different particles are commonly seen in the blood of hepatitis *B* patients (*top left*): small spheres about 22 nanometers in diameter, filamentous objects 22 nanometers wide and the double-walled hepatitis *B* virus itself (*arrow*), formerly known as the Dane particle. The plentiful 22-nanometer spheres can be purified (*top right*) and are the main source of vaccine material. The liver cells of hepatitis *B* patients contain 26-nanometer particles (*bottom left*), the unjacketed cores of the double-walled virus. The hepatitis *A* virus, unrelated to hepatitis *B* particles, is enlarged 200,000 diameters in a micrograph (*bottom right*), made by Clifton R. Gravelle of the Center for Disease Control's Phoenix laboratories, of an "immune aggregate," in which virus particles are linked by antibodies.



GEL-DIFFUSION METHOD developed by Orjan Ouchterlony of the Karolinska Institute in Stockholm detects hepatitis *B* surface antigen in blood serum and also demonstrates that there are different subtypes of the antigen. Seven wells are cut in an agar layer on a glass slide (enlarged here about seven diameters). Serum containing antibody against surface antigen is placed in the well at the center; plasmas from six individuals to be tested are placed in the outer wells. The antigens in the test serums and the antibody from the center diffuse radially through the gel, and when they meet, they form lines of precipitate. This Ouchterlony plate indicates that four of the serums contain surface antigens, three of one subtype and one of a different subtype. The origin of the precipitate lines is explained in the illustration on the opposite page.

gel and allowed to diffuse toward each other; if one serum contains an antigen to an antibody in the other, a visible line forms where the antigen and antibody meet and combine to form a precipitate. It was noted that if serums from two patients reacted with the hepatitis *B* antibody, a secondary "spur" sometimes appeared, indicating that there was more than one form of the antigen [*see illustrations on these two pages*]. Hepatitis *B* particles from any one person contain a "group" antigen common to all subtypes and at least two major subtype antigens.

The hepatitis *B* particles were soon identified. Electron micrographs revealed that serums containing the hepatitis *B* antigen contained a variety of particles whose sizes and shapes were similar to those of known animal viruses; many patients had a trillion particles per milliliter of blood. The major population in most hepatitis *B* blood samples consisted of small spherical particles about 22 nanometers (millionths of a millimeter) in diameter. There were also long, filamentous forms with the same diameter and with a length ranging from less than 100 to 700 nanometers. In 1970 D. S. Dane of the Middlesex Hospital Medical School in London observed still another form: double-shelled, viruslike particles with an outer diameter of about 42 nanometers and an inner core some 26 nanometers in diameter.

June D. Almeida of the Wellcome Research Laboratories in England provided the link between the earlier antigen-antibody observations and the new electron-micrograph observations. She reported that when a serum containing antibody against hepatitis *B* was reacted with a serum containing the various antigenic particles, all three particle forms were agglutinated: the 22-nanometer spheres, the filamentous forms and the 42-nanometer "Dane particles." If the blood samples were treated with a detergent to remove the outer shell of the Dane particle, however, the naked core was not agglutinated with the other particles. In other words, the envelope of the Dane particle was antigenically related to the 22-nanometer spheres and the filaments but the core of the Dane particle was antigenically distinct from those forms. The antigen associated with the spheres, the filaments and the envelope was designated HBsAg, or hepatitis *B* surface antigen, and the antigen of the core was called HBcAg, or hepatitis *B* core antigen. Very sensitive methods have been developed for measuring minute amounts of these two antigens. One of the most widely accepted is solid-phase radioimmunoassay, in which the antigen-antibody complex is rendered radioactive so that it can be quantified by a radiation counter [*see bottom illustration on pages 48 and 49*]; the method is as much as 4,000 times more sensitive than gel diffusion.

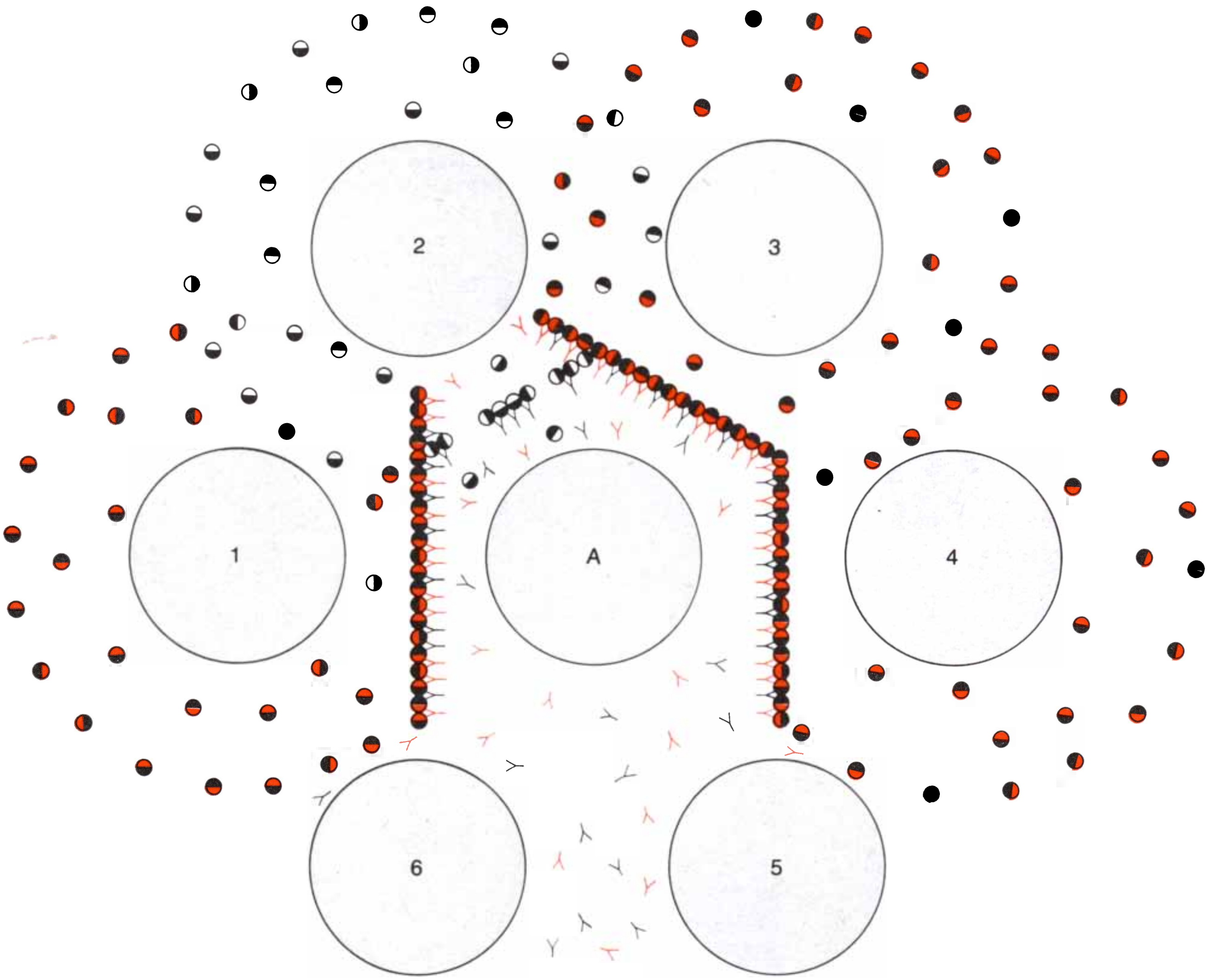
The next step was to purify and characterize the plentiful 22-nanometer spherical particles and see if they had the characteristics of an animal virus. The particles were purified in our laboratory at the Baylor College of Medicine and in other laboratories by a variety of centrifugation procedures and were then analyzed biophysically and biochemically. We found that the particles had a low density, which indicated that they contained lipid (fat) as well as protein. Further analysis revealed the presence of carbohydrate but did not reveal any nucleic acid, the genetic material that would be present in a virus. Precise measurements showed that there was a large variation in diameter, from 17 to 25 nanometers. The amino acid composition of the protein seemed to differ considerably from that of well-characterized animal viruses. Moreover, the particles had a complex biochemical nature: they consisted of about eight different polypeptides (short amino acid chains) or glycoproteins (protein-sugar complexes). All these findings suggested that the spherical particles probably represented not a virus but rather some virus-specific by-product or viral material that had been modified by the host cells in which the virus had proliferated.

Soon after the various viruslike particles had been observed in the blood of hepatitis *B* patients, investigators found antigenically related material in the cells of infected livers. The cytoplasm of the infected cells contained the surface antigen; the core antigen was detected only in the cell nuclei. The antigens were first visualized by tagging their antibodies with a fluorescent dye that could be seen under ultraviolet radiation. Their presence was confirmed by electron microscopy: infected liver-cell nuclei contained 26-nanometer particles that looked like the inner core of Dane particles, and such particles were not present in the liver cells of patients with other liver diseases.

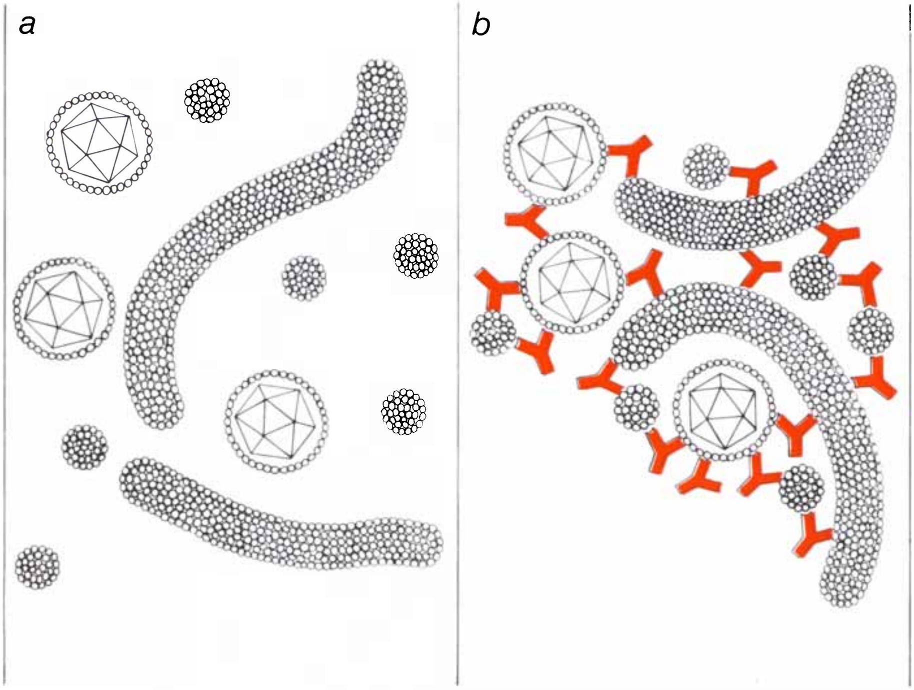
William S. Robinson and Paul Kaplan of the Stanford University School of Medicine established that the core particles isolated from blood contain a double-strand DNA and their own DNA polymerase (the enzyme required for the replication of DNA and thus for virus multiplication). We purified core particles from infected liver cells and found that they too contain double-strand DNA and polymerase. The molecular weight and the density of the cores are characteristic of certain animal viruses, and they behave like virus particles when they are subjected to electrophoresis. Most of the core preparations we have examined contain three different polypeptides, which is typical for the capsid, or protein coat, of many viruses. Taken together, these various findings strongly suggested that the Dane particle is a virus, with an inner core and with an outer shell that is at least similar to that of the "enveloped" viruses. Last year the World Health Organization recommended that the Dane particle in fact be considered the hepatitis *B* virus.

In addition to the surface and the core antigen-antibody systems a third system was noted in 1972 at the National Bacteriological Laboratory in Stockholm, the components of which have been designated the *e* antigen of hepatitis *B* and the anti-*e* antibody. The *e* antigen seems to be intimately related to the enveloped virus particle and to its infectivity. It is detected only in individuals carrying the surface antigen, and its presence is correlated with two other measures of infectivity: increased numbers of virus particles and hepatitis *B* polymerase activity. Among chronic carriers of the hepatitis *B* surface antigen those who carry the *e* antigen are more likely to have active liver disease than those whose serum contains anti-*e*. Similarly, the transmission of hepatitis to newborn infants is more likely when the mother's serum contains *e* antigen than when it contains anti-*e* antibody, and the injection of blood containing *e* antigen is more likely to produce hepatitis *B* than injection of blood containing anti-*e*. In other words, the presence of the *e* antigen in the blood of a surface-antigen carrier indicates the blood is particularly infectious; the presence of anti-*e* means it is much less infectious. The specific locus of the *e* antigen, however, is not known.

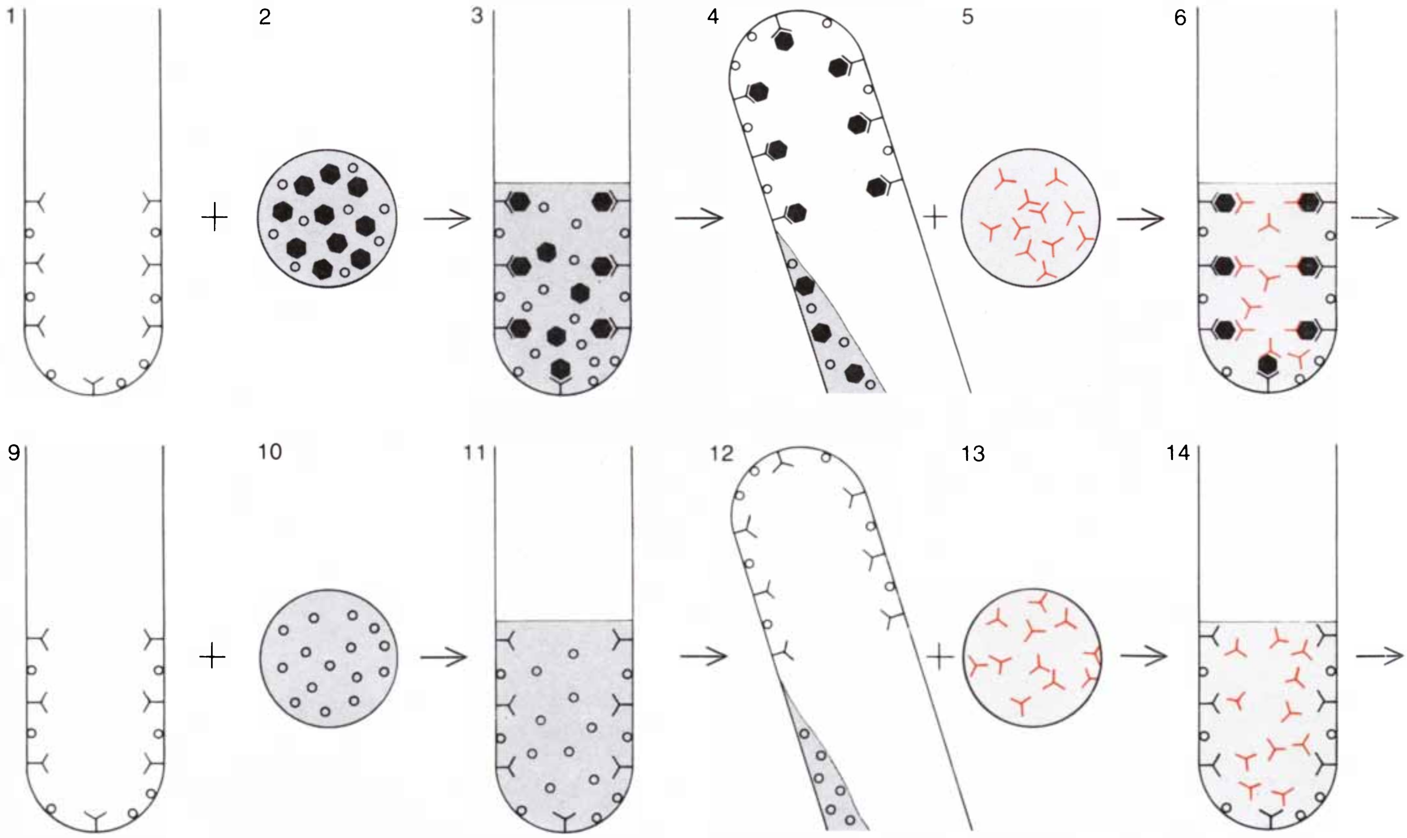
In spite of the accumulation of data concerning the hepatitis *B* virus and its antigens, the virus itself remained elusive and could not be grown in the laboratory. Fortunately an alternative possibility for a vaccine was available, however. It became evident that recovered hepatitis *B* patients were protected against reinfection by the presence in their blood of antibodies against the surface antigen. Prospective studies of hep-



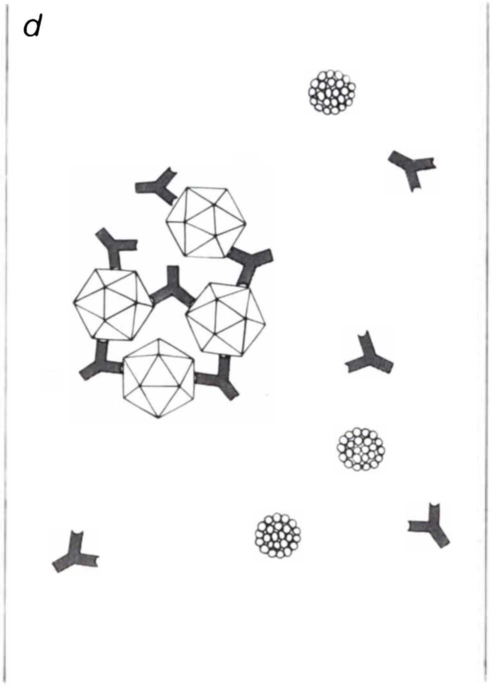
LINES OF PRECIPITATE in the photograph on the opposite page were formed by the complexing of antigen (*small circles*) with antibody (*Y's*). Serums in wells *1, 2, 3* and *4* contained surface antigen; wells *5* and *6* were negative. All antigen-containing particles have the same "group" antigenic determinant (*black*), but those from wells *1, 3* and *4* have one particular "type" determinant (*color*) and those from well *2* have a different type determinant (*white*). Antiserum from well *A* contains antibody against the group determinant and against one type determinant (*color*) but not the other. The line between wells *A* and *2* is fainter than the lines between well *A* and wells *1, 3* and *4* because, in the absence of antibody-matching antigen from well *2,* not as many antigen-antibody complexes could form there. Two "spurs" are seen, between wells *1* and *2* and between wells *2* and *3.* The reason is that type-specific antibody (*color*) continued to diffuse outward (in the absence of appropriate antigen from well *2*) until it complexed with type-specific antigen that came from wells *1* and *3.*



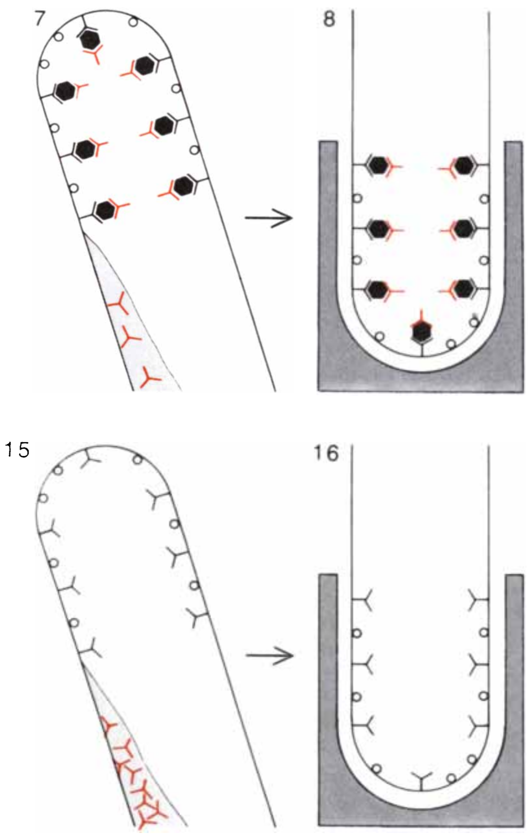
VARIOUS HEPATITIS B particles are related and differentiated by their antigenicity. The forms seen in patients' blood are the 22-nanometer spheres, the filaments 22 nanometers wide and the virus itself (Dane particle), which consists of an outer shell and an inner core (*a*). All three forms are characterized by the presence of the hepatitis *B* surface antigen, as indicated by the fact that the antibody (*color*) to that antigen combines with and agglutinates them all (*b*). Adding detergent to virus particles strips the surface-antigen outer



ANTIGEN can be detected by the technique of solid-phase radioimmunoassay. Antibody (*black Y-shaped forms*) specific for the desired antigen is adsorbed onto a solid support, in this case the wall of a plastic test tube (*1*), along with nonspecific proteins (*open circles*). The material to be tested (*2*) is added; if it contains the antigen (*black hexagons*), the antigen combines with the antibody (*3*). Unattached antigen is washed out (*4*). In order to detect the bound antigen more antibody, now labeled with a radioactive isotope (*colored Y's, 5*), is added. It binds in turn to the antigen, which is in effect sandwiched between antibody layers (*6*). The tube is washed again, leaving only



shell off them, yielding free cores (*c*). These are agglutinated by a different antibody (*dark gray*), the antibody to core antigen, distinguishing them from the surface particles (*d*).



the antibody-antigen-antibody sandwiches (*7*). The radioactivity of the tube is measured in a radiation counter (*8*) and is compared with that derived from testing a negative control sample (*9–16*) to quantify the antigen.

atitis caused by the transfusion of blood containing surface antigen indicated that none of the recipients whose blood contained preexisting antibody to the surface antigen developed overt hepatitis *B*, whereas a significant number of recipients who lacked the antibody did develop the disease. On the other hand, the presence or absence of antibody to the core antigen seemed to have nothing to do with protection against infection.

Two cardinal conclusions emerged from the investigations we have described. One was that antibody to surface antigen protects a person against infection by the hepatitis *B* virus. The other was that 22-nanometer surface-antigen particles could be harvested from the blood of apparently healthy carriers in a highly purified form and in large enough quantities to provide the material for a hepatitis *B* vaccine. An experimental animal for testing such a vaccine became available when it was discovered that chimpanzees are susceptible to the hepatitis *B* virus. Sure enough, animals inoculated with purified 22-nanometer particles that had been treated with formaldehyde did develop antibody to the surface antigen and subsequently resisted infection by an injection of blood containing highly infectious hepatitis *B* virus. Moreover, it turned out that antibody against the broad group antigen was protective, so that the animals were immune to the virus even when its surface antigen was of a subtype different from the antigen with which they had been vaccinated.

Last December, Philippe Maupas of the University of Tours reported preliminary findings of a study in which human volunteers were vaccinated with partially purified surface antigen treated with formaldehyde. Approximately 90 percent of the vaccinated individuals developed a specific immune response to the antigen. Of 80 vaccinated volunteers who worked in an artificial-kidney unit, where there was a continuous, high incidence of hepatitis *B* among patients and personnel, only one developed the symptoms of hepatitis *B*. Further tests of surface-antigen vaccines are being conducted by Maupas and by investigators in this country.

It is generally agreed, however, that vaccination with the 22-nanometer particle must be approached with caution. There are two troubling observations. One is that the complexity of the various viral components that apparently derive from a hepatitis *B* infection is more than can be accounted for by the amount of genetic information in the DNA of the core particle. The other is that inoculation of lower animals with the surface-antigen particles produces a vigorous antibody response, whereas human beings (the naturally infected host) and chimpanzees, even after infection, produce low levels of antibody to surface antigen. These observations suggest that the 22-nanometer particles may contain antigenic cell components (in this case perhaps certain liver proteins) derived from the host rather than purely "foreign" protein associated with the virus itself. As a result some individuals might develop an autoimmune response to, say, their own liver proteins. In support of this nagging fear several laboratories have reported that patients with chronic active hepatitis *B* develop an immune response to normal liver lipoprotein (complexes of protein and fat).

In view of these concerns it seems advisable to consider whether there may be a safer source of vaccine. Our laboratory and others are therefore investigating the immunogenic properties of individual polypeptide components of purified 22-nanometer particles. It appears that most of the individual polypeptides do induce an immune response in experimental animals and that animals injected with only one of the polypeptides produce an antibody that reacts with the intact 22-nanometer particle. The animals are being watched carefully to see whether they develop autoimmune responses to normal liver-cell material; only polypeptides that fail to induce such responses would be considered as candidates for a vaccine. Preparations of small, purified polypeptides would have the additional advantage of being free of viral or cellular genes that might present infectious hazards. If we find that a single-polypeptide vaccine has advantages over one made from the whole 22-nanometer particles, effective and economical methods will have to be developed for manufacturing sufficient quantities of the polypeptide from the particles.

The virus associated with hepatitis *A* is completely unrelated to that of hepatitis *B*, and very little is known about it as yet. It was first visualized in 1973 by S. M. Feinstone and his colleagues at the National Institutes of Health, in stools obtained during the acute stage of the disease from volunteers who had been experimentally infected with serum from hepatitis *A* patients. The virus particle has a diameter of 27 nanometers and, unlike the hepatitis *B* virus, does not have an outer envelope. The purified virus has been reported to have a wide range of densities, probably because different virus particles contain different amounts of nucleic acid; some of the particles seem to be so defective that they contain no nucleic acid at all. It has not yet been possible to establish what the genetic material is in these viruses. Some preliminary findings suggest that rather than being DNA, as in hepatitis *B*, the nucleic acid of the hepatitis *A* virus is RNA, as in poliovirus and a number of related viruses. Other properties, however, such as the relative heat stability of the virus and its wide range of densities, suggest that the virus may be closely related to a particular group of small viruses that contain DNA. Definitive studies will require a source of hepatitis *A* material adequate for biochemical analysis, which probably means they must await the development of a system in which the virus can be grown in the laboratory.

The epidemiology of hepatitis *A* and *B* and the serological changes observed in the two diseases are quite different, as befits infections caused by different viruses. Clinically, however, it is extremely difficult to distinguish between the two in their early stages. In either case the typical patient develops such symptoms as malaise, lassitude, muscle aches, loss of appetite, nausea and abdominal pains and runs a low to moderate fever. The urine turns brown and then the skin turns yellow; sometimes the other symptoms diminish after the jaundice appears. (In some hepatitis *B* cases there is inflammation of the joints, blood vessels and the kidney glomeruli, sometimes associated with a rash, and no jaundice; a diagnosis of hepatitis *B* can nonetheless be confirmed by examining the blood for the various serological markers and for biochemical evidence of liver dysfunction, notably a high level of the liver enzymes called transaminases.)

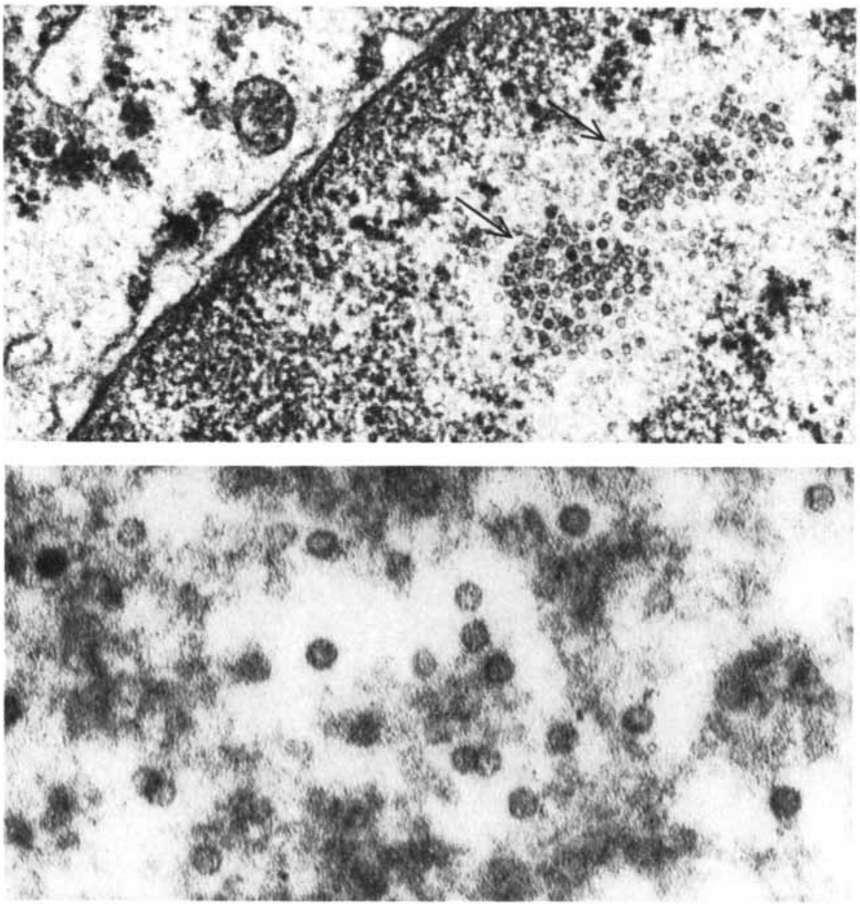
Recovery is generally complete within two to four weeks for hepatitis *A* and within four to eight weeks for hepatitis *B*. About 10 to 15 percent of the hepatitis *B* patients develop a severe variant of the disease, however. In somewhat less than 2 percent of all cases the disease progresses to rapid and severe failure of liver function, and for these cases mortality rates range from 70 to 100 percent. The overall fatality rate for all reported cases is .5 percent for hepatitis *A*, 3 to 4 percent for transfusion-associated hepatitis *B* and .8 percent for drug-abuse hepatitis *B*. These rates may of course be far too high, given the serious underreporting of mild infections.

The treatment for viral hepatitis is primarily "supportive," which means that there is not much that can be done. Patients are told to rest and to abstain from alcohol and other liver poisons. Adrenocortical steroids have been tried but have not been shown to be effective in uncomplicated cases, and they can actually cause a relapse if they are tapered off too rapidly. Steroids and concentrated hepatitis *B* immune serum globulin have been administered in severe cases, and again there is no convincing evidence of effectiveness.

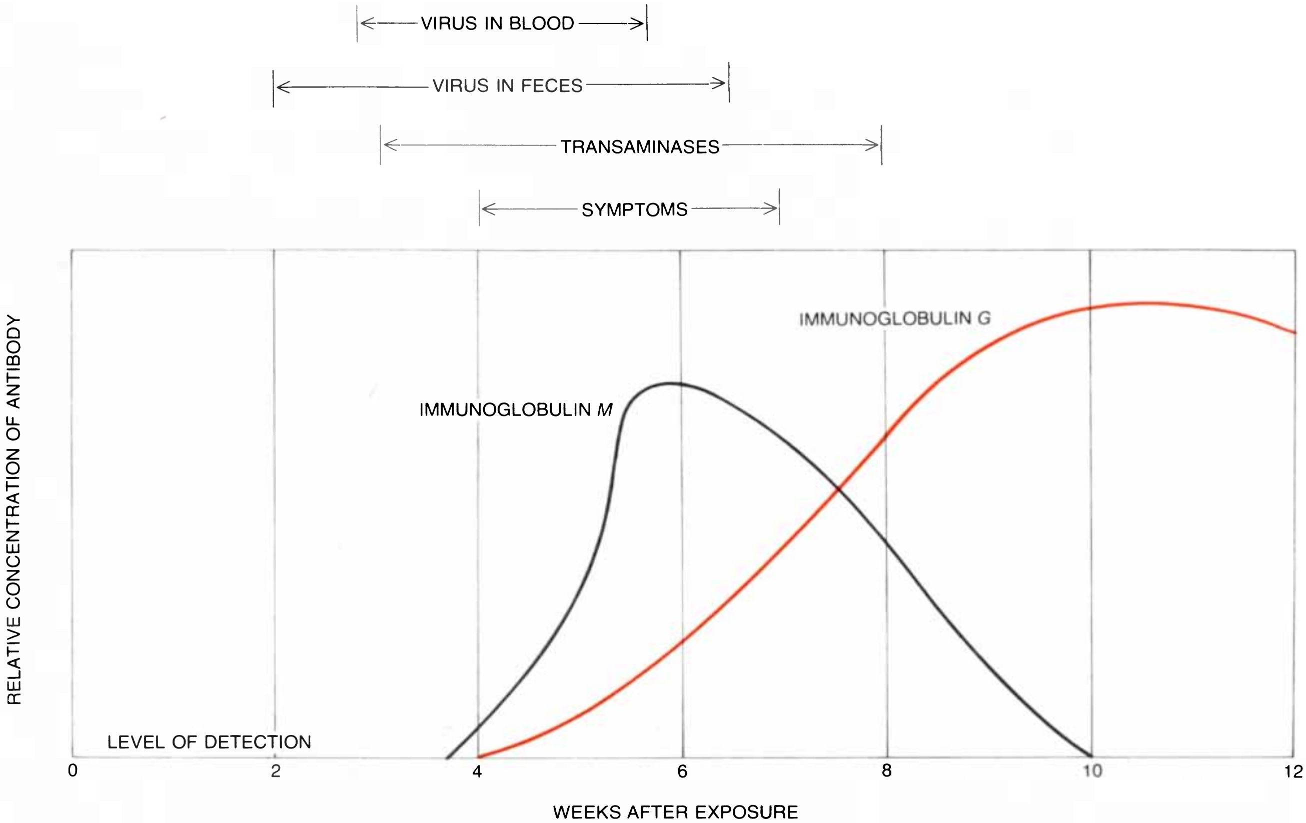
Therapy with interferon, the antiviral substance secreted by cells under virus attack, is currently being investigated. In preliminary trials at the Stanford University School of Medicine modest success has been reported in a limited number of patients with chronic hepatitis *B*. There was a transient suppression of the various virus markers present in the blood of chronic patients, but no definite effect on the liver-disease process itself was observed. The results imply that interferon may be effective in limiting the infectivity of individuals who are carriers of the virus or even in eradicating the chronic infections that afflict some carriers, but more extensive double-blind studies are required.

The biological and immunological events associated with hepatitis *A* are different from those of hepatitis *B*. In hepatitis *A* the virus probably multiplies in the lining of the intestine and then spreads to the liver by way of the blood or lymphatic vessels. Fecal excretion of virus or viral products has been reported as early as two or three weeks before the onset of jaundice, and the disease appears to be most communicable near the end of the incubation period (before symptoms appear), when the virus is most concentrated in the stools; by three or more weeks after the onset of symptoms the stools are usually noninfectious. In the early stage of the infection there are low levels of virus in the blood and urine, but the disease is seldom transmitted from those sources, and it is not thought to be spread by way of droplets from the mouth or nose. Several studies have shown a direct relation between the prevalence of hepatitis *A* (as measured by the prevalence of antibody against the virus) and both increasing age and poor environmental or personal hygiene, a correlation that is common to many virus infections of the intestinal tract.

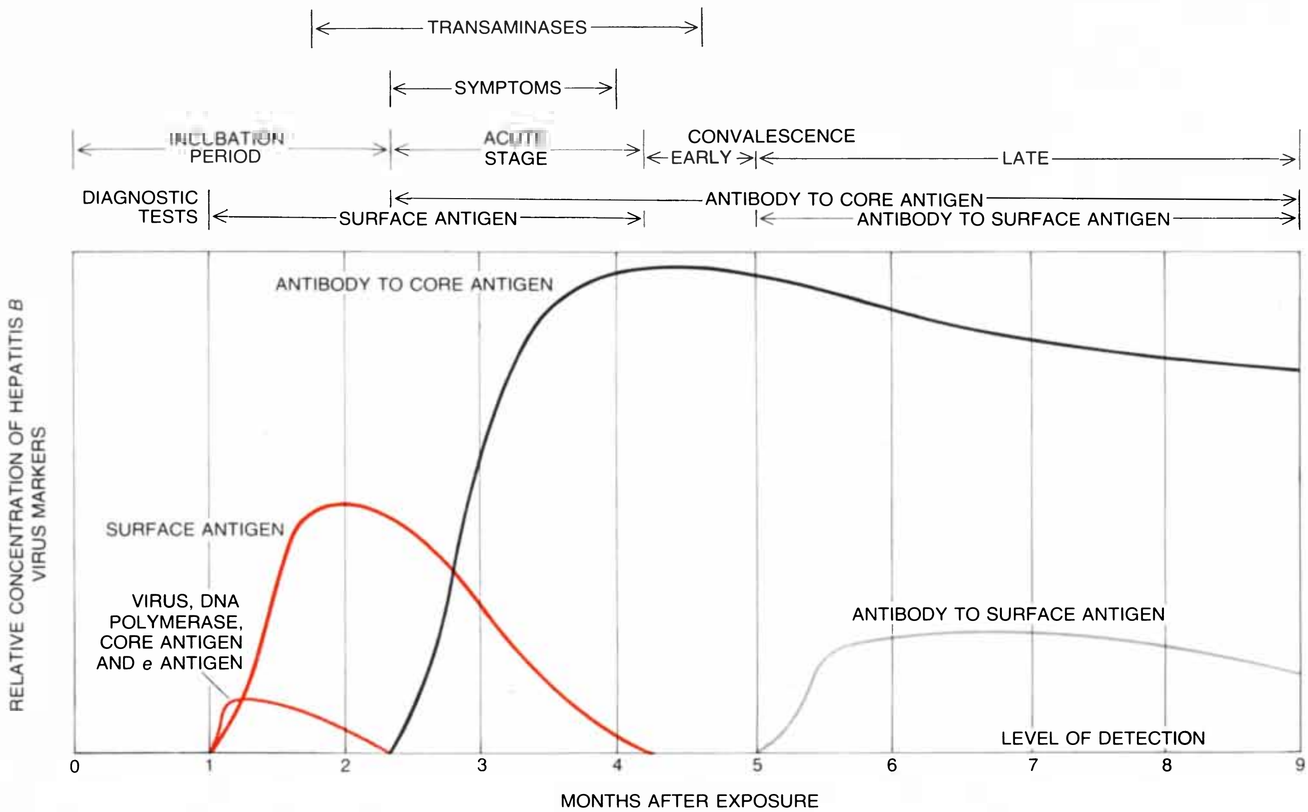
The clinical and immunological events that follow hepatitis *B* infection reflect a complex interaction among the host, the virus and particular viral antigens. In a typical acute case surface antigen appears in the blood one or two months after infection; it may persist for a few days or for several months. Virus particles, *e* antigen and polymerase activity appear in the blood along with the surface antigen. As the infection progresses, liver cells break down, releasing into the blood liver transaminase; elevated transaminase measurements foreshadow the appearance of hepatitis symptoms. Although surface antigen is ordinarily the best evidence for hepatitis *B*, it may disappear from the blood un-



NUCLEI OF LIVER CELLS of a chronic hepatitis *B* patient who died of kidney disease contain hepatitis *B* core particles. Clusters of particles (*arrows*) are enlarged 55,000 diameters (*top*) and particles in another nucleus are enlarged 200,000 diameters (*bottom*) in electron micrographs made by Ferenc Gyorkey of the Veterans Administration Hospital in Houston.



HEPATITIS A has about a four-week incubation period between exposure and appearance of symptoms, during which the virus and liver enzymes (transaminases) are detected in the blood. Curves trace the concentration of antivirus antibody in two kinds of immune globulin: immunoglobulin *M*, an acute-phase antibody, and immunoglobulin *G*, which dominates during convalescence and persists for decades.



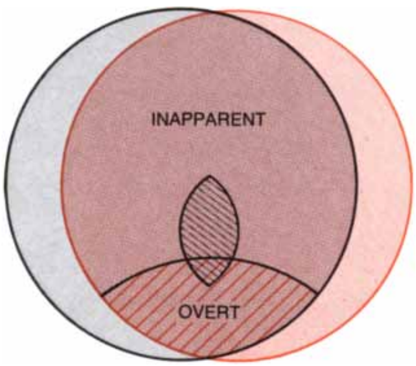
HEPATITIS B has a longer incubation period than hepatitis *A*. The various hepatitis *B* antigens and antibodies become detectable and increase and decrease in concentration at different times, and the effectiveness of various serological diagnostic tests varies accordingly.

usually early; in such cases the diagnosis can be established by the presence of antibody to core antigen, which becomes detectable along with the clinical or biochemical signs. Antibody to surface antigen is also detectable after the antigen disappears, but only by sensitive tests.

In patients who develop chronic hepatitis or become symptom-free carriers the surface antigen persists in the blood, its antibody does not appear and high concentrations of the antibody to core antigen are detected. In chronic hepatitis *B* patients the blood frequently contains *e* antigen, DNA polymerase and virus particles, whereas the blood of apparently healthy carriers is more likely to contain only the antibody to the *e* antigen. If someone immune to hepatitis *B* is reexposed to the virus, there is a rapid "recall" immune response, with antibody to the surface antigen appearing within two weeks; the surface antigen itself is not detected and the person is not infected.

The incidence of hepatitis *B* in the general population is greatest in the 15-to-29-year age group, presumably because of drug abuse. The incidence is higher, however, in specific groups such as artificial-kidney patients and their attendants. Until recently it was very high in recipients of blood or plasma, but the incidence of hepatitis *B* infection associated with transfusion has been reduced almost 75 percent since it became mandatory to screen donors by sensitive tests such as radioimmunoassay.

The natural spread of hepatitis *B* by contact (rather than only by skin puncture) appears to be commoner than it has generally been thought to be, particularly in the case of sexual partners of carriers of the virus (whether the carriers have chronic symptoms or not) and among residents of crowded institutions. The fact that surface antigen is



HEPATITIS B INFECTION is represented here as the intersection of virus (*color*) and host population (*gray*). Most of the infections are inapparent, but some are overt (*colored hatching*). In both instances some infected individuals become carriers (*black hatching*).

regularly detected in various body excretions and secretions supports the concept that hepatitis *B* can be transmitted by routes that do not involve penetration of the skin. Within families a virus carrier can transmit the disease to someone who uses his towel, razor or toothbrush, for example. (In any such case, to be sure, the virus could actually have penetrated the skin or mucous membrane.) In general hepatitis *B* is more likely to be communicated by a patient in the acute stage or a person who has the chronic disease than by an asymptomatic carrier; for an asymptomatic carrier to transmit the disease may require repeated exposures.

Not everyone who is exposed becomes infected; of those who do become infected only a small proportion develop the clinically observable disease, the percentage depending primarily on the dose and the route of infection: small doses of virus and infection by a route other than through the skin result in a longer incubation period and a small likelihood of overt symptoms, with jaundice appearing in perhaps only one in 100 cases. Larger doses and infection through the skin can decrease the incubation period and raise the jaundice-to-nonjaundice rate to one in five. People who are infected but do not develop symptoms and those who become carriers (defined as those whose blood level of surface antigen remains about the same for more than four months) are particularly significant sources of further infection.

In several studies a high prevalence of surface antigen has been observed in individuals who have primary carcinoma of the liver: in 80 percent of such cancer patients in Taiwan (compared with 15 percent in a control sample) and in 35 percent of such patients in Spain (compared with 1 percent of the controls). Similar surface-antigen rates have been observed in parts of the U.S. in patients with liver-cell carcinoma that has developed from nonalcoholic cirrhosis of the liver. Several possible explanations have been offered for this association of hepatitis *B* and liver cancer. One explanation is that patients with liver cancer or cirrhosis are particularly susceptible to hepatitis *B* infection or to becoming chronic carriers. Another is that the infection leads to cirrhosis, which is the actual precursor to the cancer. A final possibility is that the hepatitis *B* virus may transform the liver cells it invades, in other words, that it is a tumor-causing virus (influenced, perhaps, by genetic, environmental and other factors). If either of the last two interpretations is correct, the control of hepatitis *B* by immunization should reduce the incidence of liver cancer.

Some transfused patients continue to develop viral hepatitis in spite of the careful screening of donors and their blood products for indications of the *A* and *B* viruses, and without the appearance of such indications in the patients' blood. In the absence of methods for identifying the agent or agents of non-*A*-non-*B* hepatitis (which now accounts for more than 80 percent of the transfusion-associated hepatitis in the U.S.) not much can be learned about its transmission and immunological characteristics. A few things have been noted, however. The clinical and epidemiological patterns are similar to those of hepatitis *B*; the disease appears to be common and to have a carrier state, and recipients of blood from donors with an elevated transaminase level apparently have an increased risk of developing this newly recognized form of hepatitis.

Until more results are in hand on hepatitis *B* vaccines the effort to control viral hepatitis will have to depend on traditional methods, which means good personal hygiene, public-health measures and pooled immune globulin. The washing of hands before eating or handling food is probably the single most important preventive in the case of hepatitis *A*; for hepatitis *B*, exposure to anything that might be contaminated with blood, ranging from a borrowed towel or razor to a hypodermic needle, should be avoided. Replacing commercial blood banks and paid donors with an all-volunteer system would reduce the incidence of types *B* and *C*. Immune globulin (what used to be called gamma globulin) should be administered to people who come in contact with hepatitis *A* within a family or an institution, but probably not to routine contacts at work or in school; it is not helpful to give immune globulin more than six weeks after exposure or after symptoms develop. In the case of hepatitis *B*, standard immune globulin is probably best reserved for situations in which a small amount of virus has been inoculated accidentally. A high-concentration hepatitis *B* immune globulin has been produced recently that has given promising preliminary results in certain high-risk groups.

As for a vaccine, what we know about the hepatitis *B* antigens and about how to manipulate them suggests that the results of current tests will be favorable, and that within the next two years a vaccine will be available for immunizing people who are frequently exposed to hepatitis *B*. The past decade's advances in knowledge concerning viral hepatitis have been truly remarkable. And the techniques designed to investigate hepatitis have provided new tools for investigating a wide spectrum of human diseases that are suspected of being viral in origin but for which the standard approaches have been unproductive.

You're giving a paper and need slides?

Why? Because it's customary? Perhaps you have it all figured out: two people present who already know your whole story and question your conclusions, one other guy who knows it and questions your data, and about a dozen perfect strangers who won't have the foggiest notion of what you are talking about and therefore do not matter. Obviously, then, the slides don't matter either, as long as you have some and they do not draw too much attention by emphasis or excessive legibility.

NO! NEVER! We choose to assume you are an idealist, that your purpose is not escape for an hour or a week from your cubbyhole but to convince those twelve strangers that what you are about to reveal to them they really want to know.

Who are they? Why are they there? What do they know? What don't they know? To compose a realistic scenario about them before you ever meet them may take some thought and investigation, but such is your obligation. Recognizing that as no less an obligation than to know what you are talking about, you put yourself in their shoes as you plan the sequence and content of your slides.

For each slide that you will want, you go so far as to make out a card. A card on which you find yourself writing quite a few words or a rather elaborate sketch you generally break up into several cards for several slides. You realize that your audience cannot linger over a complexity, as one can reread a difficult passage in print.

With luck you find you can shuffle some extant slides into the deck of new ones. You do so with more regard for pertinence than convenience, since you *are* an idealist.

And since illegibility is not your intent, and since both friends and strangers insist on sitting as far back as they can get, which ought to be about 8 times the height of the screen, you judge a table or chart that you want to photograph as a slide by trying to read the finest detail on it from a distance of 8 times its height. If it's no go (and never mind that that illegible finest detail is not crucial!), you redraw and simplify or forget it.

And here below we leave a useful little gift for an idealist planning an engagement away from the ol' cubbyhole:

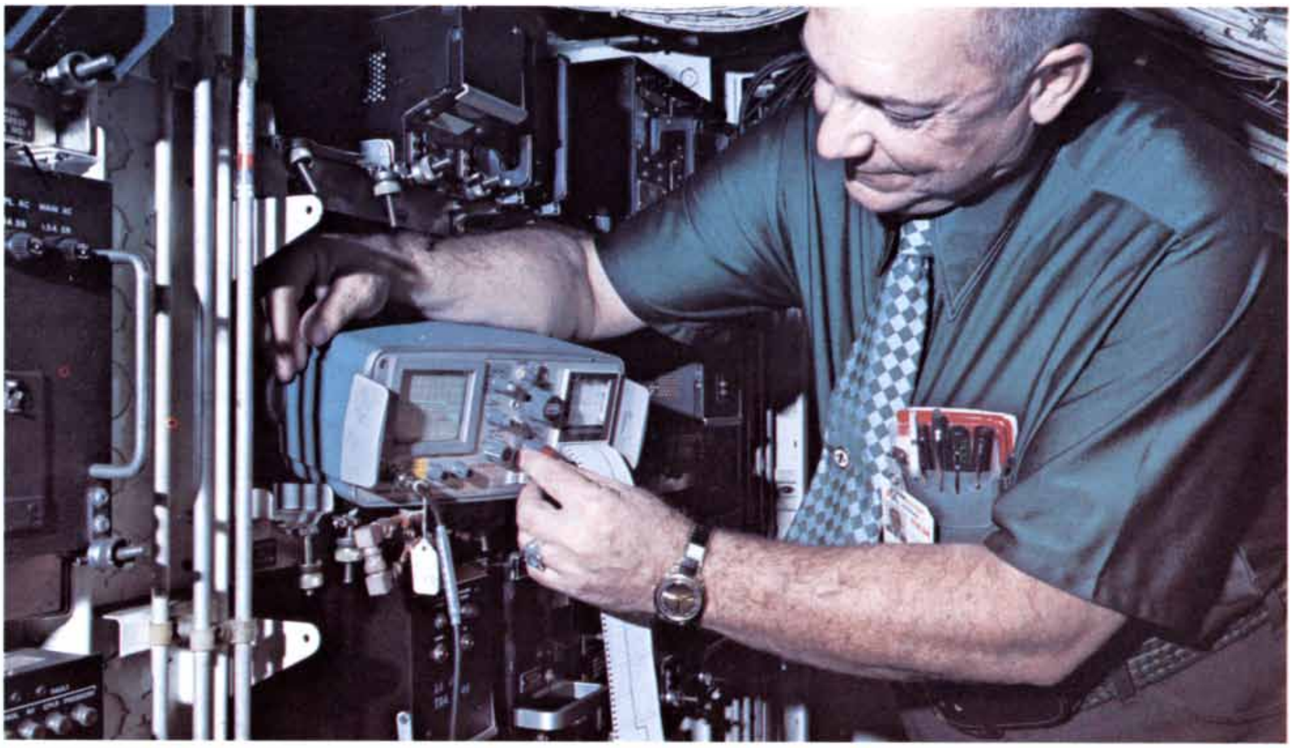
```
When making text or title slides by typewriter,
this template can serve as a guide.  Include about
3 mm outside the template lines in the camera view-
finder.  Nine double-spaced lines are maximum.  For
a projected image 1.5 m high, upper- and lowercase
elite type will be legible 12 m away.  USE OF
ALL UPPERCASE WILL EXTEND THE LEGIBILITY DISTANCE
A LITTLE FURTHER.  For true kindness, this long a
message should be divided among two or three slides.
```

But why are we telling you all this? As long as you use Kodak film for your slides, why do we care? Well, there's the possibility that we may be one of those 12 strangers in the audience.

Kodak

"With TDR* cable testing at American Airlines, we're finding cable disturbances long before they cause system failures."

Aubrey Thomas, American Airlines
Maintenance and Engineering Center, Tulsa, Oklahoma



Aubrey Thomas checks a cable and antenna by looking at their signatures on the screen of a TEKTRONIX TDR cable tester in a 747's electronic equipment bay, from which most routine avionics troubleshooting can now be done. Not only can he spot subtle changes in systems—if there is a problem, he quickly knows what it is and precisely where.

*Time Domain Reflectometer

Aubrey Thomas points out that "anybody who has had to pull panels throughout an airplane to find a cable fault will understand why we appreciate the TDR's* ability to quickly identify the nature of the problem and to tell where it is—often to the inch, or closer.

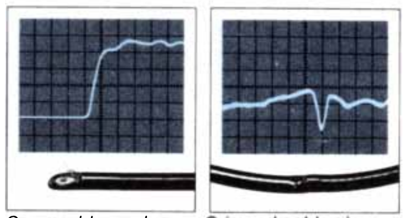
"But even more impressive is a situation where, let's say, you might have to figure out why a pilot was getting strong navigation signals close to a station, but not farther away. If you used a meter and it showed continuity in the system, you might waste hundreds of man-hours and change thousands of dollars worth of components, and still not be any better off.

"But by using the TDR in the equipment bay, you would know immediately that the problem was, let's say, six feet of corroded cable shield, starting at the vertical-fin disconnect and coming back toward the inside fuselage. That's the beauty of the TDR, it's not limited to identifying shorts or opens—it points out any disturbances."

Back in 1971 Aubrey Thomas and two other people at American Airlines were looking for a better way to test the miles of cable that carry radio, navigation, and instrumentation information, as well as power for complex basic electrical systems. "I knew that the TV industry used time domain reflectometry a lot," Aubrey says, "and couldn't see why it wouldn't work on airplanes."

They started by using a TEKTRONIX oscilloscope and plug-in TDR unit, on a roll-around cart. But because work quarters on aircraft are often either cramped or relatively inaccessible, what they really needed was something much smaller and battery powered.

Recognizing these needs, Tektronix had been developing TDR units that were not only small, portable, and battery operated—but were also simple to use, had chart-recording capabilities, could be used in hostile environments (rain, snow, extremes in temperature and altitude), and on a wide variety of cable types.



Open cable produces this characteristic signature. Detectable by other methods, but a TDR gives the precise location.

Crimped cable signature. One example of a potential problem virtually impossible to detect by conventional methods.

What an avionics tech sees

A time domain reflectometer (TDR) works in a manner similar to radar. It generates repetitive pulses of energy that are sent down a cable and displayed on a cathode-ray tube screen. Any cable faults (impedance changes) cause pulse reflections. The shape of the reflection (the signature) allows the type of problem to be identified. Distance to the fault is determined by measuring the elapsed time between the pulse and reflection. Entire lengths of cable

"Cost reductions are a major objective at American Airlines. One way we achieve this is through preventive maintenance programs that result in less down-time for airplanes."
John Hill, Supervisor, Aircraft Maintenance Support Group

can be examined, and when a disturbance is detected, that portion of the cable can be viewed in detail.

Conventional testing methods require access to a cable at both ends (a watt meter, for example, being used at one end and a signal source at the other). On aircraft, there is the sheer physical problem of getting to and removing an antenna, for example, which may be out on a wing, on top of the plane, or six stories up a tail fin. Additionally, conventional methods tell if a cable system is shorted or open, but they cannot locate the fault. Nor do they deal with the more difficult situations, such as gradual impedance changes caused by moisture and corrosion. TDR is also very useful where knowing precise cable length is critical.

American Airlines keeps Aubrey Thomas and George Beyl on 24-hour call, forming a two-man office that acts as a liaison between engineering, flight, and repair personnel—doing everything from identifying avionics problems to developing and evaluating test equipment and training maintenance personnel.

Their job is to come up with tools and procedures that solve problems quickly and eliminate guesswork. George recalls an incident that happened in the early days of TDR at American, involving another airline's plane on which American had contracted to do heavy maintenance work. "They'd been having a VHF problem, and I knew they'd been changing transceivers trying to solve it. The TDR told me that the problem was in the cable, in a relatively inaccessible place, 77 feet from the electronic equipment bay. Not a place you want to go into unless you're sure. It took a lot of convincing. And about four hours to get in there. But when the man from the other airline saw I was right, and had located the problem precisely, he looked straight at me and said, 'What is that machine? I want one of them.' "

Tektronix' 1502 TDR Cable Tester ($3,200) is recommended for cable lengths up to 2,000 feet. For longer lengths, such as telephone lines, the 1503 TDR Cable Tester ($2,985) is recommended. The optional, plug-in Y-T Chart Recorder is priced at $575.
U.S. sales price FOB, Beaverton, Oregon

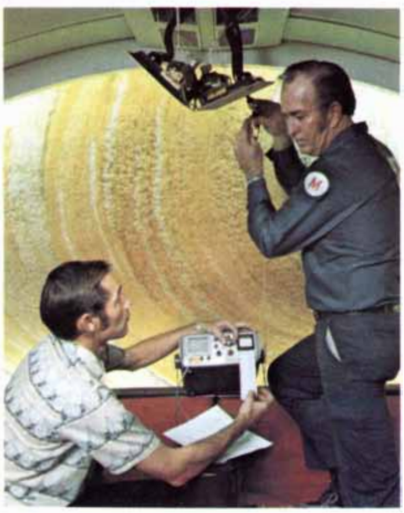
These are just three of more than 835 TEKTRONIX products that serve the test, measurement, and information display needs of thousands of customers worldwide.

For a copy of our annual report or product information, write to the Marketing Communications Department, 76-260, Tektronix, Inc., P.O. Box 500, Beaverton, Oregon 97077.
By phone—call 503-644-0161.

Tektronix
COMMITTED TO EXCELLENCE

Knowing the condition of cables is becoming even more important as avionics systems increase in complexity. Chart recordings provide useful system histories, records of unusual problems, and can be used as training aids.

Within minutes, George Beyl (left) and Sterling Griffin tested the new Omega navigation system cable. "The thing about cables on aircraft," George says, "is that they're buried. Even when you do uncover them, you can't always look at one and say it's good or bad."



"One beauty of the TDR is its simplicity," says Aubrey Thomas. "I can train a mechanic to use it in about four hours. And since it's battery operated, we can take it anywhere we need it."

SCIENCE AND THE CITIZEN

Lead Times

It has been estimated that as much as 40 percent of the cost of building a large nuclear power plant in the U.S. today is accounted for by interest and inflation accumulated during the 10 years or more that currently elapses between the date a utility decides to build such a plant and the date the plant becomes operational. In an effort to alleviate the costly lead-time problem for the construction of new nuclear power plants the Carter Administration's energy program calls for a "thorough review" of current licensing procedures for nuclear plants. As a part of that review President Carter proposed in his energy address to Congress in April that "we establish reasonable, objective criteria for licensing, and that plants which are based on a standard design not require extensive individual design studies." Even with the most thorough safeguards, he remarked, "it should not take 10 years to license a plant."

What are the prospects that the Administration will achieve its announced goal of reducing nuclear-plant lead times? According to a study conducted by the Energy and Materials Division of the General Accounting Office, the outlook is not encouraging. The GAO study noted that the total lead-time period of 10 years currently consists of four phases: the utility-planning phase (approximately two years), the construction-permit review phase (two or more years), the construction phase (about six years) and the operating-license review phase (two to three years, timed to run in parallel with the final stages of construction). Under the Energy Reorganization Act of 1974 the Federal agency responsible for conducting the two major reviews required before a utility can be issued licenses to construct and operate a nuclear power plant is the Nuclear Regulatory Commission.

In recent years, the GAO report points out, the NRC has taken certain administrative steps to reduce lead time to seven to eight years and has proposed legislation designed to further shorten it to six years. For example, the report says, "a key administrative Commission change allows utilities to begin some construction work as soon as they complete the environmental and site-suitability requirements of their construction-permit application reviews.... To further expedite construction, the Commission is proposing to review sites before receiving construction-permit applications."

According to the GAO report, other steps already taken by the NRC to reduce nuclear-plant lead times and to make its plant-design review process "more stable and predictable" include "(1) encouraging the development and use of standard nuclear power-plant designs ... and (2) more carefully controlling the manner in which new Commission safety requirements are applied to power plants in the design, licensing and construction stages."

The authors of the GAO report write: "We commend NRC's efforts but they have had little or no impact on reducing time frames. We recognize that some of the measures taken are long term and have not been fully implemented. In our opinion, however, the factors now limiting NRC's success in reducing lead times to seven to eight years will continue to do so and will also preclude utilities from achieving the stated goal of six years. We believe the six-year goal is unrealistic and the programs implemented and proposed by NRC to achieve that goal will not appreciably reduce the present 10-year lead time. In fact, we believe that both NRC and the industry will have difficulty in maintaining current time frames of 10 years."

Among the factors that limit the Administration's ability to reduce nuclear-plant lead times the GAO study lists: "Growing State and local government requirements/restrictions which are intended to lessen environmental impacts but are diametrically opposed to NRC's actions to shorten power-plant lead times.... Growing public concern about nuclear power as shown by recent State nuclear moratorium initiatives. ... Changes resulting from court decisions invalidating NRC regulations.... Changes resulting from technological advances and operating experience."

In responding to an advance draft of the GAO report, Lee V. Gossick, executive director for operations of the NRC, did not take issue either with the report's projections or with its sole recommendation ("that the Chairman, NRC, work jointly with all the States to identify and compare the various legal and procedural requirements as a first step in developing some commonality in the licensing process"). He did, however, emphasize that "many of the factors that militate against reducing lead times are, in fact, beyond the control of the Commission," adding that "we do not agree that the licensing process is the dominant factor in current power-plant lead times."

How Much Crime?

The general public perception that there is more and more "crime in the streets"—and everywhere else—has been created in part by official statistics on crime, notably the Uniform Crime Reports series published by the Federal Bureau of Investigation on the basis of data from local police departments. Such data necessarily reflect only crimes that become known to a police officer, are then recorded, classified and categorized according to varying practices and are eventually made known to the FBI. Dissatisfaction with the Uniform Crime Reports was a major reason for the establishment in 1972 of the National Crime Surveys, a "victimization" series that seeks to measure crime as experienced by people rather than as recorded by the police. The surveys, based on interviews with individuals in national and individual-city household samples, measure the extent to which people have been victimized by rape, robbery, assault, larceny, burglary and automobile theft. It was hoped that the survey results would serve as a "calibration" for the FBI reports and reveal the "actual" extent of crime.

That hope has not been fulfilled and is not likely to be, according to the report of a panel established by the National Academy of Sciences. The panel's mission was to evaluate the survey operation, not its results, and so the report deals largely with such issues as questionnaire design, sampling and the interpretation of data. The panel nonetheless has some things to say about crime statistics in general and about what they can and cannot be expected to measure.

One problem with the Uniform Crime Reports is their limited scope: they do not cover most organized-crime activities or such white-collar crimes as tax evasion and political corruption. Even for the crimes that are covered, official statistics "are necessarily an imperfect measure." As incidents work their way through the police system, at each step "some crimes are excluded (and, perhaps, some noncrimes are included)." Therefore "for most types of offenses, police statistics never can be expected to provide valid measures of the 'true' levels of crime." This is so "no matter how much they might be supplemented (or calibrated) by victimization surveys." The National Crime Surveys victimization studies, like the police reports, exclude white-collar and organized crime, and they have their own sources of inaccuracy. For example, the results depend on personal perception and memory, which apparently cause certain crimes to be underreported; the results are skewed by the treatment of multiple offenses against a single interviewee; there are "numerous problems of sampling and response," including underenumeration of certain population groups.

One thing that victimization surveys do suggest, according to the panel, is that "for the majority of Americans, crime of the type surveyed in [the Na-

Amédée Gordini, known as "The Sorcerer" for his magic with racing cars, presents one of his greatest feats.

Throughout Europe the cars of Amédée Gordini represented France in Grand Prix and sports car racing for many years. The French bestowed the name "The Sorcerer" upon him because of his inimitable touch with racing engines and high performance cars. And he bestowed his name on this coupe-convertible.

At the heart of the Gordini is the engine designed by Gordini. This 1647-cc, four-stroke engine with a high compression ratio and electronic fuel injection will more than satisfy the requirements of any driver who admires pure performance. Hooked to the single-overhead cam engine is a smooth five-speed gearbox that gives you total command of acceleration.

When you add front-wheel drive, rack and pinion steering, Michelin steel-belted radials, and a sophisticated suspension system with front and rear stabilizer bars the end result is a highly responsive sports car. A car that can cruise effortlessly all day as fast as you want to go. A car with superior handling and the road-holding ability that the French are famous for. Now add another feature that the French are famous for — style and luxury.

In front, infinitely-adjustable soft reclining bucket seats. In back, gracefully sculptured contour seats.

The Gordini's fully carpeted interior is spacious enough to seat four people very comfortably. In the rear there's a roomy 10-cubic foot trunk.

The Gordini's three-way top is something you won't find in any other car at any price. The sturdy fiberglass hardtop lifts off to reveal a soft, supple convertible top that is electrically operated.

Brock Yates of Car and Driver concluded "with a price tag under $7,000 the Gordini provides power, speed, quality and handling generally unknown in automobiles costing less than $10,000." Now that's some trick. Call 800-631-1616 for your nearest Gordini dealer.

STANDARD FEATURES	Gordini	BMW 320i	Porsche 924
Rack & pinion steering	YES	YES	YES
5-speed transmission	YES	NO	NO
Front-wheel drive	YES	NO	NO
Electronic fuel injection	YES	YES	YES
Electric sun-roof	YES	NO	NO
PERFORMANCE			
*0-60	9.2	10.5	9.9
*Standing ¼ mile	17.1	17.3	17.3
**EPA fuel estimates	36 H'wy 21 City	29 H'wy 20 City	31 H'wy 17 City
†BASE PRICE	$6998	$7990	$9395

*Test results reported in July, August and December 1976 issues of Car and Driver magazine. Cars tested were a 1976 Gordini, 1976 Porsche and 1977 BMW.
**The actual mileage you get will vary depending on the type of driving you do, your driving habits, your car's condition and optional equipment. Not available in California. †East Coast P.O.E.

The Great Gordini

by Renault

tional Crime Surveys] is not, in fact, an important personal problem—compared with issues such as inflation, unemployment, educational costs or race and sex discrimination." Nevertheless, "concern about crime is real. Crime is thus likely to remain an important fact of political life." Even if the surveys cannot measure "actual" crime, they can provide important information about the distribution and social consequences of crime and also data relevant to specific issues such as gun control and compensation for victims of crime. And by exploring public attitudes about crime and criminal justice "the surveys could help to dispel the ignorance, misunderstanding and irrational fear that now so often characterize public debate and discussion of crime."

Leveling the Load

It is well known that the demand for electric power reaches a minimum in the hours before dawn and typically rises to a peak during the first several hours of the working day. In the summer months, when air conditioners are widely used, the consumption peak may be shifted to the late afternoon. Since electric utilities must install enough capacity to meet peak loads, with a margin to spare, it might seem that they could achieve substantial savings if they could slice off the peaks of daily demand and somehow fill in the valleys. A recent issue of the *EPRI Journal,* published by the Electric Power Research Institute, describes the industry's growing interest in load leveling but concludes that dramatic savings cannot be expected.

Electric utilities distinguish between daily and annual load factors. The daily load factor is the ratio of average demand to peak demand for one day; typical values range from 75 to 90 percent. The annual load factor is a lower number because its numerator is an average of all daily average demand values for a year, whereas its denominator is the highest one-hour demand during the same period. For U.S. utilities the average annual load factor is 61 percent.

There are two general approaches to leveling power demand: one is to provide the customer with price incentives to stagger his consumption, the other consists of measures the utility can take on its side of the electric meter. Two kinds of price incentive are common among large commercial and industrial users. Utilities may charge reduced rates for power consumed at certain times of day; alternatively the unit rate for all electricity consumed in a month may be determined by the highest half-hour demand on the utility system, a method called peak-load pricing. (In a few cases large institutional users will agree to accept interruptible power in return for extremely low rates.) Perhaps the fastest-growing of the price-incentive systems is peak-load pricing because it requires no change in the usual working day. To take advantage of such a pricing system the consumer can buy control instruments that act on programmed instructions to shut off low-priority electrical equipment when total demand approaches a preset ceiling.

On its side of the meter the utility can install similar instruments that respond to time, temperature or the state of operating equipment and activate selected customer circuits or loads. EPRI and the Energy Research and Development Administration (ERDA) are cooperating in a $7-million test of two-way communication and control hardware for load-management schemes under direct control of the utility. The primary goal of load leveling is to make existing installed capacity suffice while average demand slowly increases. In this way the utility can defer capital expenditures for new additions to capacity. As the new higher loads are spread out to attain a high load factor, however, the less efficient generators used for intermediate and peak loads are run for longer periods, with resulting higher fuel and operating costs. Thus load management becomes a proposition covering 20 to 30 years, during which the utility optimizes its generating mix for a higher load factor, thereby allowing a rising fraction of its output to be provided by the most efficient base-load generators.

A study by the Edison Electric Institute suggests that savings from load management will be modest. Using actual 1970–73 data for an oil-fired utility system as the basis for comparison, the study shows that the maximum saving effected by achieving a daily load factor of 100 percent (that is, a situation in which all daily loads are completely leveled) amounted to only .3 mill per kilowatt hour, a cost reduction of 1 percent (from 3.26 cents to 3.23 cents). Two alternative models, one leveling all individual weeks and the other leveling the entire year, showed smaller savings. The EPRI article concludes that load management "is likely to become commonplace, though not solely for the intended reason of shifting peak demand to save operating or capital costs." It will increase the awareness of buyers and sellers alike that the cost of electric service varies with the time of day and the season. It should also lead to a more equitable rate structure for all consumers.

Supergravity

A major ambition of modern physics is to unite the general theory of relativity, which describes the gravitational interactions of matter, with the quantum theory, which provides a context for dealing with all the other known interactions: the strong, the electromagnetic and the weak. Separately each of the theories must be accounted a magnificent success, but so far all attempts to join them have merely made their incompatibility more apparent. Now a new approach to unification has evolved. It begins by postulating a deep connection between certain categories of elementary particles that had seemed quite disparate.

General relativity and quantum theory have in common the concept of a field, but they differ in how they describe the action of that field. In general relativity the gravitational field is described geometrically: a massive object curves space (or space-time). The quantum theory is more mechanistic: the effect of a field is expressed by the exchange of particles. For example, two electrons can interact electromagnetically by exchanging photons, the quanta of the electromagnetic field.

A strategy for quantizing gravitation has long been apparent: the gravitational field must be "personified" in a quantum particle. From the field equations of general relativity the properties of that particle can be deduced: it must be without mass and it must have an intrinsic spin angular momentum, measured in fundamental units, of 2. Such a massless spin-2 particle is the graviton. It is the hypothetical quantum of the gravitational field, analogous to the photon, which is also massless but has just one unit of angular momentum.

Difficulties appear as soon as one tries to construct a theory with gravitons. When the gravitational interaction between two massive particles is calculated, mathematical infinities appear where the results should be well defined and measurable. Similar problems have been encountered in constructing other quantum theories, but by algebraic manipulation it has always been possible to eliminate them. In theories of quantum gravity the infinities do not go away. The root of the problem is that the strength of the gravitational force on a particle increases in proportion to the mass, or energy, of the particle. Other forces are independent of the mass; electromagnetism, for example, depends only on electric charge, and mass has no effect on the strength of the interaction. In the calculation of quantum-mechanical effects exchanges of high-mass or high-energy particles are the most important ones, and the infinities are therefore much more disruptive in gravitational theories than they are in theories of electromagnetism or other forces.

The new approach to quantum gravity springs from a realm of physics that may seem quite remote from general relativity: the classification of elementary particles according to their spin angular momentum. One of the most fundamental distinctions in nature is that between particles whose spin is measured in half-integral units (such as 1/2, 3/2 and 5/2) and those with integral spin (such as 0, 1 and 2). These two classes of particles are respectively called fermions and bosons. The electron and the hypothetical quarks are fermions with a spin of 1/2; the photon (spin 1) and the graviton (spin 2) are bosons.

What has led to a new gravitational theory is the possibility that a symmetry principle relates fermions and bosons. Since this symmetry connects such disjoint classes of things it has been given the name "supersymmetry."

Any symmetry can be regarded as a set of operations that leaves a system of things effectively unchanged. The most familiar symmetry operations are coordinate transformations, or in other words motions: translation, rotation and reflection. The symmetry operations of the hypothetical supersymmetry are more abstract. They assert that the fermions and bosons in a system of particles could be interchanged without substantially altering the system. Hence particles with integral spin and those with half-integral spin can be regarded as manifestations of a single underlying state of matter.

A gravitational theory can be constructed by applying the concept of supersymmetry to particles with a spin of 2. The theory that is of most immediate interest is one postulating a symmetry between gravitons with a spin of 2 and hypothetical fermions with a spin of 3/2. This theory has been given the name "supergravity"; the spin-3/2 particles might be called "gravitinos."

The possibility of a supersymmetry relating fermions and bosons was first noted six years ago, and a full development of the idea began about two years later. A general connection with gravitation was noted then, but the goal of a concrete supergravitational field theory was elusive. Such a theory was obtained in the spring of 1976 by Daniel Freedman and Peter van Nieuwenhuizen of the State University of New York at Stony Brook, working with Sergio Ferrara of the Frascati National Laboratories near Rome. A simpler reformulation was later found by Stanley Deser of Brandeis University and Bruno Zumino of the European Organization for Nuclear Research in Geneva. More than a dozen theoretical physicists have now taken up the problem, and a number of papers reporting their results have been published in *Physical Review* and *Physical Review Letters,* among other journals.

For all long-range interactions of matter supergravity agrees with the predictions of general relativity; this is a necessary feature of any gravitational theory, since general relativity has been thoroughly tested at long range. The two theories differ in their predictions only on the microscopic scale, as in a calculation of the feeble gravitational interaction between two electrons. The discrepancies at short range result from the introduction of the spin-3/2 fermions. In a supergravitational interaction both ordinary gravitons and these fermions can be exchanged; because the fermions have an inherently short range they are important only when the interacting particles are close together.

The most important property of supergravity, and one that has surprised even those who invented the theory, is that it promises to abolish the troublesome infinities in the calculation of gravitational interactions. Preliminary indications suggest that the infinities may simply disappear. Because of the high degree of symmetry in the theory, the infinities seem to cancel in a "miraculous" way, without any of the explicit programs of correction that have been required in other theories.

The total interaction between two particles is found by adding up the contributions of all possible modes of interaction, including those with "loops" formed when the particles exchange more than one quantum. All theories are finite for interactions without loops. Supergravity has been proved finite for all interactions having one loop or two loops; that is more than has been achieved with any other quantum-mechanical theory of gravitation. Work is now under way on the calculation of interactions with three loops. Ultimately a general proof of finiteness for all numbers of loops must be sought.

Supergravity alone is not a comprehensive theory of nature; it must be extended to include particles with spins of less than 2 or 3/2. A range of possible theories having room for photons and quarks and other elementary particles has been outlined, and the mathematical construction of these theories appears to be feasible. Even when that has been accomplished, however, supergravity will remain at the speculative frontier of physics. The theory may prove to be consistent, but this shows only that it is possible to build a universe according to the premises of supergravity. Only experiment can show that the real world is constructed in this way.

Chain of Events

It might seem improbable that the status of a tiny fish could have major nutritional, economic and political consequences the world over. Yet that is the case with the Peruvian anchovy, *Engraulis ringins,* which until recently thrived in vast numbers along Peru's 1,500-mile coastline, providing the largest fishing catch of any marine species. Today the anchovy is vanishing, and Peru's fishing industry—the mainstay of the country's economy—is threatened with collapse. The crisis appears to have resulted from the combined effects of overfishing and a change in the movement of ocean currents.

The anchovy is familiar to most city dwellers as a garnish on pizza and salads, but by far its most important commercial use is in the form of fish meal as a feed supplement for poultry, cattle and hogs. Processors remove the oil from the anchovy and then cook, dry and grind the fish into a meal that contains 60 percent protein by volume. Until 1972 the 10 million metric tons of Peruvian anchovies caught each year were converted into two million tons of fish meal, accounting for more than a third of the world's fish-meal production and two-thirds of the quantity entering international trade.

The unusually favorable conditions that enable Peruvian anchovies to multiply in such vast numbers have been intensively investigated. In normal years the cold waters of the Humboldt Current run north from the Antarctic along the coast of Peru. Trade winds from the southeast drive the current offshore, whereupon it is replaced by upwelling water from the depths, rich in nitrates and phosphates. These nutrients fertilize the growth of plant plankton, on which the anchovies graze in vast numbers. The fish, which are only some 10 centimeters long when adult, are fast-growing and short-lived. Normally the offspring from the major spawning in September enter the fishery at the beginning of the following year and are caught in significant quantities over the next 18 months.

Periodically the trade winds off the Peruvian coast weaken, and warm currents penetrate south of the Equator. When that happens the anchovy population is drastically reduced for two reasons: the lack of fertile upwelling slows the growth of the anchovies' plankton food source and the warm water interferes with their reproductive cycle. The warm current from the north, whose origin is still a mystery, is known as El Niño (the Christ child) because it typically appears around Christmas, when it is summer in the Southern Hemisphere. El Niño usually occurs at about five-year intervals and lasts for six months, after which the current patterns return to normal and the anchovy population rapidly recovers its former strength.

A typical Niño, accompanied by a sharp fall and recovery of the anchovy population, occurred in 1965. Seven years later, in 1972, the warm currents returned. The first signs were observed in the extreme north of Peru at the end of January; then the Niño extended southward along the entire Peruvian coast and remained longer than ever before—more than a year and a half. One result of the situation was that the fish were concentrated in the coolest areas at a relatively short distance from the coast and were therefore particularly vulnerable to depletion by fishing.

In an effort to protect the very small fish that had been born the preceding fall the Peruvian government closed the fishery until March of 1972. When fishing resumed, good catches of older fish

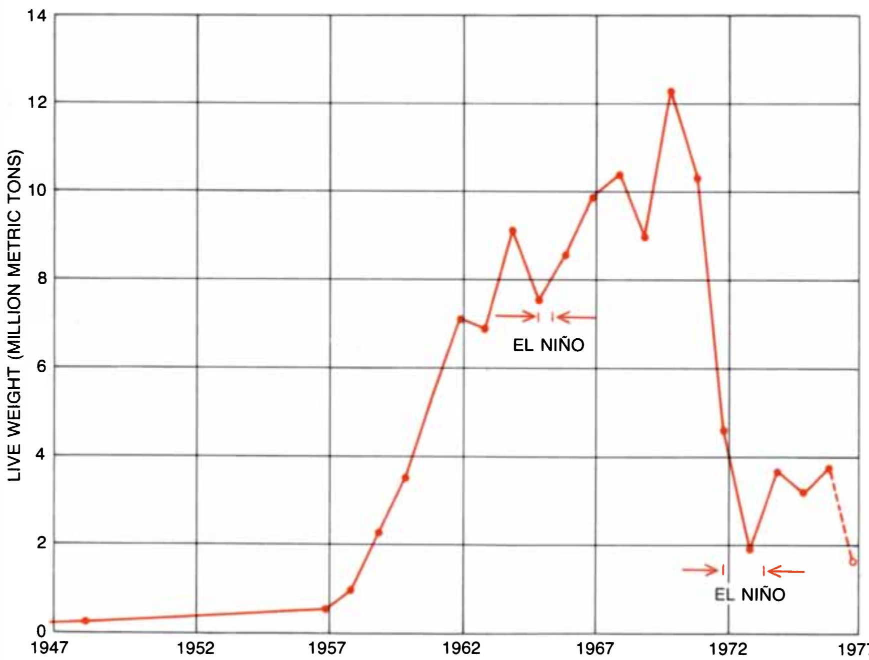

Minolta by Minolta by Irving Penn.
SRT 202
minolta
MINOLTA
LENS MADE IN JAPAN

(born in 1970) were taken, but it soon became clear that very few fish from the 1971 spawning had survived. Catches declined dramatically in April and May as the surviving older fish were caught, and the government closed the fishery again at the end of June. As a result the anchovy catch, which had reached a record level of 12.4 million metric tons in 1970, plummeted to 4.4 million tons in 1972. The following year short periods of fishing were allowed in March and April, when 1.7 million tons were taken. Studies during this period showed that the stock of spawning fish was still at a low level and that the young fish, born in 1972, were again very scarce. In March, 1973, the Peruvian Sea Institute expressed grave concern over the dwindling stock of adult anchovies and recommended that full-scale fishing be halted until the stock had recovered.

The effects of the sudden anchovy decline were dramatic and far-reaching. Peru's entire fishing industry, in some cases far overextended on credit, went into bankruptcy and had to be nationalized under a government monopoly named Pesca Peru. Forty percent of the country's foreign-exchange earnings were lost, and the Peruvian economy suffered a heavy blow. In the U.S., West Germany and the Netherlands, livestock producers found almost half of their customary feedstuffs eliminated and began heavy buying of soybean meal and grain. The sudden demand for soybeans drove up the world price, with widespread economic and political repercussions. In the U.S. the heavy buying into the already tight grain supply helped to send wheat prices soaring. Since 1973 increased soybean production has lowered prices somewhat, but the economic effects of the fish-meal shortage are still being felt.

When El Niño departed in mid-1973, there was widespread hope that the anchovy population would return to its normal size, but that did not happen. The annual yield rose slightly to 3.5 million tons in 1974 and did not increase through 1976, mostly because of warm currents from the Pacific (not El Niño) that appeared at different stages of the season. This year, according to Carlos Mendoza, the Peruvian fish-meal attaché to the U.S., fishing was allowed until mid-May, with a catch of 1.2 million tons. Then the fishery was closed because of continued declines in the anchovy population. Fishing will probably not be resumed until fall.

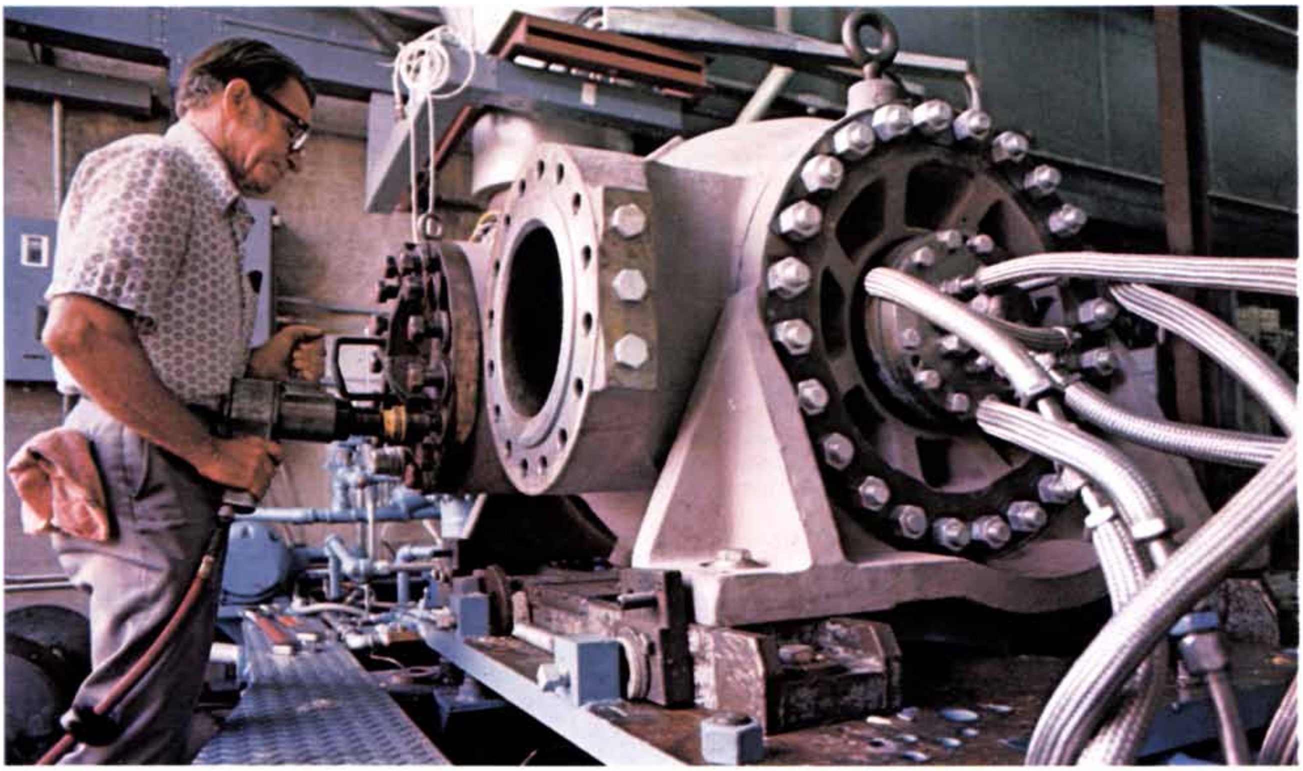
No one really knows why the anchovies did not return to their normal strength after the 1972–73 Niño, or why the broods born in 1971 and 1972 failed. A report issued by the United Nations Food and Agriculture Organization's Advisory Committee of Experts on Marine Resources Research states that a combination of factors was most likely responsible for the anchovy decline. Moderately low levels of adult stocks resulting from heavy fishing can still give rise to adequate numbers of offspring under normal or favorable conditions, but a combination of overfishing and high water temperature may have weakened the anchovy population to such an extent that it was unable to fully recover. The decision of the Peruvian government to keep catches at the lowest possible levels during the present crisis may succeed in keeping the Peruvian anchovy from following into commercial extinction the California sardine and the Norwegian herring.



Rapid rise and equally rapid decline of the Peruvian anchovy catch from 1947 to the present

DP Dialogue

Notes and observations from IBM that may prove of interest to the engineering community.



Gas turbine-driven compressors must be custom-configured for each application. Engineers at Solar use an IBM computer interactively to find optimum specifications in minutes.

Solar Finds Best Pipeline Design with TSO

"When a customer asks us to recommend a compressor, one of our engineers enters the details at a terminal, and the computer immediately selects several suitable configurations. A computer-driven plotter then draws performance curves for each one, so the customer can compare them graphically."

Douglas McKerrow, manager of systems analysis for the Solar Group of International Harvester at San Diego, is describing Solar's use of Time Sharing Option (TSO). This IBM facility allows engineers to interact directly with the company's System/370 Model 158, operating their application programs through six visual display terminals.

Among other products, Solar makes turbine compressor systems which propel natural gas through cross-country pipelines. Each system is individually configured to meet customer needs. Taking into account all possible variations in components, Solar can build an eight-stage system in as many as 9,000 different configurations.

"Before we used TSO and interactive computing," McKerrow continues, "we applied a customer's specifications manually to formulas and sets of tables—and came back to him several days later with one configuration for his job. Now we give him a choice."

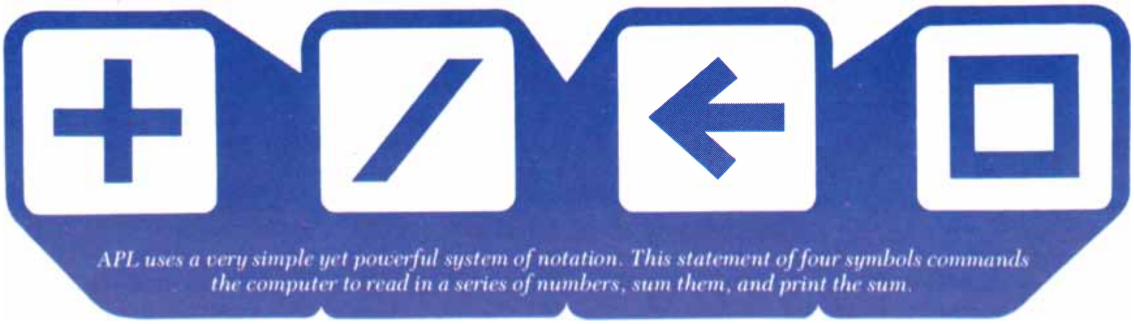
Solar's interactive facility goes far beyond that. The company can find the best design for an entire gas pipeline of which their compressor systems will be a part.

A customer describes in detail what the pipeline is intended to do. Then an engineer at a terminal interacts with the computer to reach an optimum solution. A pipeline is complex, with a vast range of possible tradeoffs. For example, the pipe diameter and the spacing between compressors affects the pressure rise at each compressor. Finding the best balance between these two factors is just one such tradeoff.

"Only a few years ago," McKerrow explains, "it took pipeline designers eight or nine months to achieve a satisfactory result. With interactive computing, we can specify a better engineered, more economical pipeline in two or three weeks.

"Moreover, we're constantly improving the quality of our work with interactive computing. Today, we use six times the data we used two years ago, which makes possible even greater precision in our recommendations.

"We have built a powerful selling tool using interactive computing—one which helps our customers use Solar products better."



APL uses a very simple yet powerful system of notation. This statement of four symbols commands the computer to read in a series of numbers, sum them, and print the sum.

APL: A Flexible Problem-solver for Engineers

With APL, an interactive language from IBM, engineers can interact directly with the computer, entering a problem at a terminal and receiving prompt results. APL suits very small calculations—including those "one-time" solutions usually approached manually—as well as large computations such as heat transfer and structural analysis.

Engineers using APL have experienced major gains in productivity; applications have often been developed in one third or less the time required by conventional methods, and at corresponding savings in cost.

The user expresses his problem in familiar language, interacting creatively with the computer in the search for a solution. Tentative results can be presented graphically at an alphanumeric terminal, for quick evaluation, using facilities within APL for generating curves and charts.

The language uses standard mathematical notation in a concise form, capable of specifying extensive procedures in brief statements. For example, consider some vector arithmetic. If A, B and C are each a sequence of numbers (a vector) of 50 elements, which has previously been entered, then typing:

$$D \leftarrow 1.125 \times A + B + C$$

at the terminal causes the computer to add the three vectors together element by element, multiply each element of the sum vector by 1.125, and assign the resulting 50-element vector to D.

An APL user can be given controlled access to any data in the system, including APL and non-APL files. The language can readily be added to any System/370. It is quickly learned, becoming productive almost immediately. Economically priced, APL can pay for itself many times in faster and better engineering solutions.

ATMS Gets Specifications Out Faster for Ferguson

"ATMS saves us at least half the manhours we would otherwise need to produce 150 pages a day of construction specifications," says Robert C. Lord, Sr.

Lord manages the specification-writing department of the H.K. Ferguson Company, Cleveland, where he uses IBM's Advanced Text Management System (ATMS), which helps users prepare many kinds of business and technical documents.

One of the 20 largest industrial plant construction companies in the U.S., Ferguson specializes in building plants for heavy manufacturing, chemical and food processing companies.

"Each process plant we build requires massive documentation to spell out materials and methods for fabricating buildings, piping, wiring and so forth," Lord says. "A project will typically involve 75 to 100 separate specifications, averaging 10 to 15 pages each.

"We keep the entire text of each specification in ATMS," Lord notes. "To enter revisions, the writer works at an online IBM terminal, using the ATMS editing facilities. Our System/370 Model 125 automatically takes care of any necessary renumbering of paragraphs and pages, and then prints out the entire revised specification at a high-speed printer. We don't need to key in and proofread any parts of the text which were not altered. Since we average three or four revisions of each specification, this means major manhour savings."

"We store generalized master specifications in ATMS which include alternative wording covering most requirements," explains Robert V. Toensing, manager of computer facilities. "The specification writer marks up copy by hand, crossing out inapplicable wording, adding material and entering variable data for the specific project. Then a clerk enters these notations at the terminal; ATMS applies them to the stored master text and generates printed text."

"This process is very rapid," Lord adds. "We produce 150 pages a day at a terminal during a normal eight-hour day.

"And the use of a master specification gives us automatic quality control, since the writers select existing wording rather than write from scratch.

"ATMS has meant great strides in both the quality and productivity of specification writing at Ferguson."

Before building a process plant, H.K. Ferguson Company fabricates a detailed model such as this one of a chlor-alkali plant. Such models facilitate planning the piping layout and serve as a guide during construction.

Computer Guides Oil Exploration Teams To Pay Dirt

Drilling for oil is expensive. It can cost $60,000 a day to keep an offshore drilling platform in operation. Once a rig is in place and drilling has started, the crew needs guidance fast. Should they continue? Should they drill straight down or deflect the hole? What is the next move?

At Amoco Production Company Research, scientists and engineers use interactive computing on an IBM System/370 Model 158, under the Virtual Machine/Conversational Monitor System (VM/CMS) for better insights into oil exploration. Well logging and seismic analysis are two exploratory tools in which the computer plays an important part.

Well logs are vital to evaluate what is found in a drill hole, to determine its value or where to drill next based on that hole. "Without well logs, we just have a hole in the ground. With them, we gather vital data about the rocks, fluids, and geological conditions found in that well-bore," says Michael Waller, vice president, research and director of the Amoco Research Center, Tulsa, Oklahoma.

It has been difficult to clarify visually the irregular contours of subsurface geology. Here Amoco scientists examine a three-dimensional model created on multiple sheets of clear acetate.

Sensors measure earth composition—density, gamma ray intensity, porosity, resistivity, and water saturation—and readings are logged at the surface.

Correlation analysis of these variables can then describe the nature of the subsurface geology and the presence of hydrocarbons.

"With so many variables," Waller explains, "there are an enormous number of correlations we could plot. That's why we developed Inlan—Interactive Log Analysis. It permits a geologist working interactively at a computer terminal to explore possibilities and zero in on the most promising."

The computer is also used in the exploration process before drilling begins. Seismic excitation—a shock wave from an explosive charge or a mechanical "thumper"—is applied to the earth. Reflections from the interfaces between underground geological formations are detected by an array of sensors distributed over the exploration site.

Converting these readings into accurate three-dimensional maps of the subsurface configuration is a complex task. Amoco scientists use interactive techniques and build successive computer models of the geological layers as they expect them to be. They also calculate the sonic reflection path and compare it to the field recording, thus working toward a convergence of the field data with their idea of structure shape.

James G. Steward, manager of computing research, directs a professional group which studies the application of the computer to oil production, conducts program development, and supports the interactive computing environment which gives scientists at the center direct access to computing power through terminals.

"With 300 people authorized to use the terminals," Steward says, "we now average over 400 'log-ons' a day. As many as 70 people may be online simultaneously. The beauty of the Conversational Monitor System for us is that our people communicate with the system in the terms of their science, not in computer talk. People find that they can often develop computer programs or test scientific ideas in a few minutes with interactive computing. It might have taken a week with conventional approaches—or would not have been possible at all."

Software Products from IBM

Three products available from IBM fill important needs for engineers and scientists:

1. General Purpose Simulation System V (GPSS V) Models any physical or logical system which can be represented as a series of discrete stages, processing points or transaction nodes.

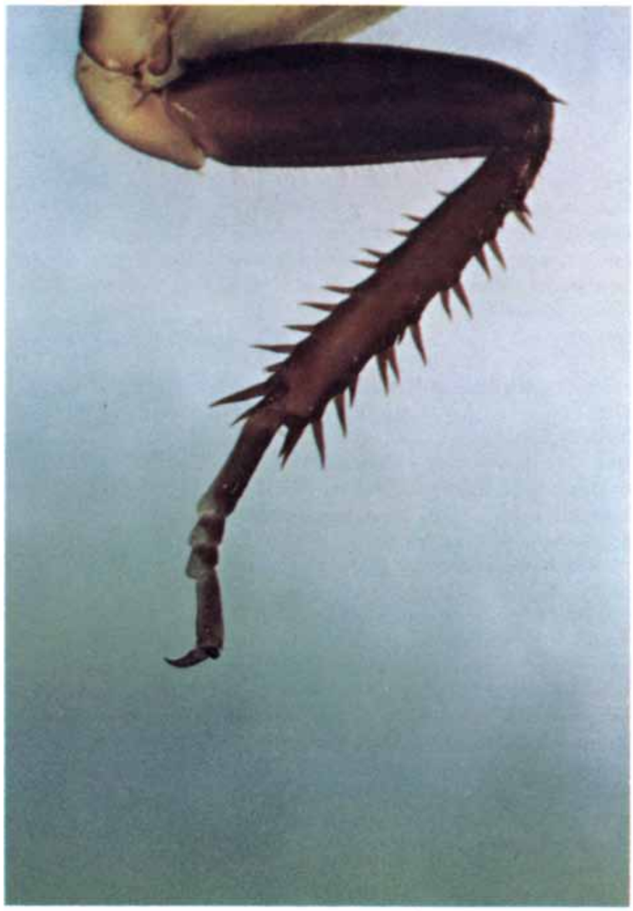
2. Continuous System Modeling Program III (CSMP III) Models a continuous process expressed in the form of either an analog block diagram or a system of ordinary differential equations.

3. Graphic Analysis of Three-Dimensional Data (GATD) Graphically represents and analyzes three-dimensional data.

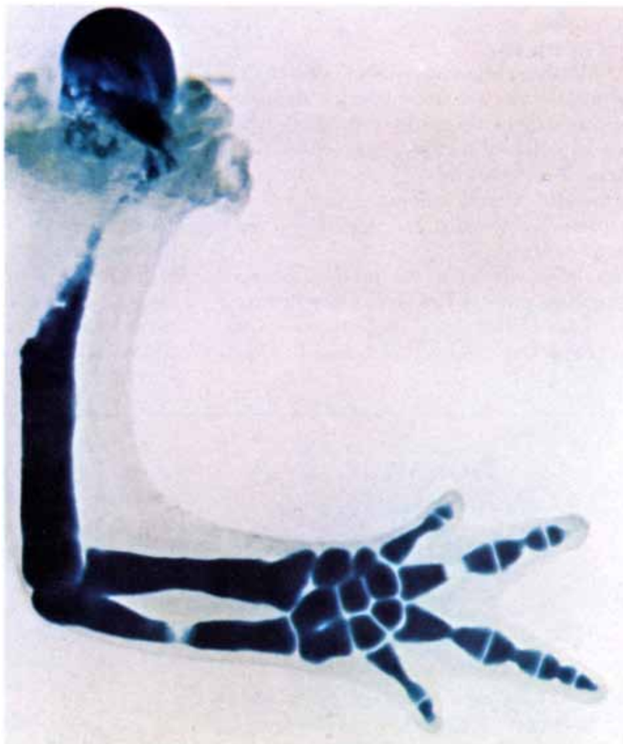
For more information on these products, contact your local IBM branch office or write to the Editor of DP Dialogue at the address on the right.

DP Dialogue is designed to provide you with useful information about data processing applications, concepts and techniques. For more information about IBM products or services, contact your local IBM branch office, or write Editor, DP Dialogue, IBM Data Processing Division, White Plains, N.Y. 10604.

IBM®
Data Processing Division



REGENERATION OF COCKROACH LEG is documented. A normal leg like the one at left was amputated between the femur and the trochanter. The leg regenerated by the next molt (*right*). It is identical with a normal leg except that one tarsal unit is missing.



REGENERATION OF NEWT FORELIMB is apparent because a stain has been applied that turns cartilage darker than bone. The leg at left is normal, with cartilage confined to jointed areas. When, as in this instance, a newt leg is amputated through the humerus, it will regenerate the amputated portion over a six-week period (*right*). At first regenerated skeleton consists of dark-staining cartilage alone.

Biological Regeneration and Pattern Formation

Certain animals can grow new legs and even extra ones. Studies of how they do so are revealing much about the basic principles of the organization and growth of complex structures in animals

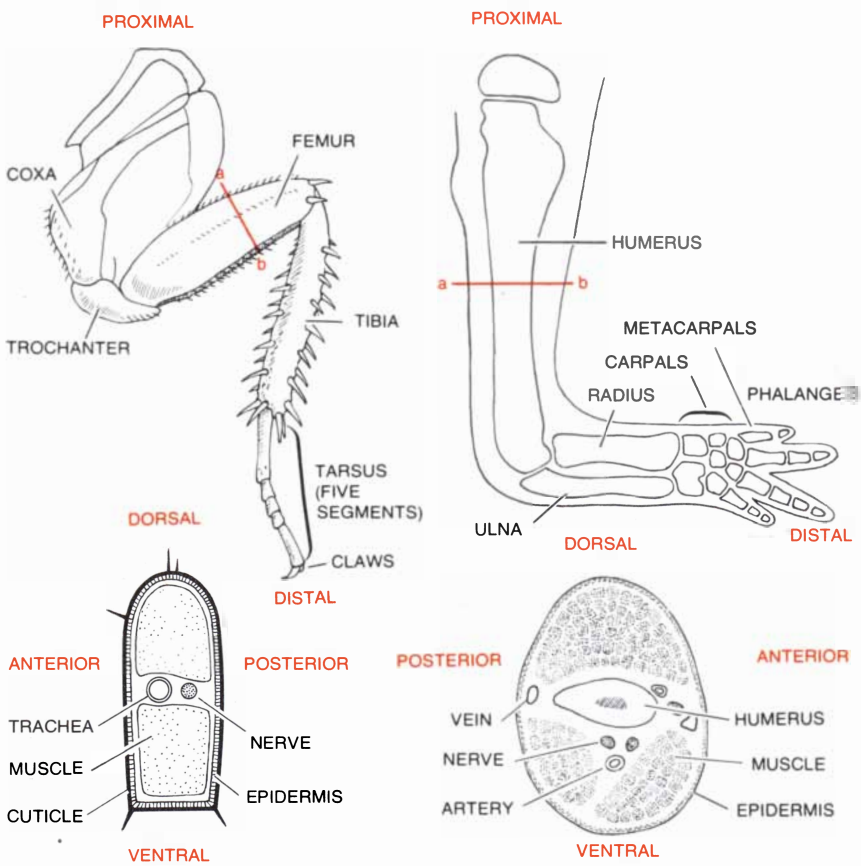
by Peter J. Bryant, Susan V. Bryant and Vernon French

One of the least understood phenomena of animal life is that in every generation the union of an egg cell and a sperm cell leads to the formation of an embryo endowed with a predictable spatial organization. This is to say that embryonic cells differentiate into distinct cell types, arranged in recognizable and reproducible patterns to form specific tissues and organs. How these spatial patterns are formed is one of the great enigmas of developmental biology, but embryologists are now beginning to realize that the process must involve mechanisms whereby cells can assess their physical position within developing cell populations. The idea that cells acquire positional information and respond to it by differentiating in appropriate ways has been formulated largely by Lewis Wolpert of the Middlesex Hospital Medical School in London.

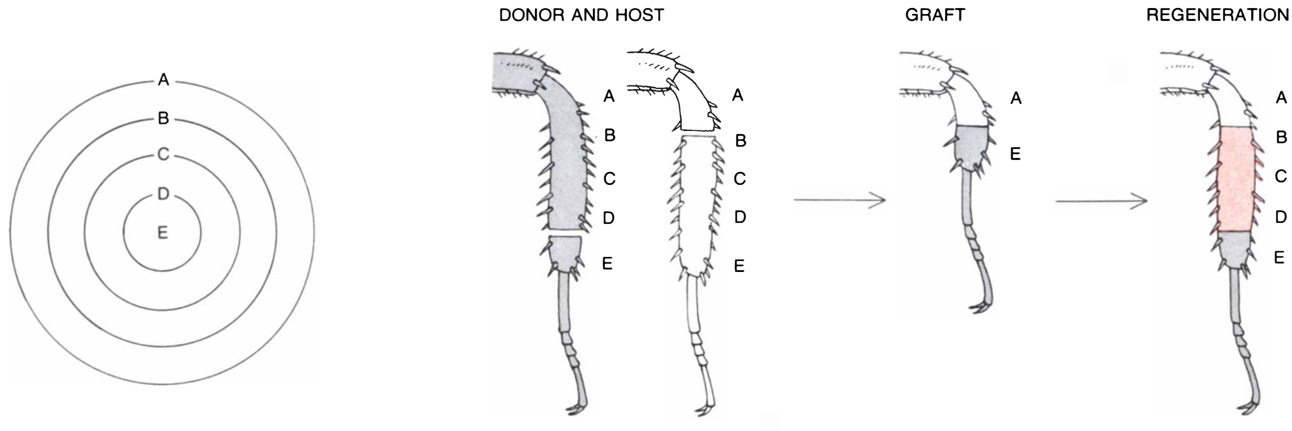
A clue to the way in which spatial organization develops is provided by the finding that developing animals are subdivided into "fields," regions that are to a certain extent developmentally autonomous. The concept of fields was developed mainly by Paul A. Weiss. An embryonic field can be defined as a region within which the fate of different areas can be changed by surgical intervention; in other words, fields are regions within which embryonic regulation can occur. For example, as Hans A. E. Driesch demonstrated late in the 19th century with the sea urchin, parts of the early embryo of various animals can be removed and the remaining parts will embryonically regulate to form a normal whole. Hence the entire early embryo can be classified as a single primary field.

As development proceeds, however, surgical intervention has much more localized effects. For example, in the early part of this century Ross G. Harrison showed what happens in early amphibian embryos when all the region destined to form the forelimb and shoulder is removed from one embryo and grafted onto a second embryo. At the end of embryonic development the donor embryo will lack one of its four limbs and the host embryo will have grown a fifth limb from the graft. If only a part of the donor's limb region is removed, however, the remaining tissue will regulate and the donor will develop its usual number of limbs.

Studies of this kind demonstrate that the limb region has become independent of the rest of the embryo. Such independent regions are known as secondary fields. Within a secondary field the consequences of surgical manipulation are limited to that field. Weiss recognized that as development proceeds primary fields are replaced by more numerous secondary fields. In vertebrate animals secondary embryonic fields include re-



COMPARATIVE ANATOMY of a cockroach leg and a newt leg is outlined to facilitate comparisons with the legs in the photographs on the opposite page. Both of the legs are shown from above and then in cross section along the *a-b* axis indicated (*color*). The most distal (farthest from the body) of the major cockroach leg segments, the tarsus, normally consists of five units, as shown. Newt forelimb shows typical amphibian bone sequence from humerus to phalanges.



ORGANIZATION OF A LEG SEGMENT in the cockroach, as observed in the two experiments shown in this illustration, suggests an ordered array. The concentric circles diagram such an array along one axis, from *A*, the most proximal level (closest to the body), to *E*, the most distal level. When, in the first experiment, an *E*-level portion of a tibia is removed from a donor and is grafted onto the *A*-level portion of the stump of a tibia on a host, local growth at the graft junction produces intercalary regeneration of the missing levels of

gions destined to form the limbs, the heart, the ears and the eyes. In insects secondary fields include regions destined to become the wings and the legs. The characteristic properties of secondary fields persist in some animals even after embryonic development, enabling portions of the body to regenerate. This is the case in some insects during their larval stages, and it is true throughout the life of some crustaceans, worms and amphibians.

In a 1901 treatise on regeneration Thomas Hunt Morgan recognized two basic responses to the removal of parts of animals. In one type of response the remaining portion of the field is remodeled to form a complete but miniature pattern. This process, termed morphallaxis, does not require the proliferation of cells. In the other type of response, termed epimorphosis, cells do proliferate, and new pattern elements are added on during the growth of the tissue. As a general rule most primary fields, such as early embryos, appear to regulate by morphallaxis whereas secondary fields do so by epimorphosis.

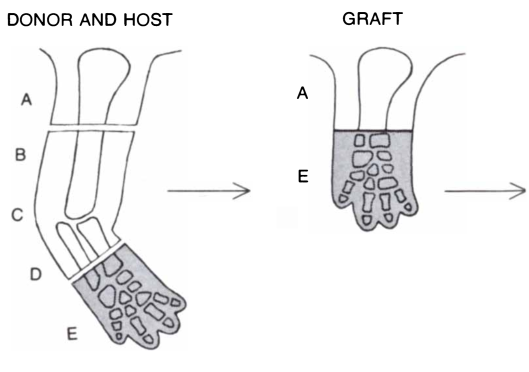
The final pattern of the embryo, then, develops from an initial primary field, typical of the egg and the early embryo, through the establishment of secondary fields by groups of cells at certain positions in the primary field. Here we shall be discussing some ways whereby secondary fields might be initiated in the early embryo, but most of what we can deduce about the process comes from studies of much later stages of development and involves the manipulation of established secondary fields.

Let us consider in two animals, the cockroach and the newt, the remarkable capacity of being able to regenerate a lost leg. In the insect leg we shall be concerned with the pattern of the cuticle, or horny outer covering, which is secreted by a single layer of epidermal cells. In the amphibian leg we shall consider the pattern of bones derived from the mesodermal cells of the leg. These appendages are quite dissimilar but nonetheless show great similarities in their responses to a variety of amputation and grafting experiments. By analyzing these responses we have developed a simple model to account for pattern regulation in such epimorphic fields. We shall also show how this model accounts for pattern regulation in the fruit fly *Drosophila*.

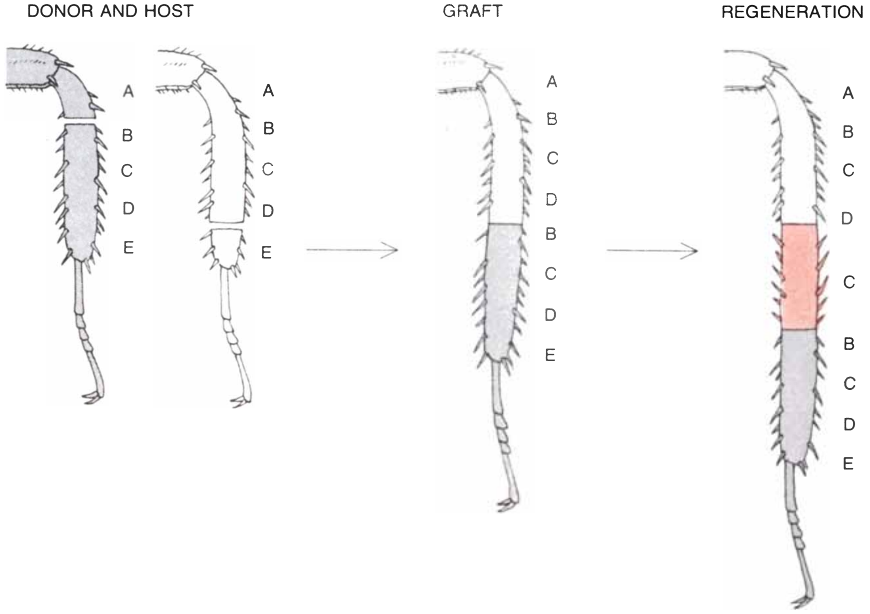
Leg Regeneration

When the leg of a cockroach or a newt is amputated at any point, the wound first heals over. The cells at the end of the stump dedifferentiate, or lose their characteristic differentiated appearance, and accumulate to form the clump of tissue known as a regeneration blastema. The blastema grows, and from it are formed all the structures that normally lie distal to the point of amputation. ("Distal" refers to farther from the body, "proximal" to closer.) The regenerate and the stump together form a complete and functional leg.

Many workers have shown that such structures can also regenerate from the cut surface of the amputated leg. For example, if a portion of the leg is grafted back onto the animal with its proximal-distal axis reversed, thereby allowing the distal parts that would otherwise be discarded at amputation to survive, regeneration from the cut end forms all the structures normally distal to that point. The result is a mirror-image duplicate of the distal parts that have been grafted in reversed orientation. With a different grafting operation Desiré Bullière of the University of Science and Medicine in Grenoble and one of us (French) have shown that if a graft junction in a cockroach leg does not heal



ORGANIZATION OF A LEG in the newt, tested in two experiments, resembles the leg-segment organization of the cockroach at the top of these two pages; the proximal-distal



organization (in this example *B*, *C* and *D*). When, in the second experiment, a proximal-level portion of a tibia from a donor is grafted onto a distal-level portion of a tibia on a host, the confrontation, although it still induces intercalation, gives rise to a reversed intermediate organization level (in this example *C* between *D* and *B*) where bristles grow up rather than down.

properly, both the host component and the donor component can regenerate distal structures. The result is an elongated leg that has three sets of distal parts.

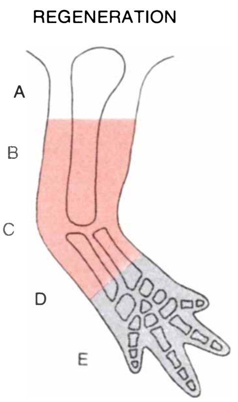
In the 1920's Hans Przibram of the University of Vienna suggested that the bizarre multiple legs insects and amphibians sometimes grow spontaneously may result from a similar process. In Przibram's view a partly broken leg would regenerate extra copies of the portion of the leg beyond the break from each of the surfaces exposed by the break. Since the portion of the original leg beyond the break did not fall off, two supernumerary distal parts are forced to grow out of the side of the original leg.

Proximal-Distal Interactions

By discovering the rules governing the behavior of cells from different regions of a field when they are made to interact, we have been able to find out much more about the regeneration and spatial organization of cockroach and newt legs. For example, both Horst Bohn of the University of Munich and Bullière have shown that when different proximal-distal levels of a cockroach leg segment, say the tibia, are grafted together, there is localized growth at the graft junction. The new tissue forms the structures found in the normal leg between the proximal-distal levels of the stump and the graft. This type of tissue growth is termed intercalary regeneration.

A graft between a proximal level of a host cockroach leg and a distal level of a donor leg gives rise to a normally oriented intercalary regenerate and forms a normal leg segment. What about the reverse combination? Bohn and Bullière have grafted donor legs amputated at a proximal level onto distal-level host stumps. The same localized growth results and an intercalary regenerate is formed. Under these circumstances, however, the new structures show a reversed proximal-distal orientation. Although there are no discontinuities in the resulting segment, it is far from normal. Both investigators have also shown that the proximal-distal organization is repeated in each segment of the cockroach leg. For example, grafting the middle part of a tibia onto the middle part of a femur does not lead to the intercalary regeneration of the missing parts of the tibia and the femur. Grafting the proximal part of a tibia onto the distal part of a femur, on the other hand, results in the formation of an intercalary regenerate that is approximately half the length of a normal segment.

One of us (Susan Bryant), working in collaboration with Laurie E. Iten and David L. Stocum, has performed similar experiments on the limbs of newts and salamanders by grafting regeneration blastemas, formed after amputa-



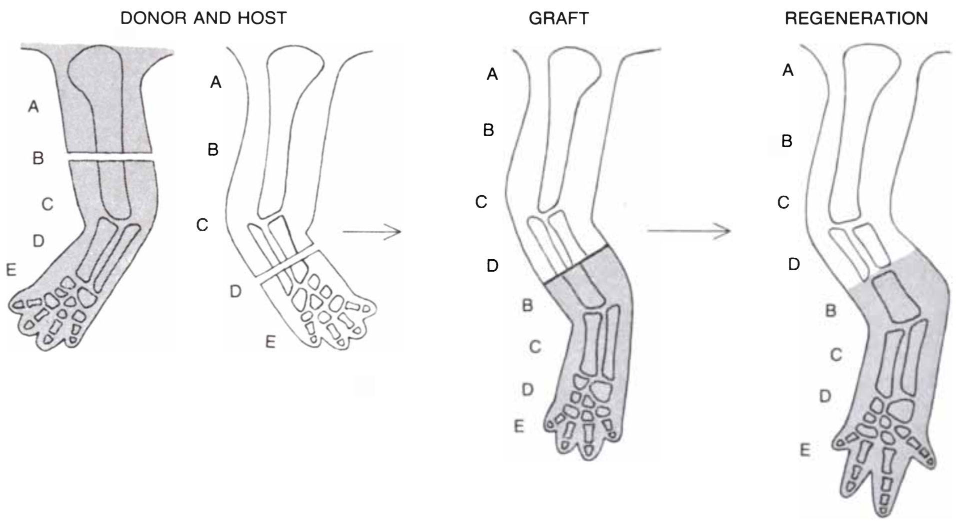
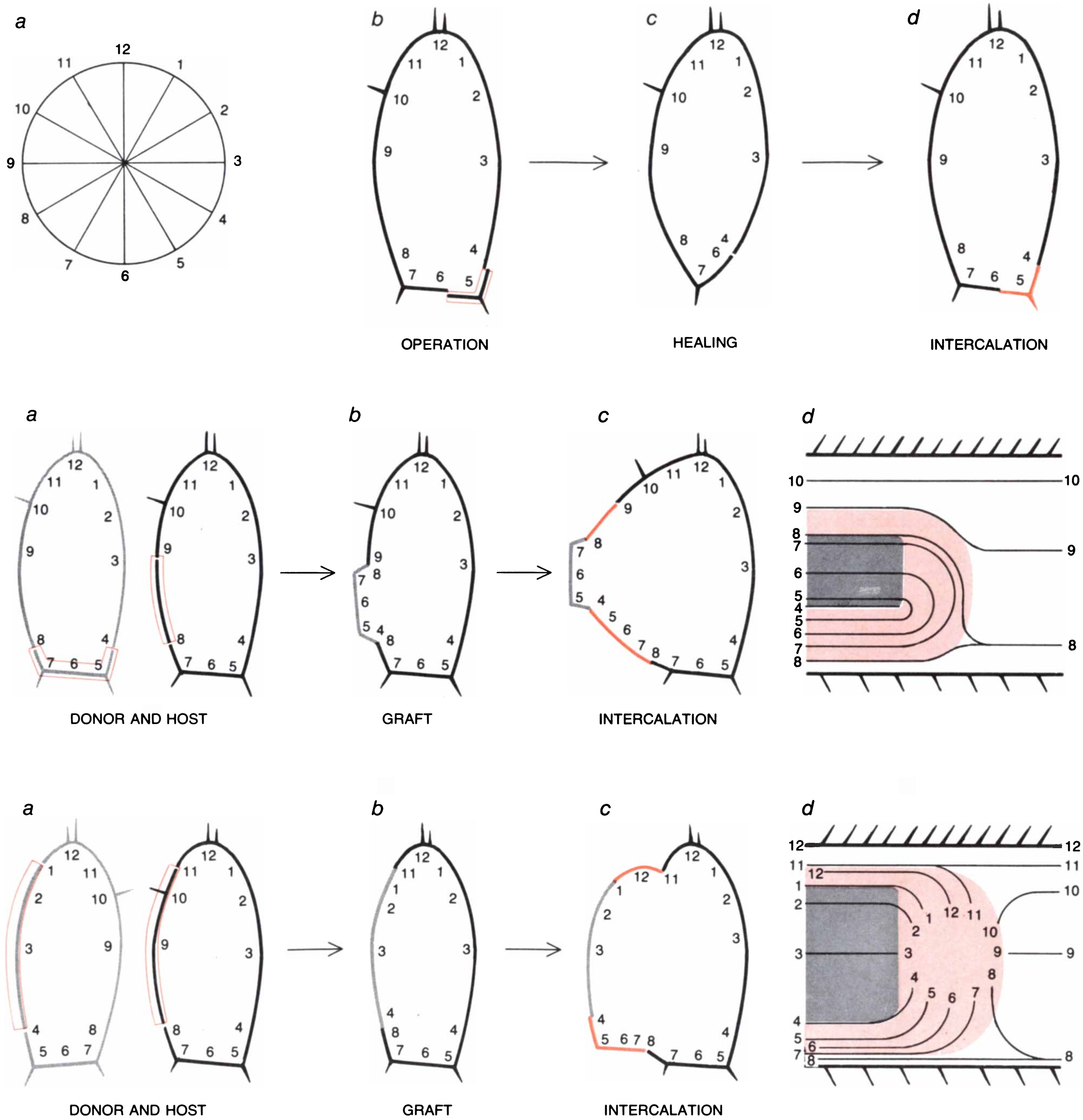


diagram applies, however, to the entire leg and not to each segment. When, as at the left, intermediate organization levels are surgically removed and a regenerating hand is grafted onto the middle of an upper arm, connecting an *E*-level distal portion to the *A*-level stump, intercalary regeneration at the grafting point reproduces the missing levels. When, however, as at the right, a regenerating leg amputated at a proximal level is grafted onto a distal-level stump, the intervening level between *B* and *D* is not intercalated, as in cockroach tibia.

tion at one proximal-distal level, to host limb stumps amputated at a different level. It was found that grafting a forearm blastema onto an upper-arm stump led to the growth and intercalary regeneration of the missing intermediate levels. For some unknown reason the reciprocal graft of an upper-arm blastema onto a forearm stump does not lead to intercalary regeneration. The resulting limb exhibits a discontinuous pattern.

Taken together, these experimental results indicate that cells at different proximal-distal levels of the cockroach leg segment and the regenerating amphibian limb possess different positional information. Furthermore, the interaction of cells with such different positional values leads to a local stimulation of growth and the intercalation of intermediate values within the new tissue. A



ORGANIZATION OF A COCKROACH LEG around a second, meridional axis is revealed in the experiments shown in this illustration; the numbers around the circumference of the diagram identify organizational zones. Circumferential regeneration intercalates the missing positional values. At the top, following the surgical removal of organization Zone 5 (*b, color*) the cells of the adjacent zones 4 and 6 heal together (*c*). Within two molts intercalary regeneration (*d*) replaces the extirpated circumferential zone with its characteristic bristle row. In the middle, when a donor graft with positional values from 4 to 8 is made to a host between positional values 8 and 9 (*a, b*), appropriate intercalary regeneration appears on both sides of the graft (*c*); the intercalation is shown diagrammatically in lateral view (*d*). At the bottom a donor graft to a host (*a, b*) replaces the circumferential values between 8 and 11 with values 1 through 4. Intercalary growth (*c*) on both sides of graft results in formation of a complete circumference at the ends of the graft; lateral view (*d*) shows the distal circumference diagrammatically. An intercalated circumference is equivalent to a circumference formed by amputation and gives rise to the same regenerative response: the production of whatever leg structure would normally exist distal to the circumference.

logical extension of these experiments is to perform grafts that confront cells from the same proximal-distal level but different circumferential positions.

Transverse Interactions

One of us (French) has conducted such experiments with cockroach legs, initially in collaboration with Bullière. For example, the removal of a narrow longitudinal strip of integument (both epidermis and cuticle) from a cockroach femur results in a healing of the wound and a confrontation between two groups of epidermal cells that had originally been some distance apart around the circumference of the leg. The interaction stimulates localized growth and intercalary regeneration; the missing intermediate structures are created, and the leg regains its normal circumference. This result is the same regardless of where on the circumference of the leg the strip of tissue is removed.

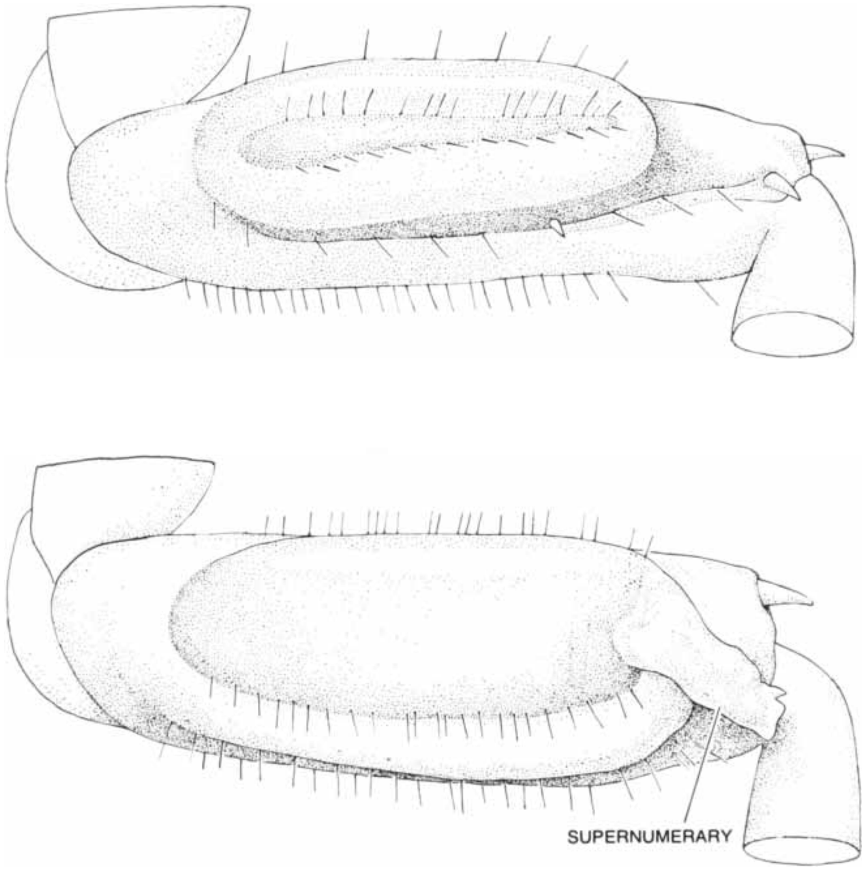
In a variation on this experiment a rectangular portion of integument is taken from a donor leg and grafted into an abnormal circumferential position on a host leg without altering its proximal-distal position. Epidermal cells from different circumferential positions again confront one another, now along both edges of the graft. This stimulates localized growth at the junctions; the structures that normally lie between the host and graft positions are intercalated in the new tissue. The intercalated structures, however, are only those that are normally located on the shorter of the two possible routes around the leg circumference. If such a graft confronts host and donor cells from diametrically opposed positions on the circumference of the leg, the subsequent intercalation of tissue may form either of the normally intervening half-circumferences.

Even when strip grafts of this kind are performed between different segments of the cockroach leg, or between different proximal-distal levels within the same leg segment, the results are similar. This strongly suggests that cells differ in their positional values according to their circumferential positions, regardless of proximal-distal level. As with grafts that affect the proximal-distal axis of the leg segment, the interaction of cells with different positional values around the circumference of the leg segment leads to a local stimulation of growth and the intercalation of the intermediate positional values.

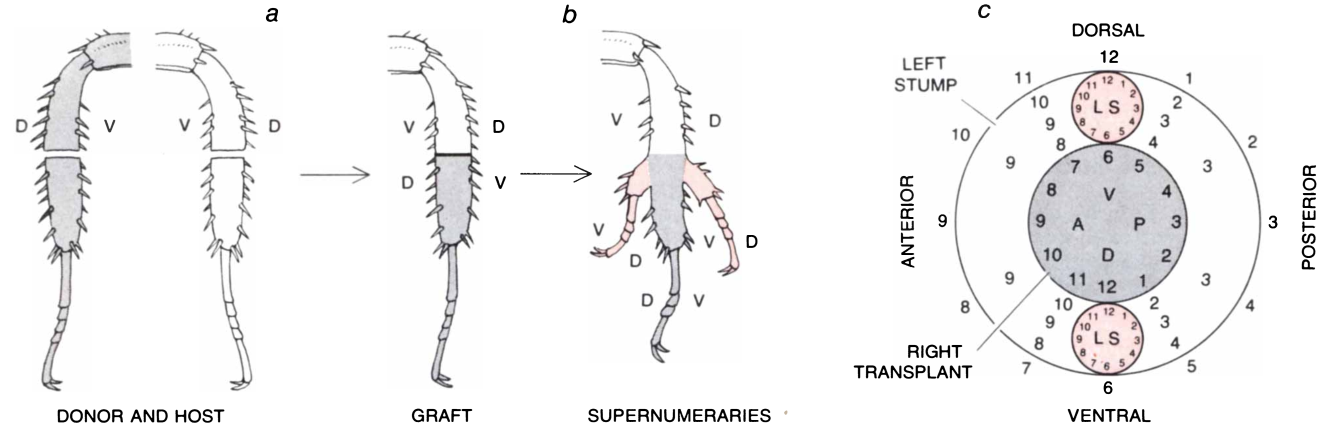
Supernumerary Structures

Many grafting operations on cockroach and newt limbs give rise to bizarre multiple regenerates. A study of these supernumerary structures provides further information about the rules for cellular interactions in the leg. Work along these lines has been done with the cockroach by Bohn, Bullière and one of us (French) and with the newt by Iten and another of us (Susan Bryant). As an example, when a longitudinal strip of cockroach leg integument is grafted into a host at an abnormal circumferential position, a supernumerary leg will be formed at each end of the graft in some combinations but not in others. Analysis of these results shows that the supernumerary legs are produced only when the end of the graft, the adjacent host tissue and the new tissue intercalated between the host and the graft form a complete circumference. This observation suggests an important conclusion: Distal structures will regenerate from a complete circumference not only at an amputation site but also at any other site where a complete circumference can form. Furthermore, experiments suggest that a complete circumference is a necessary prerequisite for distal regeneration. For example, when Bohn amputated cockroach legs in such a way that the cut end consisted of two mirror-image copies of half of the circumference and was lacking the other half of the circumference, he found either no regeneration or incomplete regeneration. We have recently done similar experiments on both cockroach and newt limbs and have obtained similar results.

Regeneration from complete circumferences explains the supernumerary legs that appear after other kinds of grafting operations on cockroach and newt limbs. When an amputated cockroach left leg, or the regeneration blastema from the stump of a newt left leg, is grafted onto the stump of a right leg, the switch from left to right reverses one of the transverse axes of the graft with respect to the transverse axis of the stump. It has been found that after a period of time two supernumerary sets of distal parts arise from the junction between the stump and the graft. If it is the anterior-posterior axis that has been reversed in grafting, then the supernumerary leg parts arise anterior and posterior to the junction. If instead it is the dorsal-ventral axis that has been reversed, the supernumerary leg parts arise on the dorsal and ventral sides of the junction. It is characteristic of these supernumerary regenerates in all instances where "handedness" can be determined that



ZONES OF INTERCALARY GROWTH on the circumference of cockroach legs appear in these drawings. These relate to the two diagrammatic lateral views of intercalary regeneration shown in the illustration on the opposite page. In the upper drawing intercalary regeneration appears on both sides of graft. In lower drawing, where intercalary growth has resulted in formation of a complete circumference, a supernumerary limb distal to femur has developed.



SUPERNUMERARY LEGS appear in many cockroaches when grafting experiments lead to abnormal cell confrontations. In the experiment at left a distal half of a donor right tibia is grafted onto the stump of the proximal half of a host left tibia (*a*). This places host and donor cells with different circumferential values in opposition to one another at all but two points. Two supernumerary legs appear (*b*), one at each of the two points of maximum incongruity. The supernumerary legs are mirror images of the donor tarsus, and they form because intercalary growth in the graft junction has given rise to complete circumferences at positions of maximum incongruity be-

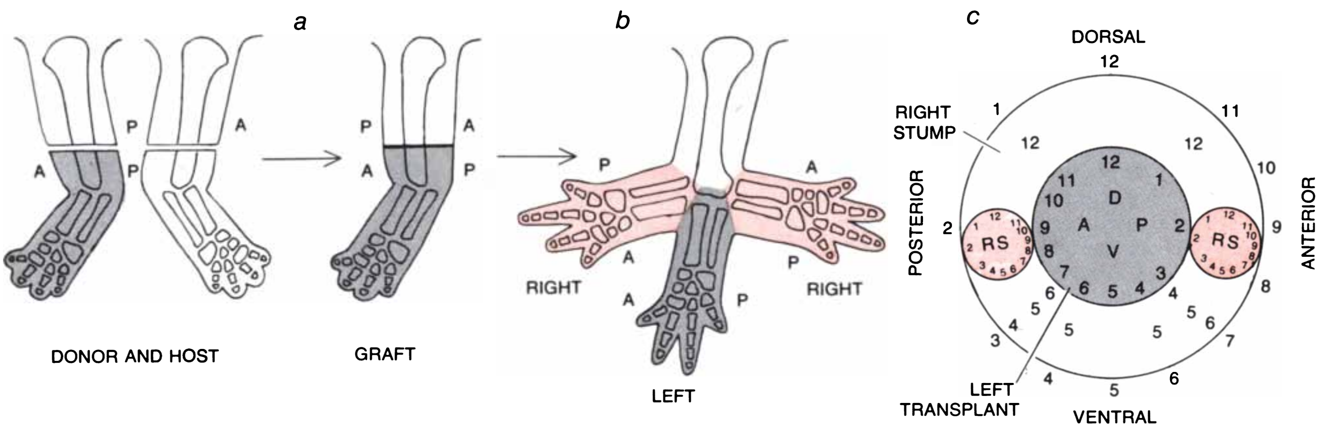
they exhibit the same handedness and orientation as the stump. This type of graft confronts stump cells and graft cells at all points around the graft-host junction, and only at two places will the stump cells and the graft cells have corresponding positional values. Everywhere else the cells will differ in their positional values, and their interaction will lead to the intercalation of cells that have the shorter set of intermediate values. The result is the formation of a complete circumference at each of the two points of maximum incongruity between stump and graft. Regeneration from these complete circumferences will lead to the formation of supernumerary limbs in the observed positions and with the observed orientation and handedness.

When an amputated cockroach leg is grafted onto its own stump after having been rotated 180 degrees, either no supernumerary legs form or two of them form at various positions around the circumference of the junction between the stump and the graft, and with various orientations. Where handedness can be determined one of the supernumeraries is a left leg and the other a right leg. If the leg is rotated by only 90 degrees before being replaced on its stump, no supernumeraries are formed.

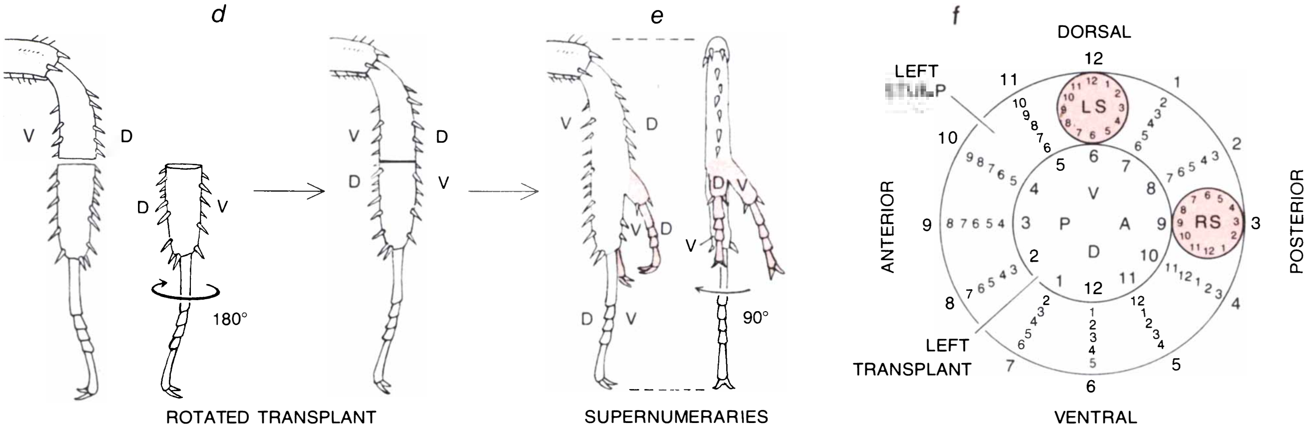
A comparable 180-degree graft of a newt-limb blastema results in the formation of one left-handed supernumerary limb and one right-handed one. Unlike the cockroach supernumeraries, however, their location is constant. These results can be explained by the same principles as intercalation and distal regeneration from complete circumferences, given the assumption that the distribution of positional values around the circumference of the cockroach leg is equal whereas the distribution around the circumference of the newt leg is slightly unequal.

The Polar-Coordinate Model

The data on limb regeneration clearly indicate that in spite of the anatomical differences between cockroach and newt legs the two epimorphic fields obey the same fundamental rules for cellular interaction. The rules can be summarized in the form of a simple model as follows. The limb can be represented as a cone (or in two dimensions as a disk). The apex of the cone (or the center of the disk) represents the distal end of the limb; the base of the cone (or the circumference of the disk) represents the proximal boundary of the epimorphic field. The two separate components of positional information that were identified experimentally thus correspond to a polar-coordinate system for specify-



NEWT SUPERNUMERARY LEGS also appear in response to grafts that lead to abnormal cell confrontations, indicating that the same organization around a meridional axis is an attribute of regenerative growth in newts as it is in cockroaches. In the experiment at left a regenerate from a proximal level in a left leg is grafted onto the stump of a right leg, reversing the anterior-posterior orientation (*a*). The result, as with the cockroach graft, is the development of two supernumerary limbs that are mirror images of the graft (*b*). The in-



tween the host and graft cells (*c, color*). The direction of shortest intercalation on each side of these positions of maximum incongruity determines the placement of positional values on these complete circumferences, and hence determines the handedness and orientation of the resulting supernumerary limbs. In the experiment at right the amputated tibia is rotated 180 degrees and regrafted onto the donor (*d*). Two supernumerary tarsus segments may now appear anywhere along the circumference of the graft; here one has appeared at outside position and one at posterior position (*e*). Intercalation (*f, color*) has produced complete circumference at each point of regeneration.

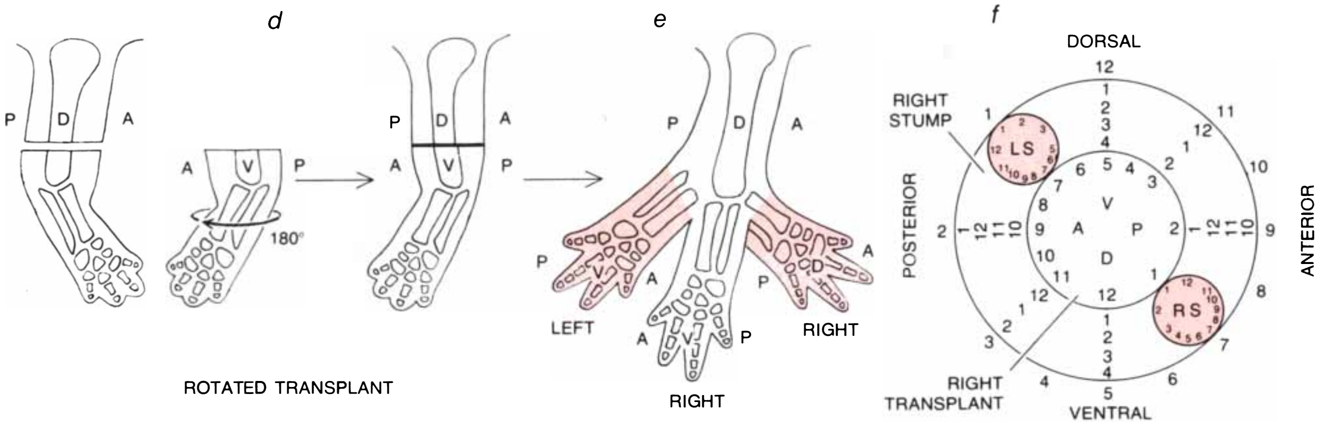
ing position in the field. Each cell in the field has a unique two-component positional value. One component defines the circumferential position of the cell in terms of 12 meridional coordinates numbered clockwise from 1 to 12. The other component defines the proximal-distal level of the cell in terms of radial coordinates lettered *A* through *E*. Two coordinates suffice to describe the position of any cell in the two-dimensional model.

Two rules govern the behavior of the cells in the field. The first we call the shortest intercalation rule. It states that if cells with normally nonadjacent potional values confront one another, the incongruity induces growth at the junction that continues until cells with the intermediate positional values have been intercalated. As a corollary to this rule, with circumferential confrontations it is always the shorter of the two possible sets of intermediate values that is intercalated. For example, when the two values No. 3 and No. 6 are grafted together, the two intermediates, No. 4 and No. 5, will be intercalated instead of the eight intermediates that separate No. 6 from No. 3 in the other direction.

The second rule we call the complete-circle rule for distal regeneration. It states that cells with positional values equivalent to any proximal-distal level can give rise to new cells with all the positional values distal to their own value. This, however, can happen only when a complete sequence of circumferential positional values has been exposed, either by amputation or by intercalation.

Our model with its two simple rules is an aid to understanding the complex regulative phenomena that can be observed in the established secondary fields of two very different organisms. To be sure, the legs of both cockroaches and newts are three-dimensional structures, whereas this model is only two-dimensional, lacking the inside-outside component. In the cockroach leg, however, the cuticular pattern is established by the secretions of a single layer of epidermal cells, and there is no evidence to suggest that the internal tissues of the leg carry positional information relevant to regeneration. It is possible that the pattern of the muscles is derived from a primary pattern in the epidermis. A two-dimensional model therefore suffices to describe the cockroach leg.

The more complex amphibian limb has a central core of bone, surrounded by muscle and covered by dermis and epidermis. Experiments suggest that here also positional information may be restricted to two dimensions. For example, experiments by Bruce M. Carlson of the University of Michigan have



tercalated circumferences (*c, color*) are at the points of maximum incongruity. In the next experiment (*right*) 180-degree rotation and regrafting of a regenerate onto the right limb (*d*) can give rise to either one or two supernumeraries at predictable locations (*e*). If a regenerate appears in the posterior-dorsal quadrant, it will be a left limb, with the ventral surface of the limb up; if it appears in the anterior-ventral quadrant, it will be a normally oriented right limb. Once again regeneration is a product of complete circumferences (*f, color*).

shown that in salamanders the manipulation of bone or epidermis after the amputation of a limb does not lead to the regeneration of multiple limbs, whereas the manipulation of muscle or dermis does. This suggests that only muscle and dermis carry positional information. It does not seem too much of an oversimplification to assume that positional values in amphibian legs are arranged in a cone with tissue to the outside and inside lacking positional information but deriving its pattern from the primary pattern in the dermis and muscle.

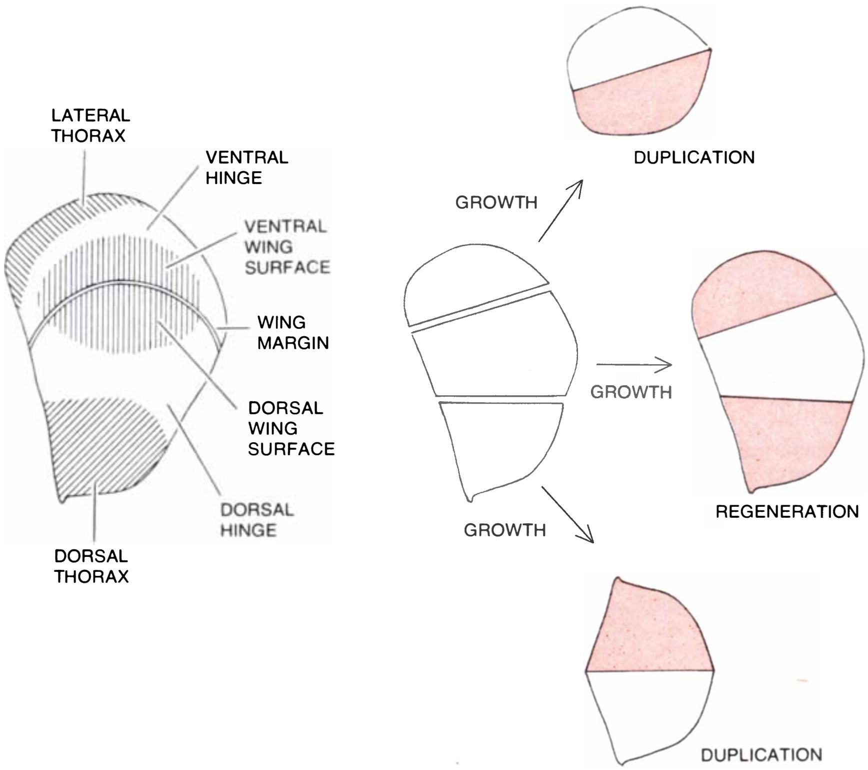
Imaginal-Disk Regulation

A suggestion made by Wolpert has become virtually an article of faith for many developmental biologists. This is that many different organisms have the same basic mechanism for specifying positional information and that the differences between organisms arise because their cells interpret positional information in different ways when they differentiate. This tenet of the universality of positional information was suggested partly by the discovery of universality in other basic biological mechanisms and partly by the results of grafting experiments between related species. The tenet has now received strong additional support from our recent finding that the polar-coordinate model can account not only for the regulative abilities of the limbs of such diverse organisms as amphibians and cockroaches but also for the regulative behavior shown by a completely different set of secondary fields: the imaginal disks of the fruit fly *Drosophila*. This favorite laboratory animal of geneticists was first chosen by Thomas Hunt Morgan because of its suitability for genetic manipulations at a time when his interests were shifting from regeneration and embryology to genetics. It is appropriate that *Drosophila* is now beginning to provide unique opportunities for developmental biologists, often for the very reason that its genetics are by this time well understood.

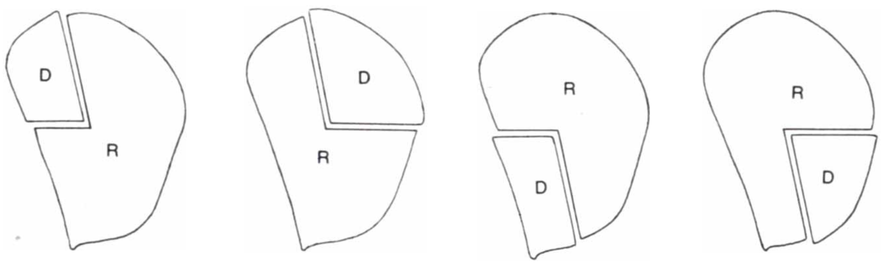
The immature larval stage of *Drosophila* has no appendages. The wings, the antennae, the legs, the balancers, the associated parts of the head and thorax, and the genitalia of the future adult, however, are all represented inside the growing larva as the small pieces of growing but undifferentiated tissue known as imaginal disks. It is not until the larva goes through metamorphosis that the imaginal disks assume their function of giving rise to the various parts of the adult. They secrete cuticle-bearing surface structures such as bristles, hairs and sense organs in intricate and reproducible patterns.

The imaginal disks of the fruit fly are an entire order of magnitude smaller than the limbs of the cockroach and the newt. They can nonetheless be subjected to various surgical manipulations that demonstrate their regulative abilities. For example, when a piece of leg imaginal disk is implanted in the abdomen of a host larva of the same age, it metamorphoses with the host and differentiates into recognizable parts of a leg. Gerold A. Schubiger of the University of Washington has shown that the specific structures formed depend on which part of the disk is transplanted; this has enabled him to construct a complete "fate map" of the leg disk. Similar fate maps have been constructed for all the other major imaginal disks, for example the wing disk.

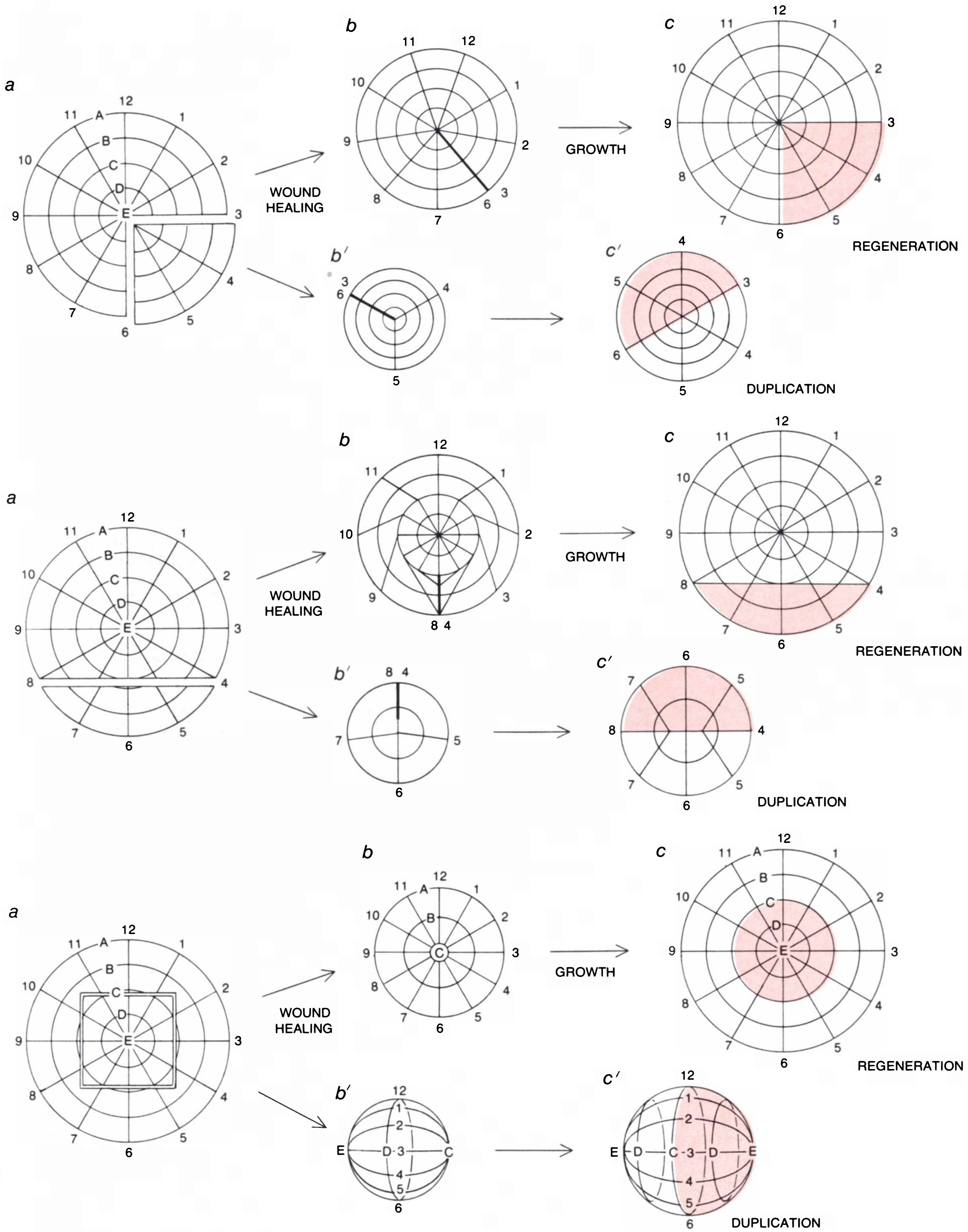
In this kind of experiment it was found that complementary parts of imaginal disks give rise to complementary sets of adult structures. In other words, there is no developmental regulation when disk fragments are subjected to immediate metamorphosis in a host larva. The reason presumably is that imaginal disks are epimorphic fields and fragments are unable to regulate unless they are given an opportunity for growth before metamorphosis. If, however, a disk fragment is implanted in an adult host rather than a larval one, the hormonal conditions allow it to grow without metamorphosing until it is retransplanted into a larval host. In short-term cultures, where disk fragments are grown for a week or less in an adult fly before they are retransplanted, they show a kind of regulative behavior that at first seemed to be rather unusual. For example, it



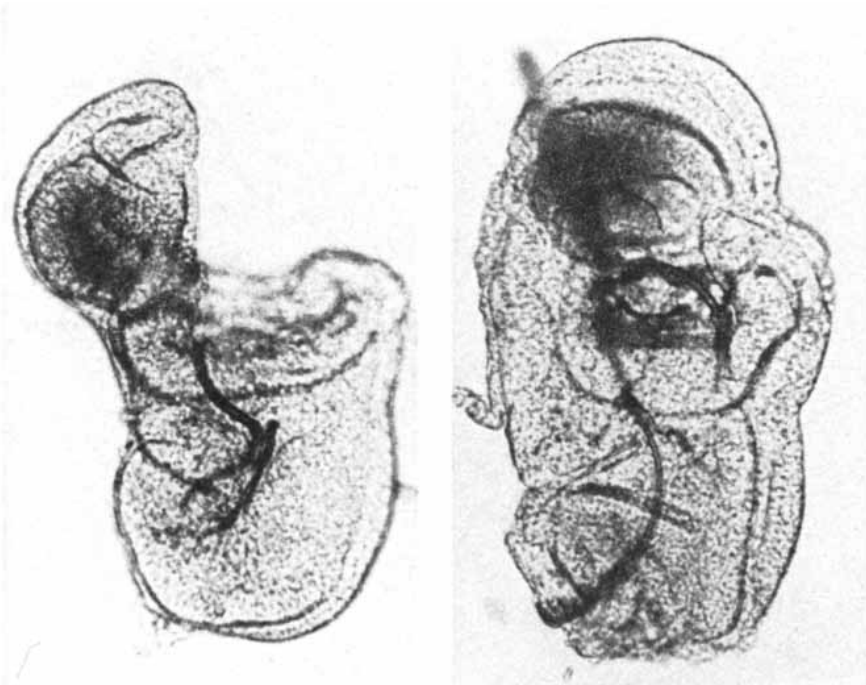
IMAGINAL DISKS in the larva of the fruit fly *Drosophila* also possess regulative abilities. Shown at the left is the imaginal disk that gives rise to a wing when the larva undergoes metamorphosis; labels indicate the normal fate of its different parts. When such a wing disk is cut into three pieces (*center*), cultured and transferred to larval hosts for metamorphosis, central piece regenerates outward in both directions (*color*). Edge pieces only duplicate themselves.



REGENERATION EXPERIMENTS with imaginal wing disks include these variations on division into a 270-degree sector and a 90-degree sector. Regardless of how disks were cut, after culture and transfer the larger of two fragments regenerated and the smaller duplicated.



INTERACTION RULES of a polar-coordinate model apply to the results obtained by culturing fragments of imaginal wing disks. At the top 270-degree and 90-degree sectors are cultured (*a*). In the larger and smaller pieces alike the cut edges heal together (*b, b′*). This brings Position 3 and Position 6 together in both instances. The rule of shortest intercalation governs, and the missing values, 4 and 5, are intercalated (*c, c′, color*). For the 90-degree sector this means duplication; for the 270-degree sector it means regeneration. In the middle the imaginal disk is bisected (*a*). Both fragments heal by contraction along the cut edges, bringing positions 8 and 4 together (*b, b′*). Intercalation again gives rise to a regenerate (*c*) and a duplicate (*c′*). At the bottom a central fragment and the periphery of a wing disk are cultured (*a*). Distal transformation results in the duplication of the central fragment (*b′, c′, side view*) and the regeneration of the peripheral fragment (*b, c*). A complete circular sequence of positional values is present at each cut edge of a contracted fragment.



HEALING OF WING-DISK FRAGMENT is seen in these micrographs. The process begins in the first day or two after the fragments are transplanted into the abdomens of adult fruit flies. Here the cut edges of a 270-degree sector (*left*) have fused (*right*) prior to intercalation.

was shown by Schubiger that the medial half of the leg disk could regenerate the missing lateral half, but that when the lateral half was cultured, it did not regenerate the medial half but instead underwent duplication to produce two copies of the lateral structures of the leg.

Many fragments of leg, wing and other imaginal disks have since been tested and have been found to either regenerate or duplicate. Furthermore, the results show an intriguing pattern: if one fragment of an imaginal disk has the ability to regenerate, the complementary fragment of that disk usually duplicates. The regenerative abilities of imaginal disks led us to regard them as typical secondary fields, but until recently the property of duplication seemed rather mysterious.

A detailed study of the regulative ability of the wing disk by one of us (Peter Bryant) showed that its behavior (and probably that of the other imaginal disks) could be easily understood in terms of the polar-coordinate model. When the disk was cut into three pieces, the central piece regenerated outward and the two edge pieces duplicated. If fragments were made by cutting along radial lines from the center of the disk to the edge, they also regenerated or duplicated. It was found that each of four 90-degree sectors of the wing disk duplicated and that the complementary 270-degree sectors regenerated.

These results can easily be understood in terms of the "shortest intercalation" rule. In order to invoke this rule there has to be a confrontation of cells with different positional values in the tissue, and it now appears that this indeed happens in imaginal-disk fragments as a result of wound healing. The 90- and 270-degree sectors heal by fusing the two orthogonal cut edges, bringing into contact cells with positional values that are 90 degrees apart around the circumference of the disk. The shortest intercalation between these confronted values then accounts for the observed regeneration of the 270-degree sectors and duplication of the 90-degree sectors. Similar reasoning accounts for regeneration and duplication in fragments bound by straight cut edges, where wound healing results in contraction of the cut edge and apposition of the positional values along it. In general this line of reasoning predicts that fragments larger than half the disk will regenerate, whereas the smaller edge pieces will duplicate. That is what is observed.

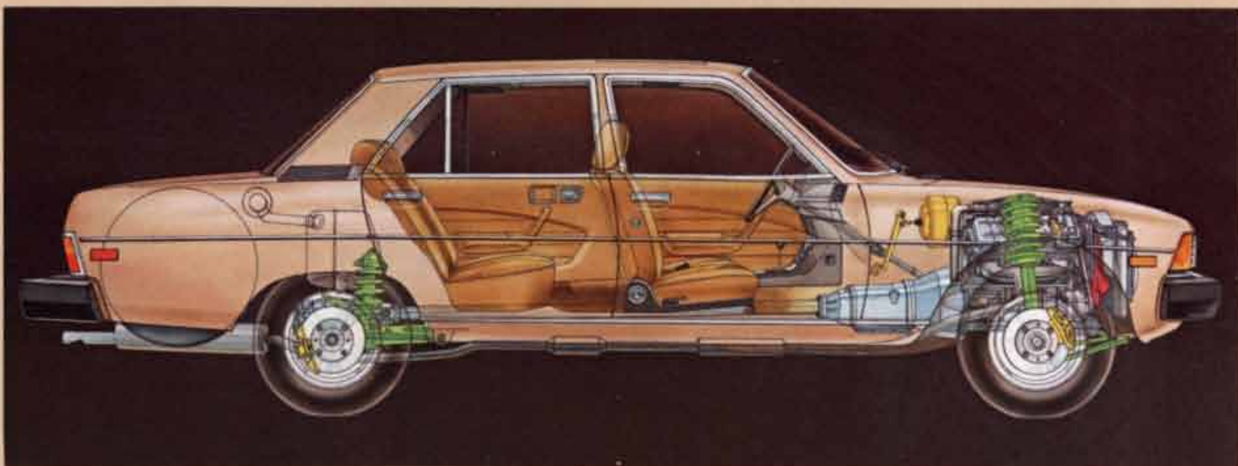
The polar-coordinate model also accounts for the behavior of central and peripheral fragments of the wing disk, which seemed to be inconsistent with the behavior of other fragments. As we have seen, fragments bounded by two cut edges usually regenerated outward in both directions, and this was true for cuts in several different directions: longitudinal, transverse and diagonal. A central square was therefore expected to regenerate outward in all four directions, but when this fragment was tested, it duplicated rather than regenerating. Furthermore, the remaining peripheral fragment was able to regenerate the center, thus maintaining the rule that complementary pieces regenerate and duplicate. This paradoxical behavior, however, is exactly what is expected from the "complete circle" rule. Both the central square and the peripheral fragment have a complete circle of positional values exposed at the cut edge, and so distal transformation is expected in each case. The addition of the structures destined to become distal corresponds to regeneration for the peripheral fragment and duplication for the central fragment.

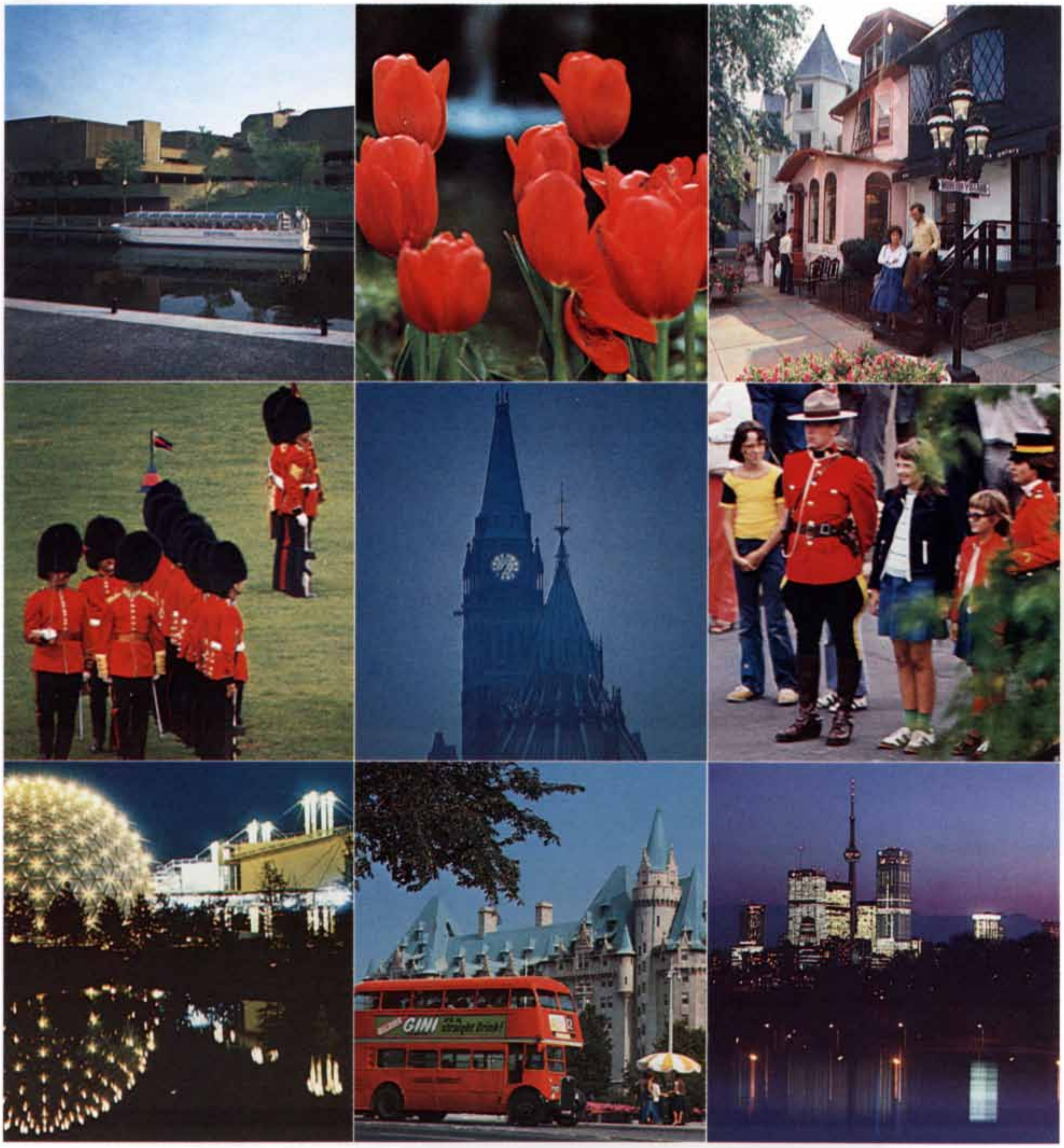
Recently one of us (Peter Bryant), in collaboration with John L. Haynie, has been able to show directly that intercalary regeneration can occur in the imaginal wing disk. We took two fragments from opposite ends of the wing disk; one end was destined to give rise to the dorsal part of the thorax and the other end to give rise to the lateral part. When either of the two fragments was cultured alone or mechanically mixed with identical fragments, it duplicated. When the two pieces from the opposite ends of the disk were mixed together, however, they were able to produce the missing intermediate structures by intercalary regeneration.

Establishment of Secondary Fields

Our model provides clues to how the pattern within epimorphic secondary fields becomes established initially in the embryo. In the 1910's and 1920's Harrison and his colleagues demonstrated that in amphibian embryos the presumptive limb region (the region that will later produce a leg) behaves in many ways like a typical epimorphic field. This is particularly clear when the presumptive limb region of one embryo is grafted in one of a variety of orientations either onto the presumptive limb site of another embryo or onto the flank of the embryo behind the presumptive forelimb site. In addition to the leg that develops from the graft itself one or two supernumerary legs may arise. At the time these grafts were first performed it appeared that almost any disturbance to a presumptive limb would cause it to duplicate. We now know, however, that the extra legs form only in situations where they would be predicted if the presumptive limb region acted as an established secondary field, according to the principles we have outlined.

At the earliest stages examined the presumptive limb region of an early embryo already has a stable anterior-posterior (front-back) organization. This organization is retained during subsequent development even if the region is removed and grafted into an abnormal orientation on a host embryo. The dor-

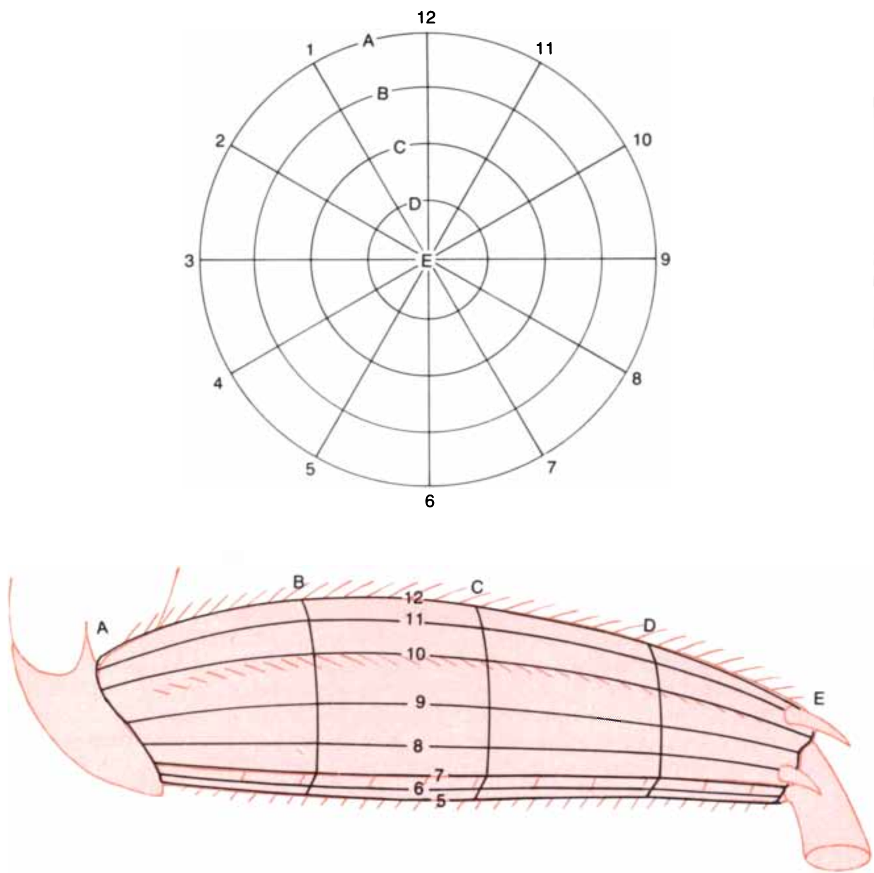



GINI

sal-ventral (top-bottom) organization of the presumptive limb region, however, is not finally established until a much later stage: immediately before the limb bud begins to grow out from the side of the body. A similar sequence of axis establishment is shown by other secondary fields such as the eye and the ear.

At first the sequence of establishing an anterior-posterior axis and then a dorsal-ventral axis suggests that the field is organized according to a system of orthogonal coordinates for positional information, that is, with the second axis perpendicular to the first. It is possible, however, to interpret the sequence so that it is consistent with our model for secondary fields as follows. Assume that the positional values corresponding to the anterior and posterior of the most proximal part of the field are the first to be established, and that the positional values corresponding to the dorsal and ventral parts of the most proximal regions are established at a later stage. Then intercalation by the shortest route between the existing positional values leads to the formation of a complete circumference at the most proximal part of the field. Thereafter distal transformation from the proximal circumference establishes all the more distal positional values of the field.

We have seen that even in diverse organisms the same basic rules of cellular behavior can govern the regeneration of larval and adult patterns, and possibly the formation and growth of those patterns in the early embryo. By means of simple surgical experiments we have identified what we believe are the kinds of cell interaction that underlie the development of form and pattern in secondary fields. So far, however, we have no idea of the mechanism of these cell interactions and know nothing of the nature of the information or signals that pass between cells in a developing organ. Discovering the nature of these interactions represents one of the major remaining challenges in molecular biology. We believe the elucidation of these developmental mechanisms will prove to be as exciting and important as the elucidation of genetic mechanisms has been.



POLAR-COORDINATE SYSTEM is represented here as a disk (*top*). Twelve meridians are identified numerically; the central, peripheral and intermediate radial levels are identified alphabetically. The system makes it possible to specify positional information in an epimorphic field; in combination with the authors' two rules it is a model of the behavior of cells in such a field. As example of cockroach femur (*bottom*) indicates, the model represents proximal-to-distal positional values in its periphery-to-center divisions and dorsal-ventral and anterior-posterior positional values in its circumferential divisions. Application to imaginal disks is new.

The Solidification of Cement

What happens after water is added to cement to transform it from a soft slurry to a rock-hard mass? Chemical reactions give rise to a system of interlocking particles that knit the material together

by D. D. Double and A. Hellawell

In its commonly accepted sense cement can be described as a finely powdered, calcareous material that, when it is mixed with water, forms a plastic paste that sets and eventually hardens to a rocklike consistency. This much is commonly taken for granted, but what is perhaps not so well appreciated is that the development of strength is the result of chemical reactions between the cement constituents and the added water. The consolidation process occurs by the formation of a rigid interlocking matrix of hydration products, which gradually replace the water between the cement grains and finally bind the composite cement mass together. The processes involved in hydration are quite complex. The purpose of this article is to describe some recently acquired knowledge in this field based on our own studies at the University of Oxford and those of workers elsewhere.

To gain perspective one might recall that the use of cement as an adhesive filler between bricks and stones in the construction of buildings and engineering works is by no means new. Its origins can be traced back to early Egyptian and Greek times. The use of cement reached a high level of expertise during the Roman period.

These early cements would now be more accurately described as lime mortars, that is, mixtures of quicklime (calcium oxide, or CaO) and sand. They depended for their development of strength and their long-term durability not so much on the initial slaking reaction—written $CaO + H_2O \rightarrow Ca(OH)_2\downarrow$ to show that on the addition of water, calcium hydroxide precipitates out and forms a colloid, a system in which fine particles of the material are dispersed in the water—as on the subsequent aging process by which the colloidal hydroxide was slowly converted to calcium carbonate by reaction with atmospheric carbon dioxide [$Ca(OH)_2 + CO_2 \rightarrow CaCO_3$]. The high quality of Roman cements, which is evident in the number and solidity of the Roman structures still standing, was due in large measure to the added discovery that lime mixed with reactive siliceous material (in the form of crushed tiles or volcanic ash) gave a cement that developed superior strength and water resistance. "Roman cement" made in this way enjoyed wide prestige and retained its popularity with little improvement or development until the end of the 18th century.

Although there is some dispute about it, the invention of portland cement as it is now known is usually attributed to Joseph Aspdin, a builder from Leeds who took out a patent in 1824 for "a cement of superior quality resembling Portland stone." His cement was prepared by sintering fixed proportions of limestone or chalk (calcium carbonate) with clay (aluminosilicates) in a kiln at very high temperatures. This is the basis of the manufacturing method that is still in service today. In the course of firing, the raw materials combine to produce clinkers (small rounded lumps that are subsequently ground to yield the cement powder), which contain reactive calcium silicates. The silicates give the cement its hydraulic character, that is, the property of hardening by reaction with water. Unlike the original lime-based cements, Aspdin's type of cement would harden under water and also be resistant to water for a long time. One of its first large-scale uses was in the construction of the Thames Tunnel from 1825 to 1843 by Marc Isambard Brunel and his son Isambard Kingdom Brunel.

Ordinary portland cement is still by far the most important cement in terms of the quantity used for construction purposes. In its most widespread form it is mixed with inert filler material (sand and graded aggregates) to produce concrete. A number of variations in type have been developed to meet specific requirements (such as rapid hardening, low evolution of heat and resistance to sulfates), but they involve relatively minor modifications in composition or method of preparation.

A notable exception in this context is high-alumina cement, which also has hydraulic properties although it is quite different in its chemical composition and hydration characteristics. For this reason it provides an interesting comparison with portland cement. High-alumina cement, which was developed in France and came on the commercial market after World War I, is manufactured by firing limestone with bauxite (hydrated aluminum oxide).

Unlike portland cement, which was arrived at more or less by trial and error, high-alumina cement was specifically designed to resist sulfates. Indeed, it proved successful in overcoming a number of construction problems in environments where portland cement was unsuitable because of its vulnerability to corrosive attack by sulfates. Another advantage that quickly came to be appreciated was that high-alumina cement has a very rapid rate of hardening. This latter quality has made it an attractive alternative to the slower-hardening portland cement in such applications as rapid repair work and the fabrication of precast units, where a higher production rate provides an economic advantage that outweighs a relatively higher cost.

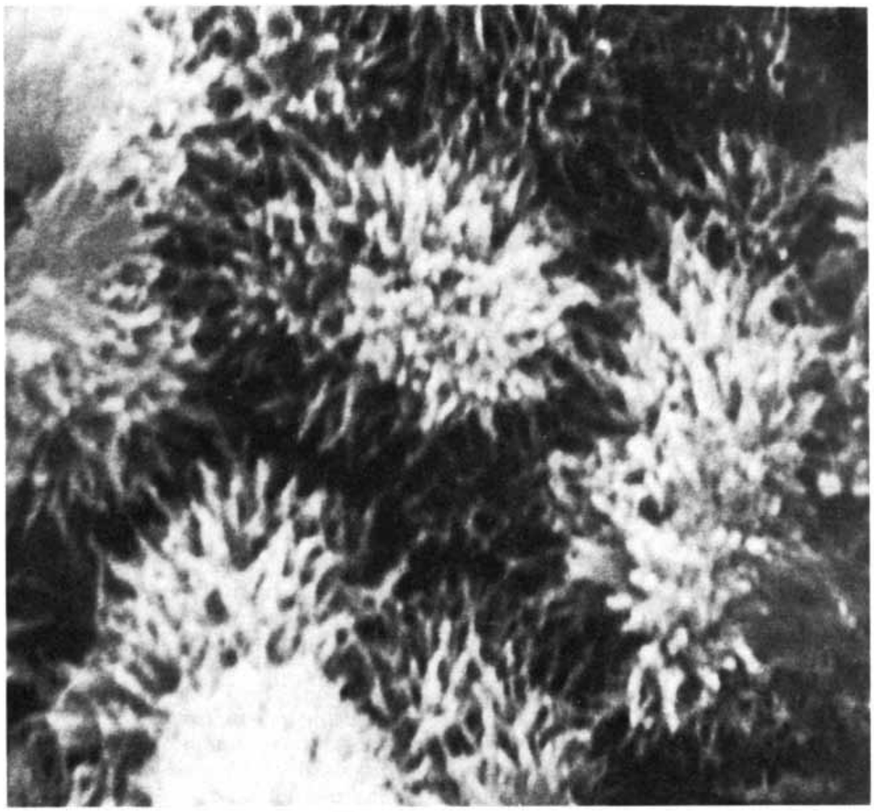
Nevertheless, the accumulation of experience with concretes made from high-alumina cement has shown that it can be defective in its long-term strength. As a result the use of this cement in any load-bearing capacity has come to be regarded as a highly questionable practice. Most countries now avoid using high-alumina cement for structural purposes, and some of them have imposed prohibitions or restrictions to this effect. Frequently these actions have followed incidents in which structural members and even entire buildings containing high-alumina cement have collapsed, primarily because of the deterioration of the concrete.

In this article the main features of the hydration and development of strength of portland cement and high-alumina cement are described in terms of a comparison between the two types. Both cements are based on the ternary system of oxides of calcium, silicon and aluminum ($CaO–SiO_2–Al_2O_3$). Commercial cements tend to vary in composition over a small range and also contain a number of impurities, notably iron oxides (less than 10 percent in portland cement, up to 20 percent in high-alumina cement), and to a much lesser extent free lime and associated oxides. The impurities do not, however, substantially affect the basic properties of the cements.

Portland cement is basically a calcium silicate mixture predominantly containing tricalcium and dicalcium silicates, which can be written in abbreviated form as C_3S and C_2S. (We employ here a simplified notation wherein oxides of calcium, silicon, aluminum and iron are represented respectively by C, S, A and F and water is represented by H. Whereas dicalcium silicate would normally be written $2CaO \cdot SiO_2$, for example, in the simplified system it is C_2S.) In addition there are also present smaller proportions of tricalcium aluminate (C_3A) and a calcium aluminoferrite approximating the formula C_4AF. Typical weight proportions in an ordinary portland cement are 50 percent C_3S, 25 percent C_2S, 10 percent C_3A, 10 percent C_4AF and 5 percent other oxides. High-alumina cement, as its name suggests, contains mainly monocalcium aluminate ($CaO \cdot Al_2O_3$, written CA) and relatively little silicate material. A variety of less well specified compounds are also present, but CA is the main reactive ingredient.

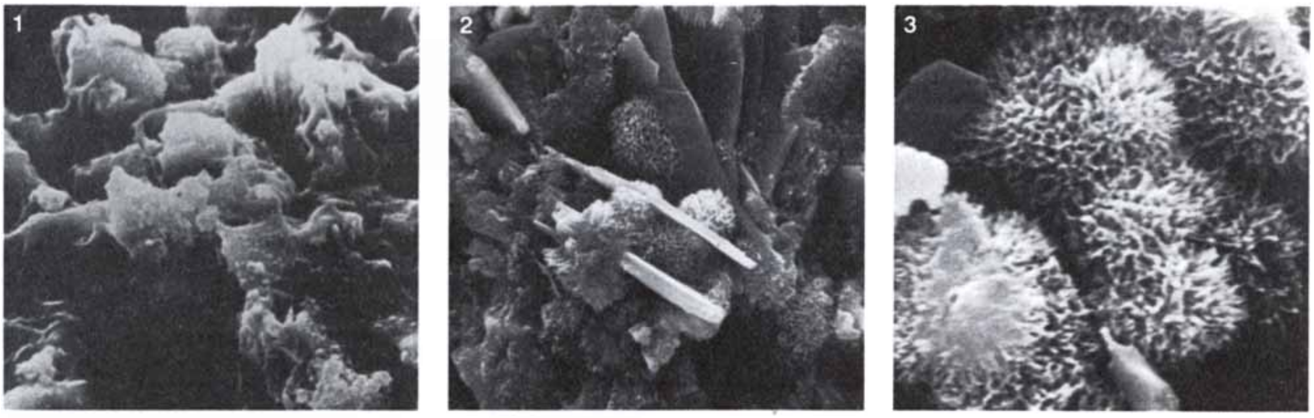
In both cements the principal anhydrous constituents show hydraulic properties: in a finely divided form they react with water to yield a paste that sets and eventually hardens. Moreover, because the cements are chemically dissimilar, the way and the rate at which their strengths develop are different. Both cements can achieve about the same ultimate strength, but high-alumina cement initially hardens much faster than portland cement. In practice high-alumina cement can develop in a day about the same strength that a comparable portland cement would achieve in a month. The advantages of high-alumina cement in this context are obvious. It is the long-term strength of high-alumina cement that is the problem. Depending on the ambient temperature and the environment, the cement may deteriorate by a process known as conversion in such a way that the strength shows a sudden dramatic decrease. The conditions that accentuate this behavior are a high-temperature environment and too high a content of water in the cement. Although these factors would also tend to reduce the strength of portland cement at all ages, the effects are much less marked, and there is no sudden drop in strength comparable to that displayed by high-alumina cement.

Taking the composition of portland



MICROSTRUCTURE OF PORTLAND CEMENT in the early stages of the reaction with water is revealed in this scanning electron micrograph at an enlargement of 6,000 diameters. Individual grains of cement have begun to develop fibers of the silicate-gel reaction product.

HIGH-ALUMINA CEMENT, which was designed to resist attack by sulfates and which hardens faster than portland cement, is portrayed at the same magnification as the micrograph at the top of the page. The reaction products are crystalline aluminates with an interlocking platelike and needlelike form. The chemical instability of these reaction products is the cause of a problem with high-alumina cement, namely the unpredictability of its long-term strength.

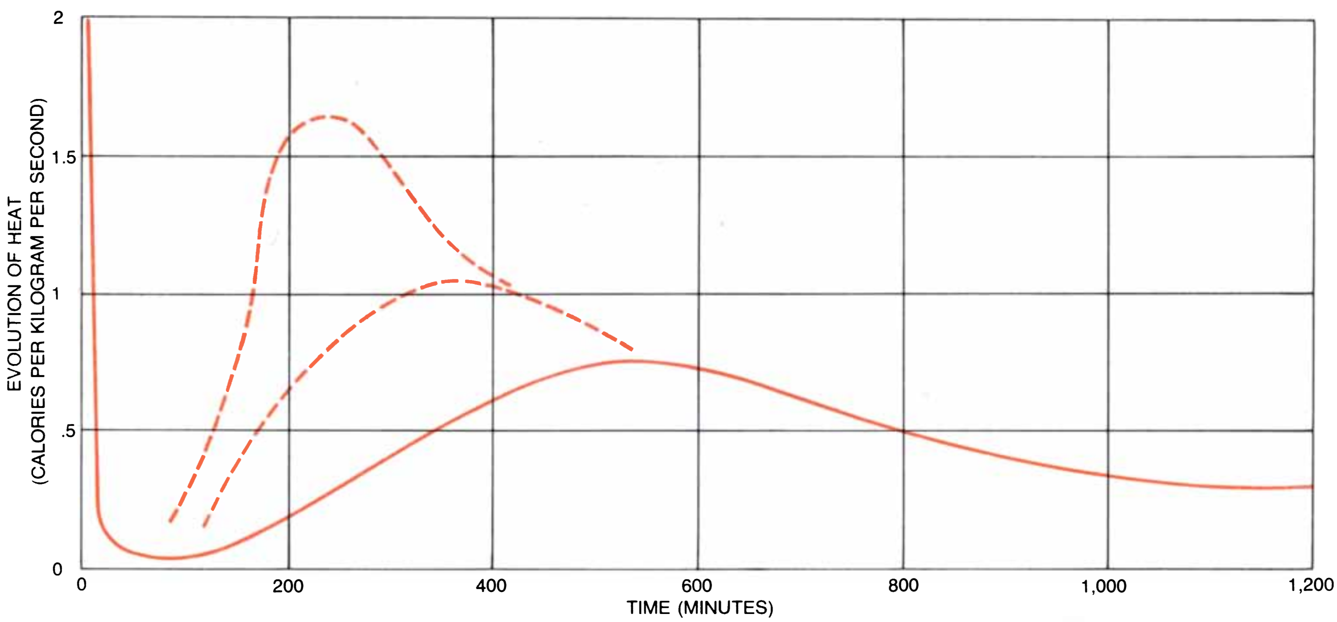


PASTES OF PORTLAND CEMENT appear in scanning electron micrographs made at various stages of hydration, that is, at various times after water was mixed with the cement. After two hours (*1*) the initial gel coatings are visible around the cement grains. After a month (*2*) the fibrils are evident, as are large platelike crystals of calcium hydroxide. Detail of same sample (*3*) shows the interlocking fibers.

cement as that given above, it is appropriate to consider what is known about the products of hydration and their proportions in the hardened paste. As might be expected, the hardening reaction in a cement-water paste is principally associated with the hydration of the two silicate compounds, C_3S and C_2S. In the anhydrous clinker they occupy about 75 percent by weight of the total; they finally provide the bulk of the hydrated material. The reaction rates of C_3S and C_2S differ appreciably, but the final product in both cases is thought to be the same: a colloidal calcium-silicate-hydrate gel of rather indefinite composition and structure. In a hardened cement this gel, in a finely divided and largely amorphous form, occupies about 70 percent by volume of the hydrated material and thus forms the main bonding agent between the residual grains of unreacted cement and the other crystalline products of hydration. The major part of the other crystalline products consists of calcium hydroxide (portlandite), the rest being composed of various complex aluminum hydrates and other minor substances.

Of the two silicates, C_3S is primarily associated with the hardening reaction that takes place during the early stages of hydration, within the first few days and weeks. The slower-reacting C_2S is probably responsible for the longer-term and continuous development of strength, which can extend for periods of months or even years. The contribution to the strength from the aluminate compounds is uncertain, but it is probably relatively small.

In a cement paste under normal conditions the hydration reaction does not usually go to completion. Even after several years a typical microstructure will show residual grains of unreacted cement embedded in a matrix composed mainly of calcium-silicate-hydrate gel. The reason for this state of affairs is that during hydration the reaction products form coatings around the cement grains, and the coatings progressively inhibit the access of water to the anhydrous material. Only if these coatings are continu-



EVOLUTION OF HEAT as portland cement sets and begins to harden is charted. The sharp first peak represents rapid heat-generating processes, including the hydration of free lime and the dissolution of impurities in the cement. The trough thereafter represents the "dormant" or "induction" period, which probably results from the inhibition of hydration by the initial gel coatings formed around the cement grains. The coatings eventually become ineffective, because the reaction rate soon rises to a second peak. At a hydration temperature of 20 degrees Celsius (68 degrees Fahrenheit) the rate of reaction is represented by the lowest curve. The broken-line curves above it show the effect on the form of the second peak of increasing the temperature of hydration to 30 and 40 degrees C. respectively.

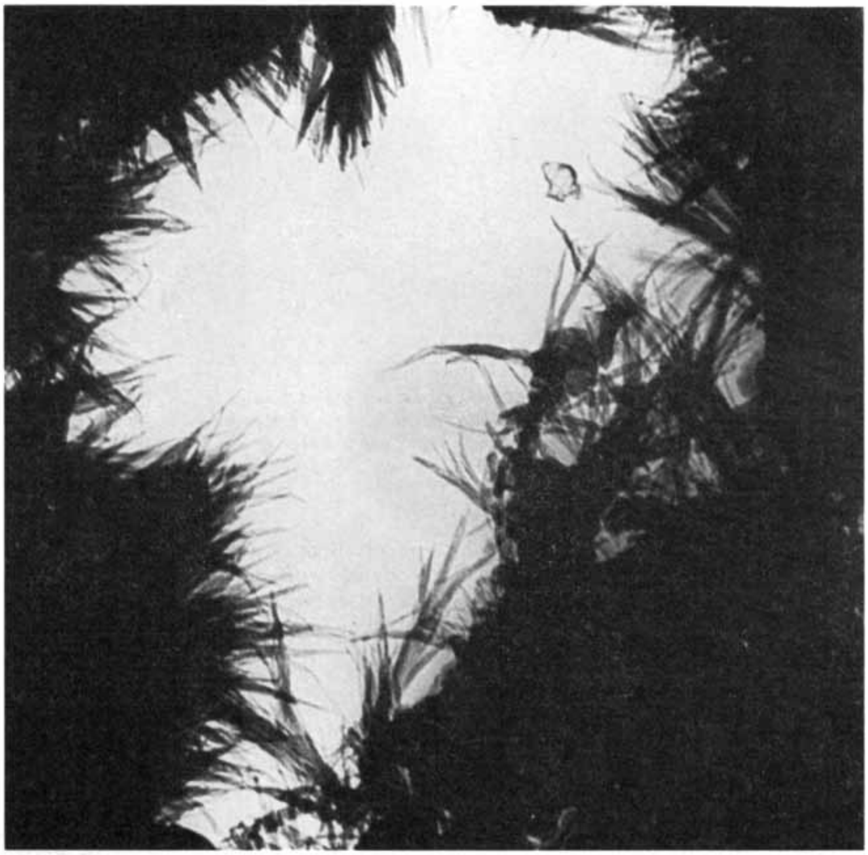
ously broken up to expose fresh reactive surfaces, as, for example, by continuous agitation of the paste in a ball mill, can complete hydration be achieved in a reasonable length of time.

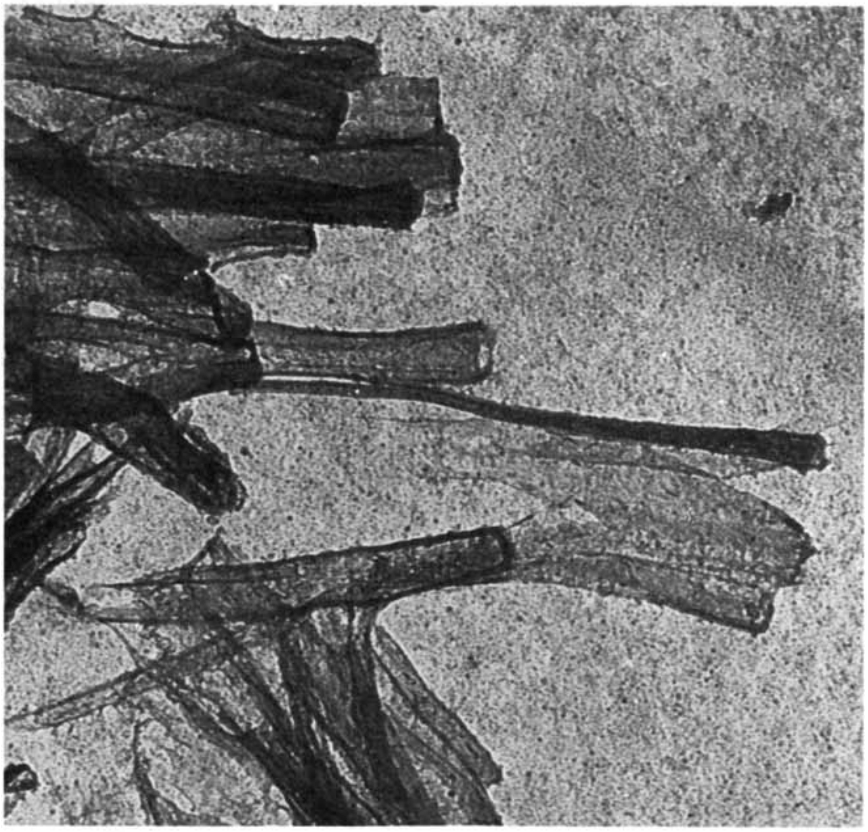
The presence of unreacted clinker material can have some interesting effects. For example, if the cement is ground up and mixed again with water, it will set a second time and develop a degree of strength that, although it is inferior to the strength of the first setting, can nonetheless be substantial. It is even possible that if fine microcracks develop in a cement structure, they can heal themselves by renewed hydration.

The reaction between cement and water is exothermic; under typical conditions the temperature of a paste may rise by tens of degrees. This phenomenon can present quite a problem in laying concrete structures of large section. It is sometimes necessary to cool the material during the setting and hardening period to prevent cracking through thermal expansion and contraction. The heat evolved is directly related to the degree of hydration of the various constituents of the cement. Because of the physical nature of the reaction, the heat-evolution rate tails off exponentially at long time intervals. Under normal conditions, however, some 50 percent of the total heat available (100 to 150 calories per gram) is liberated within the first three days of mixing and some 30 percent is liberated within the first day.

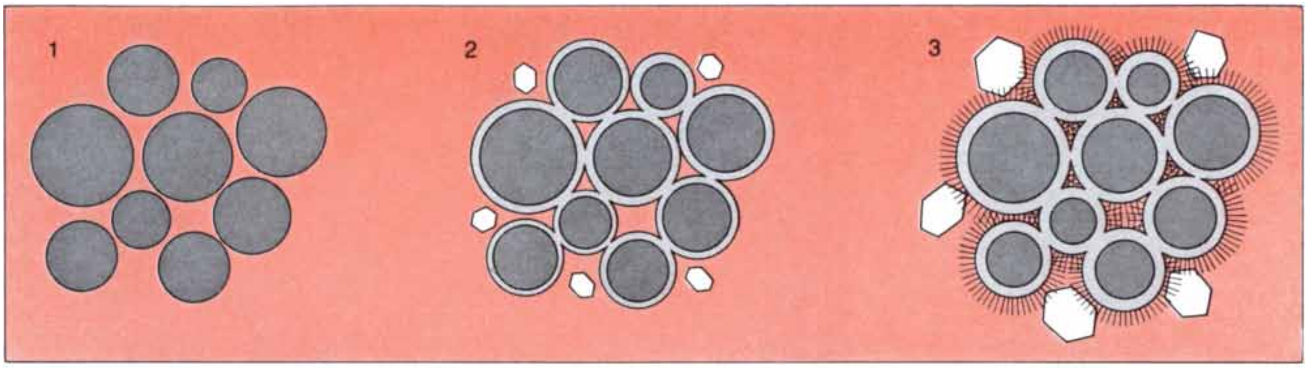
If the evolution of heat is plotted as a curve, it shows two peaks [*see bottom illustration on opposite page*]. One peak is sharp, reflecting highly exothermic processes. It reaches a maximum within 10 minutes of mixing and is mainly due to the hydration of free lime and the dissolution of other impurities that are always present in small quantities in commercial portland cement. After the first peak the rate of reaction drops rapidly to a low value, where it remains for an hour or two. This is the "dormant" or "induction" period. It is thought to be brought about by the inhibition of hydration as the cement grains become coated with the initial gelatinous products. This picture is confirmed by observations with the electron microscope.

Later the coatings apparently become ineffective, because the reaction rate begins to increase again and reaches a second (but much lower) peak in from eight to 10 hours. Thereafter it decreases gradually over a much longer period. The form of the peak is largely determined by the hydration of the major constituent of the cement, tricalcium silicate. The curve has a form that is characteristic of many phase transformations occurring by a process of nucleation and growth. As one might expect, the extent of reaction is increased by the temperature and is directly related to





HYDRATED PORTLAND CEMENT is seen in microstructure in the transmission electron microscope. The top micrograph shows a wet sample of cement after some two days of hydration; one can see a well-developed fibrillar growth of silicate gel around the cement grains. The enlargement is 20,000 diameters. The bottom micrograph (enlarged 70,000 diameters) shows a dried sample of hydrated cement that clearly exhibits the tubular nature of the fibers.



STAGES IN HYDRATION of portland cement are depicted schematically. The process begins (*1*) with the addition of water, which surrounds the individual grains of cement. Soon the coating of gel appears around the grains (*2*), and angular crystals of calcium hydroxide that develop as a by-product of the hydration of the silicates in the mix become evident. Finally the tubular fibrils develop (*3*). The interlocking of these fibrils as hydration proceeds eventually binds the cement and other components of the mix into a hard mass.

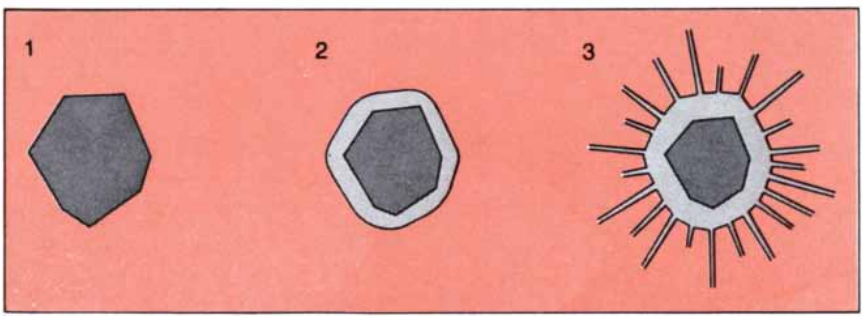
the surface area of the cement powder, that is, to how finely the cement powder has been ground. Variations in the ratio of water to cement in the range from 30 to 50 percent (which is typical of ordinary practice) do not significantly alter the pattern of hydration.

Although the heat-evolution curves give information about the extent and rate of hydration, they tell little about the structure of the material and are not necessarily related in any simple way to the development of strength. It is true that under normal conditions the setting period occupies approximately the interval of time between the dormant stage and the second peak and that the hardening reaction follows thereafter, but the setting and hardening processes are clearly dependent on the detailed development of the microstructure of the paste. Because of the fine scale of the hydrates involved the microstructure has been most usefully studied by electron microscopy.

The micrographs show that within two hours the originally angular and crystalline grains of cement become coated with a gelatinous envelope of hydrate material. The grains are bridged at their points of contact by these coatings. The network of relatively weak bonds thus formed accounts for the loss of plasticity of the paste at this time, that is, for the setting effect. The structure can be broken up easily but will quite quickly set again as the cement grains recoagulate.

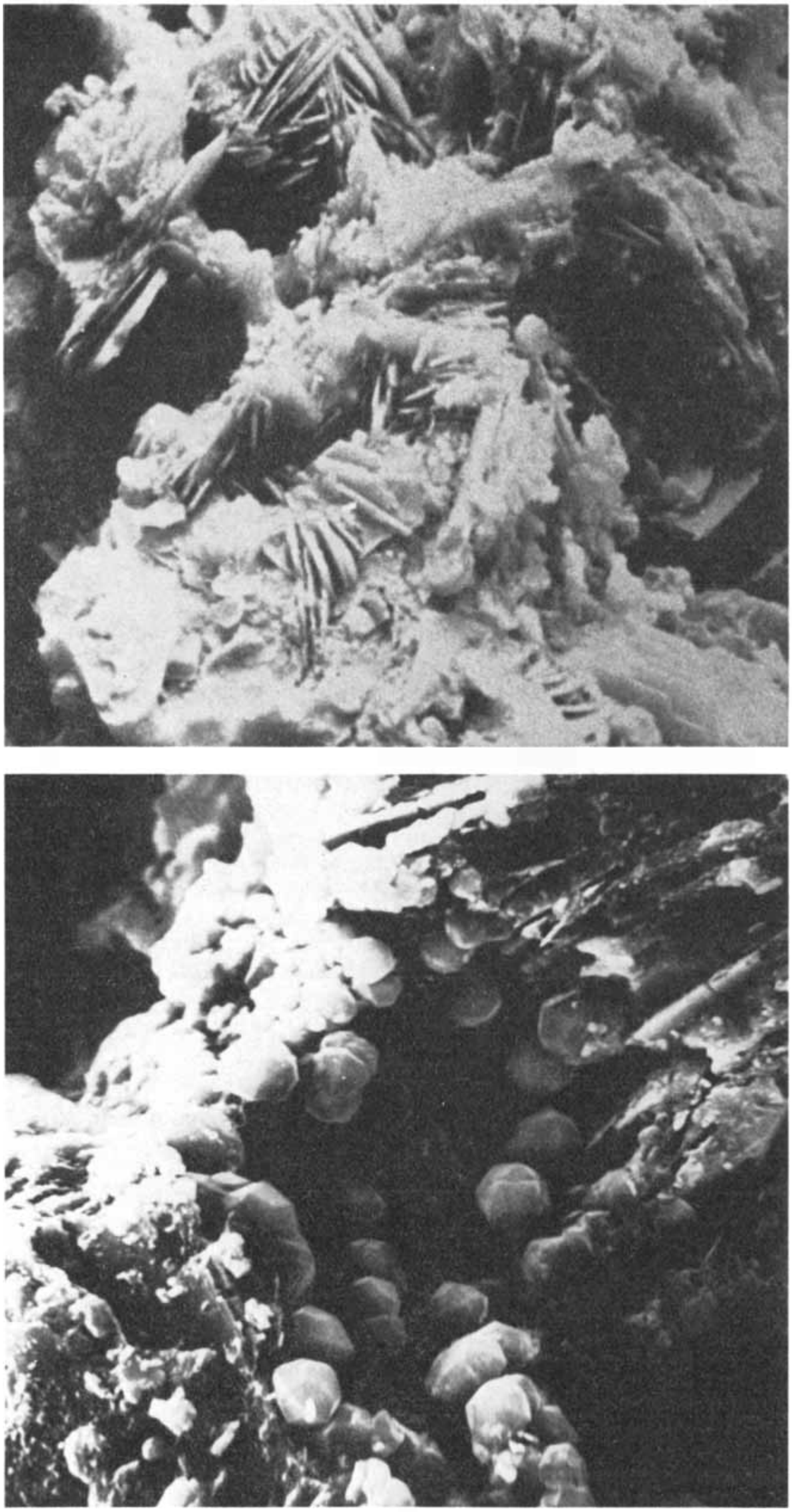
The irreversible development of strength during the hardening period, which begins after some three to five hours, is associated with a distinct change in the microstructure of the cement. The initial gel coatings develop fine surface protuberances. Over a period of time they grow into thin, densely packed fibrils radiating like porcupine quills from the individual grains of cement. This fibrillar development is the calcium-silicate-hydrate gel product of the reaction between the cement silicates and water. At the same time calcium hydroxide (portlandite) precipitates as a by-product and forms the large angular crystals that are so prominent in the microstructure. As hydration proceeds the gel fibrils gradually interpenetrate in the region between the adjacent cement grains. The interwoven meshwork thereby created eventually consolidates to form a rigid matrix that effectively binds the composite mass: the residual grains of unreacted cement, the crystalline products of hydration and, in a concrete, the sand and the aggregate material.

Although the secondary growth of the fibrils of calcium-silicate-hydrate gel is clearly of importance in the development of strength of the cement-water paste, its detailed morphology and mechanism of growth have hitherto remained obscure. Recent studies by our research group at Oxford making use of transmission electron microscopy have provided some new insight into this problem. What these pictures show is that the fibrils do not resemble the faceted regular shapes one might expect from the growth of solid crystals from aqueous solution and that the individual fibrils are not solid but appear to consist of fine hollow tubes. This information presents an entirely new aspect of the problem.



SILICA "GARDENS" provide an analogy to the growth of tubular fibrils in portland cement. A silica garden forms when water-soluble salts of certain metals are put in a solution of sodium silicate. In this example, from a crystal of cobalt nitrate, the salt crystal begins to dissolve (*1*) and a shell of insoluble silicate forms around it (*2*). Since the membrane is permeable to water, the crystal continues to dissolve. Finally osmotic pressure ruptures the membrane (*3*), causing the growth of precipitated silicate in the form of fine hollow tubes similar to those found in hydrated portland cement, although in a silica garden the fibrils are much larger.

In the context of the precipitation of solids from aqueous solution the concept of a tubular morphology is perhaps rather unusual. It turns out, however, that such growth forms are commoner than might have been expected. A familiar example of tubular growth is to be found in the silica "gardens" that develop when water-soluble salts of metals, which themselves form insoluble silicates, are placed in a dilute aqueous solution of sodium silicate. The growth pattern of a silica garden provides a striking visual analogy with what happens during the hydration of portland cement. A gelatinous silicate-hydrate envelope is first precipitated around the salt crystal. After a dormant period the gel bursts open intermittently to pro-



DEVELOPING MICROSTRUCTURE of hydrated high-alumina cement paste is seen in the scanning electron microscope. The paste was made with a high ratio (60 percent) of water to cement and stored at a high temperature (40 degrees C., or 104 degrees F.) in order to study the conversion process that often gives rise to the failure of load-bearing members made of high-alumina cement. The top micrograph shows the paste after one day; the needlelike crystals of the initial metastable hydration products are evident. The bottom micrograph, made after 14 days of hydration, shows faceted polyhedrons of the stable but weak product of conversion.

duce fine tubular growth forms with a radiating fibrillar appearance. In spite of certain obvious differences between the hydration of portland cement and the growth of a silica garden (in the scale of growth and the nature of the chemical constituents) the parallel between the two processes is so close that it suggests a similar growth mechanism operates in both cases.

The growth mechanism for the silica garden has been shown to depend on osmotic pressure. As the salt crystals come in contact with the aqueous silicate solution, they begin to dissolve and simultaneously precipitate a gelatinous silicate-hydrate envelope. This envelope is colloidal and acts as a semipermeable membrane. Because of the concentration difference in the solutions across the membrane (a concentrated solution of the metal salt inside and a dilute solution of the silicate outside) water diffuses preferentially into the envelope, causing continued dissolution of the salt crystal. The internal pressure slowly builds up, and eventually the envelope bursts, ejecting jets of salt solution into the silicate solution. There is a spontaneous reaction between the two with the precipitation of the insoluble silicate material as a continuously growing tubular wall. The initial membrane acts as an osmotic pump, the salt solution flowing up the tube and precipitating at the tip by an almost steady-state process.

So far there is little quantitative data bearing on this mode of growth, but it is likely that it is relatively common whenever the initial reaction product between two solutions is a continuous semipermeable membrane. Both the initial envelope and the tubes that grow from it are amorphous and gelatinous, although when they are extracted from solution, they prove quite hard and strong. In any given system the rate of growth is proportional to the measured diameter of the tube; smaller tubes are the result of slower growth. This finding is in the right direction qualitatively to support the analogy between silica gardens and the hydration products of portland cement. A wide variety of metal-ion salts show this type of growth. It is reassuring that calcium salts are among them, since calcium-silicate-hydrate gel is the principal hydration product of portland cement.

Using the silica garden as a model it is possible to explain the essential features of the hydration of portland cement and in particular the growth of the fibrillar calcium-silicate-hydrate gel material that is such an important factor in the development of strength. The implication of the analogy is that osmosis provides the driving force for the hydration of portland cement. Further detailed studies are being carried out in order that a better understanding of the fundamental nature of this material can be obtained.

The hydration products of high-alumina cement are more conventional crystalline forms. As we have indicated, the principal anhydrous constituent of high-alumina cement is monocalcium aluminate ($CaO \cdot Al_2O_3$, written CA), and the development of strength of the cement is derived primarily from the hydration products of this compound. Under normal conditions at room temperature the main product is the decahydrate phase CAH_{10}, which is usually associated with smaller quantities of octahydrate C_2AH_8 with alumina gel, which eventually ages to crystalline gibbsite, AH_3. The relative amounts of CAH_{10} and C_2AH_8 present in the cement paste are dependent on the temperature of hydration and the composition of the cement itself. These crystalline hydrates provide the matrix that binds the paste together. Their early production gives rise to the rapid increase in strength that is characteristic of the material.

Conversion occurs because the initial hydrates CAH_{10} and C_2AH_8 are metastable and tend to change spontaneously to the stabler hydrate C_3AH_6, with water and further gibbsite as by-products. The process is accompanied by a change in the crystal structure of the hydrates (CAH_{10} and C_2AH_8 have well-defined hexagonal crystal structures and C_3AH_6 is cubic) and, because of differences in their densities, by a substantial decrease in the volume of solids. The change amounts to about 50 percent for CAH_{10} and some 30 percent for C_2AH_8.

These changes are the basis of the problem with the long-term strength of high-alumina cement. They lead to a loss in cohesion of the microstructure and an increase in the porosity of the hardened cement paste. At room temperature and lower temperatures the rate of change is slow, particularly in dry conditions, and the effects may not be evident for years. At higher temperatures and in a moist environment the changes can occur within a few weeks or even days. In the end the result is a marked reduction in strength and an increased susceptibility to corrosive attack. High-alumina cement lasts longest when it is mixed fairly dry and is in a cold environment. The reverse combinations are disastrous for the long-term stability of the material.

The reactions can be followed by examining with the scanning electron microscope samples of cement paste at various stages of hydration. Our samples, which were made with a high ratio of water to cement, were artificially aged by being stored in a humid environment at a fairly high temperature (40 degrees Celsius, or 104 degrees Fahrenheit). Initially the cement becomes coated with a smooth gelatinous layer, not much different in appearance from the one that forms on portland cement. After a short incubation period the coatings are broken by ridges that grow rapidly into faceted crystals with hexagonal or acicular (needlelike) outlines. These crystals can be identified with the first products of hydration; it is by their complex interlocking that the cement achieves its rapid rise in strength.

Under the contrived aging conditions the sample undergoes a rapid transformation, and within 14 days it exhibits the stabler form of a hydration product: the hexahydrate in a nearly spherical form with faceted polyhedrons. Since faceted crystals are characteristic of growth from solution, the conclusion is that the change to the stable hydrate must take place in this way. In other words, the change involves a recrystallization from an aqueous solution. Evidently, then, free water is present in the microstructure of the cement, since it would be needed to provide the medium in which the stable phase could nucleate and grow. This conclusion suggests why a high ratio of water to cement in the original mix can facilitate the loss of strength in high-alumina cement.

One is tempted to ask what can be done about this problem. Can the transformation be inhibited, or could the stable product be made to form preferentially in the first place? It is difficult to see how the change could be prevented except by the accepted practice of making the mix relatively dry, so that the amount of water available to promote the undesirable transformation would be at a minimum. Perhaps if the initial hydration were carried out at a sufficiently high temperature, the stable hydrate might develop more rapidly. The necessary temperatures are probably impractically high, however, and in any case it is unlikely that a cement matrix composed of faceted polyhedrons would develop the interpenetrating network that seems to be a requirement for high strength.

In structural member, such as a beam of prestressed concrete, there are further practical problems. Although it is accepted that the best approach is to use a minimum of water during mixing of the cement, how uniform can such a mixture be in a section from 25 to 30 centimeters across and from five to 10 meters in length? It is instructive to consider an example.

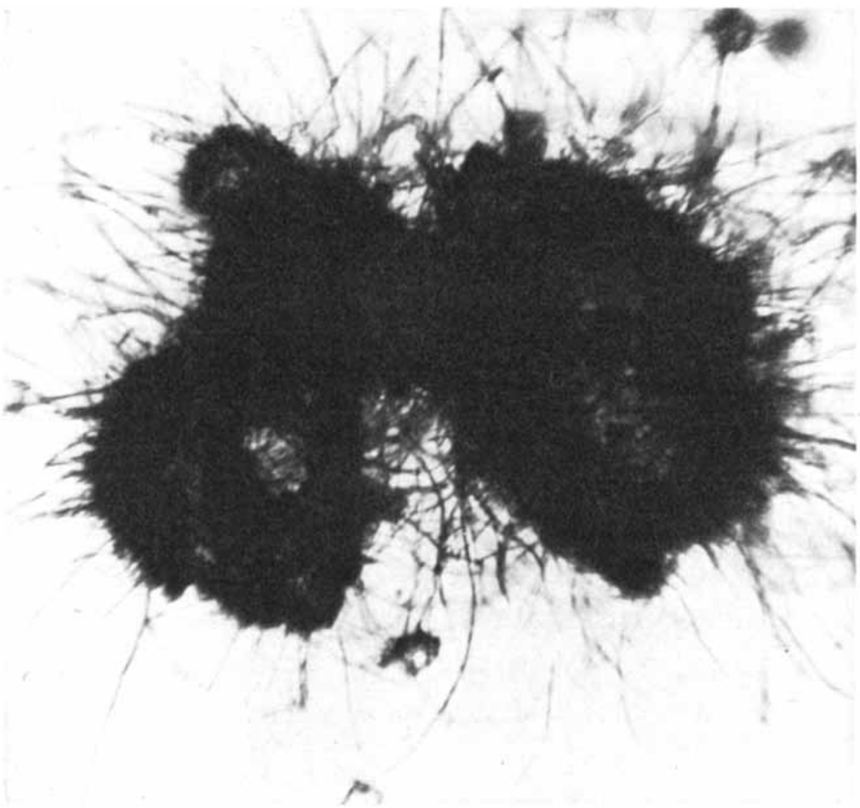
A cross section two centimeters thick was cut from a six-meter beam of high-alumina cement. The beam had a wood insert at the bottom, presumably so that objects could be nailed to the beam. The cross section was then cut into smaller pieces and the relative conversion (from metastable to stable compound) was determined for each piece by a thermochemical analysis.

As one might expect from such an inhomogeneous structure, the distribution was variable, with a few regions of low conversion near the surfaces and around the wood insert, where the concrete had been locally dried by absorption of the water. Inevitably one must assume that variations in the other dimension, along the length of the beam, would show a larger amplitude. The practical problems of ensuring homogeneity of such a material, within precise limits and over large dimensions, can be appreciated. These variations in the conversion values must also be reflected in variations of the mechanical strength of the concrete from point to point. Ultimately the strength of a given beam will depend on its weakest part.

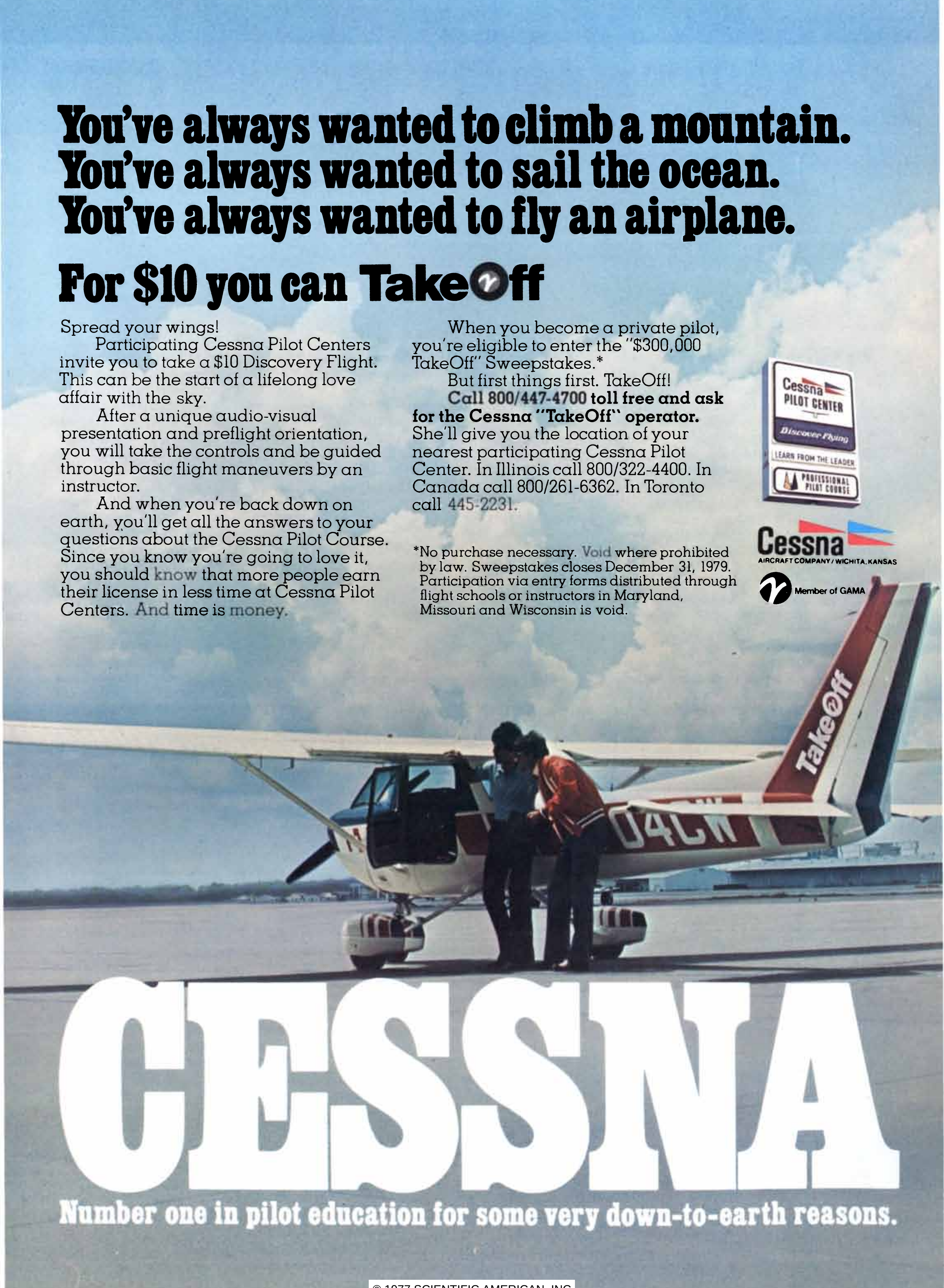
One of the characteristic features of cements (and concretes in general) is that they can develop a large compressive strength during hydration, whereas their tensile strength remains relatively very low. For this reason it is usually necessary to reinforce concrete beams or to subject them to prestressing or poststressing in compression by the inclusion of stretched steel wires or rods. In general the engineering design of a building has to be such that the concrete is not subject to an undue tensile stress. Clearly any improvement that could be achieved in the tensile properties of a cement or a concrete would have far-reaching and universal importance in building practice.

If the setting and hardening processes depend on the interpenetration and interlocking of gelatinous fibrils or faceted crystals, as the recent findings we have described suggest, it is reasonable to ask whether the tensile strength could be improved by somehow emphasizing the development of the interlocking features, perhaps by causing the growth products to interweave or interlock more closely. Modification of the microstructure of metals has been fundamentally important in metallurgical technology and is also well known in the solidification of many alloys, but it has not been tried with cement.

With the well-developed crystalline hydrates of high-alumina cement the prospects for useful modification are uncertain. With the irregular fibers of portland cement, however, some improvement might be expected if, for example, the fibers could be made less regular and more intricately woven. It is instructive to consider the long, thin and densely interwoven fibrils that grow from crystals of ferrous sulfate. Clearly it would be valuable to find out if something similar can be achieved with portland cement and, if so, how this curious form of growth can be controlled.



IMPROVED STRUCTURE that might be sought for portland cement is suggested by the fine and densely interwoven network of tubular fibrils visible in this micrograph. Crystal seen here at an enlargement of 32 diameters is ferrous sulfate that was immersed in a solution of sodium silicate. A denser interlocking of fibers would improve tensile strength of portland cement.


You've always wanted to climb a mountain.
You've always wanted to sail the ocean.
You've always wanted to fly an airplane.
For $10 you can TakeOff
Spread your wings!
Participating Cessna Pilot Centers invite you to take a $10 Discovery Flight. This can be the start of a lifelong love affair with the sky.
After a unique audio-visual presentation and preflight orientation, you will take the controls and be guided through basic flight maneuvers by an instructor.
And when you're back down on earth, you'll get all the answers to your questions about the Cessna Pilot Course. Since you know you're going to love it, you should know that more people earn their license in less time at Cessna Pilot Centers. And time is money.
When you become a private pilot, you're eligible to enter the "$300,000 TakeOff" Sweepstakes.*
But first things first. TakeOff!
Call 800/447-4700 toll free and ask for the Cessna "TakeOff" operator. She'll give you the location of your nearest participating Cessna Pilot Center. In Illinois call 800/322-4400. In Canada call 800/261-6362. In Toronto call 445-2231.
*No purchase necessary. Void where prohibited by law. Sweepstakes closes December 31, 1979. Participation via entry forms distributed through flight schools or instructors in Maryland, Missouri and Wisconsin is void.
Cessna
PILOT CENTER
Discover Flying
LEARN FROM THE LEADER
PROFESSIONAL PILOT COURSE
Cessna
AIRCRAFT COMPANY / WICHITA, KANSAS
Member of GAMA
TakeOff
04CW
CESSNA
Number one in pilot education for some very down-to-earth reasons.

Anions of the Alkali Metals

Alkali metals, such as sodium, are familiar as cations: positively charged ions formed when atoms lose an electron. Now it has been learned they can also gain an electron to form very reactive anions

by James L. Dye

The alkali metals are a group of elements whose most notable and most familiar chemical property is their eagerness to give up an electron. The alkali metal sodium, for example, readily donates an electron to chlorine, forming sodium chloride. Metallic sodium is so strongly disposed to get rid of an electron that it will even split a water molecule, displacing a hydrogen atom and forming sodium hydroxide; the reaction can be a violent one. Because the metallic form of sodium is so reactive it does not exist in nature; the element is found only as a positive ion, or cation, denoted Na^+. The cation forms when the atom surrenders one of its electrons to some other chemical species.

These properties of the alkali metals have been known since the beginning of chemistry, and it has therefore come as a surprise to learn that alkali metals can also accept an electron, performing in a role precisely the reverse of their usual one. The addition of an electron to the neutral sodium atom, for example, forms the negatively charged ion, or anion, Na^-. This anion has been known for some time to be a stable species in gaseous sodium; what is more important, it has recently been discovered in solutions and even as a component of a crystalline salt. It appears that the anions of all the other alkali metals can also be prepared; those of potassium (K^-), rubidium (Rb^-) and cesium (Cs^-) have already been observed. It may even be possible to prepare salts containing the simplest possible anion: the immobilized electron.

The key to the preparation of the alkali metal anions, curiously, is the trapping of the alkali cations in an organic molecule with a cagelike structure. Ordinarily any negative metal ions present in a solution would quickly react with positive ions to yield neutral atoms of the metal. When the cations are sequestered in an organic cage molecule, the resulting complex is so stable that the "backsliding" reaction is prevented. The negative ions and the crystals containing them are nonetheless highly reactive. They cannot be exposed to air or moisture, and they are stable for long periods only when stored at low temperature.

The alkali metals comprise six elements: lithium, sodium, potassium, rubidium, cesium and francium. (The last is a rare, radioactive species.) In the usual arrangement of the periodic table of the elements they are listed in the first column, under hydrogen. This grouping reflects similarities in the chemical properties of all the alkali metals and in their underlying electronic structures: each has a solitary electron in its outermost, or valence, shell of electrons.

In any neutral atom, of course, there must be just enough electrons to balance the electric charge of the protons in the nucleus. The electrons tend naturally to arrange themselves in a configuration of minimum energy, but they can do so only within the bounds allowed them by the rules of quantum mechanics, which confine the electrons to "orbitals." In classical physics an orbital might be said to describe the trajectory of an electron; a more precise, quantum-mechanical interpretation of the orbital considers it as defining the probability of finding the electron at a given position. Equivalently, the orbital can be said to represent the distribution of a smeared-out electronic charge.

The sequence in which the available orbitals are filled is determined by a rule that forbids any two electrons in an atom from occupying precisely the same state. The first orbital to be filled is given the label 1*s;* electrons in the 1*s* orbital are distributed throughout a small spherical volume centered on the nucleus. The distribution is completely symmetrical and defines no preferred directions in space. The 1*s* orbital can be occupied either by one electron or by two electrons; in the latter case the two electrons have opposite values of spin angular momentum. (In quantum mechanics the electron must spin on its axis perpetually, and that spin can have only two possible orientations in space.)

The second orbital to be filled is designated 2*s*. It is like the first one in shape, but the two electrons that can fill it have a higher energy and are therefore dispersed throughout a larger sphere.

It is at this point in the construction of an atom that the first nonspherical orbitals are encountered. They are designated *p*-type orbitals and there are three of them, all given the label 2*p*. Each of the three 2*p* orbitals has two lobes, one on each side of the nucleus, and they are arranged in space so that each orbital is perpendicular to the other two. The six electrons that fill the 2*p* orbitals (in pairs with opposite spin) have one unit of orbital angular momentum in addition to their spin angular momentum. In other words, in addition to rotating on their axes they revolve around the nucleus.

In order to fill the 1*s*, 2*s* and 2*p* orbitals 10 electrons are required. If any more electrons are added to the atom, they will go first into the 3*s* orbital and then into three 3*p* orbitals, which are again similar in shape to the equivalent orbitals of lower energy but are larger. Heavier atoms have still other kinds of orbitals (labeled *d, f* and *g*) with more complicated geometries and higher values of orbital angular momentum.

It is of the most fundamental importance to all chemistry that an atom with no vacancies in any of its *s* or *p* orbitals shows remarkable stability. Such atoms make up the noble-gas series of elements, whose inertness testifies to their stability. The first of these is helium, with just two electrons exactly filling the 1*s* orbital. Neon, with 10 electrons, has the 1*s*, 2*s* and 2*p* orbitals completed; argon has filled orbitals through 3*s* and 3*p*. All these elements are most reluctant either to accept or to donate an electron.

The alkali metals can be regarded as noble gases with one extra electron (and, of course, one extra proton in the nucleus). Lithium has the filled 1*s* orbital of a helium atom, and one additional electron half filling the 2*s* orbital. Sodium has filled 1*s*, 2*s* and 2*p* orbitals, as in neon, and in addition has one 3*s* electron. The remaining alkali metals are similar in structure; in each of them the outermost electron is alone in an *s*-type orbital.

The strong tendency of the alkali metals to lose an electron and form a cation can now be understood. With the removal of the valence electron, each of the alkali metals takes on the exceptionally stable structure of a noble gas. Even though the resulting ion has an unbalanced positive charge, in the presence of an electron acceptor it is stabler than the neutral atom.

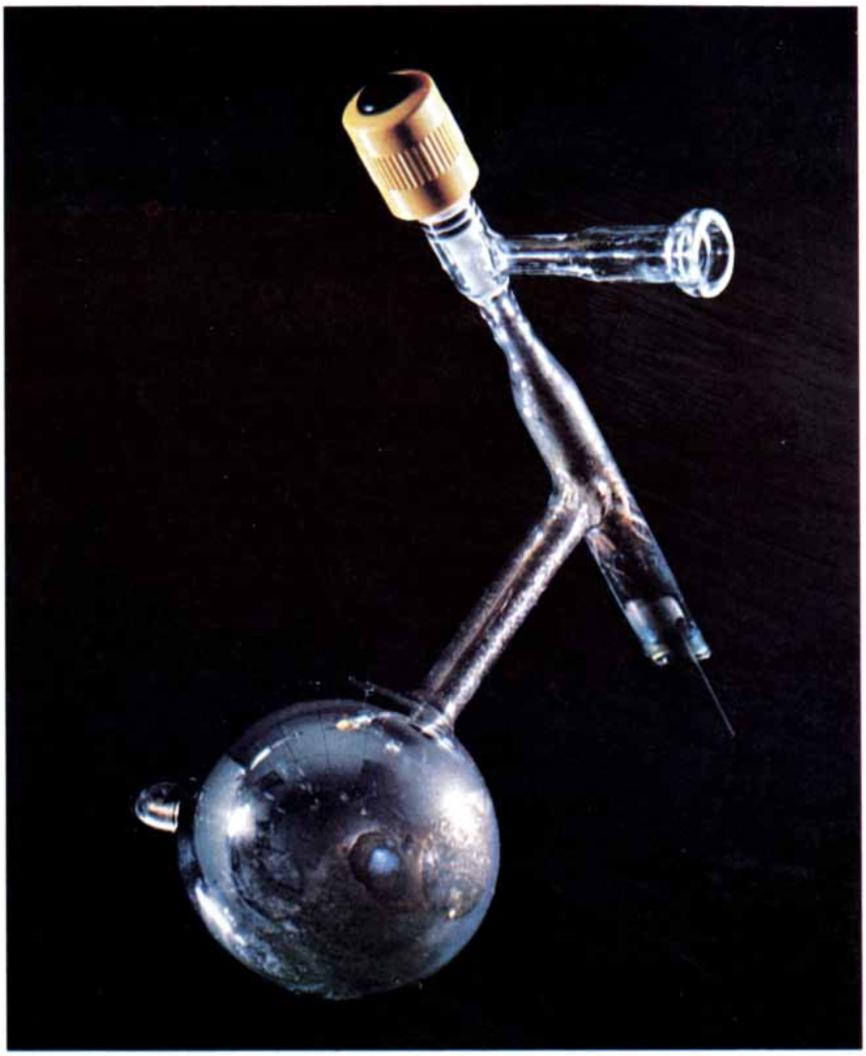
Adding an electron to an alkali metal results in a configuration that is much less strongly favored than the noble-gas orbitals of the cation. Nevertheless, the extra electron does fill an *s*-type orbital, and a filled orbital is somewhat stabler than a half-filled one. Under certain carefully contrived circumstances that slight gain in stability is enough to favor the existence of alkali metal anions.

The chemistry of ionized atomic species is most conveniently studied in solutions. In order to study the alkali metals in that way we need a solvent capable of dissolving them. The sequence of investigations that has recently culminated in the observation of alkali metal anions began in Germany in 1864 with the discovery by W. Weyl that the alkali metals dissolve in anhydrous liquid ammonia (NH_3). They yield solutions colored a beautiful deep blue.

The process by which an ionic substance dissolves depends on an intimate electrical interaction between the molecules of the solvent and the ions of the solid. The solvent must generally be a polar liquid, that is, it must be made up of molecules that, although they are electrically neutral overall, have regions of charge. Ammonia is an example: in the liquid the ammonia molecule has a configuration in which the three hydrogen atoms are clustered together on one side of the nitrogen atom. The nitrogen tends to acquire a negative electric charge, balanced by a positive charge on the hydrogens, and as a result the molecule is an electric dipole: it has a negative end and a positive end.

When a salt such as sodium chloride dissolves, it breaks up into negative and positive ions, in this case Na^+ and Cl^-. What holds these ions in solution is an interaction with the polar molecules of the solvent. The positive charge of an Na^+ ion, for example, attracts to it the more negative regions of the solvent molecules, and it is soon surrounded by a layer of oriented molecules. The negative ion attracts a similar coating of solvent molecules, but of course they are oriented in the opposite way. It is these layers of oriented solvent molecules that stabilize the solution and prevent the ions from aggregating to form a salt.

Most metals are much less readily dissolved than salts, simply because there is no easily formed anion. The alkali metals dissolve in ammonia because of their extraordinary eagerness to cast off an electron. The resulting solution contains metal cations, such as Na^+, indis-



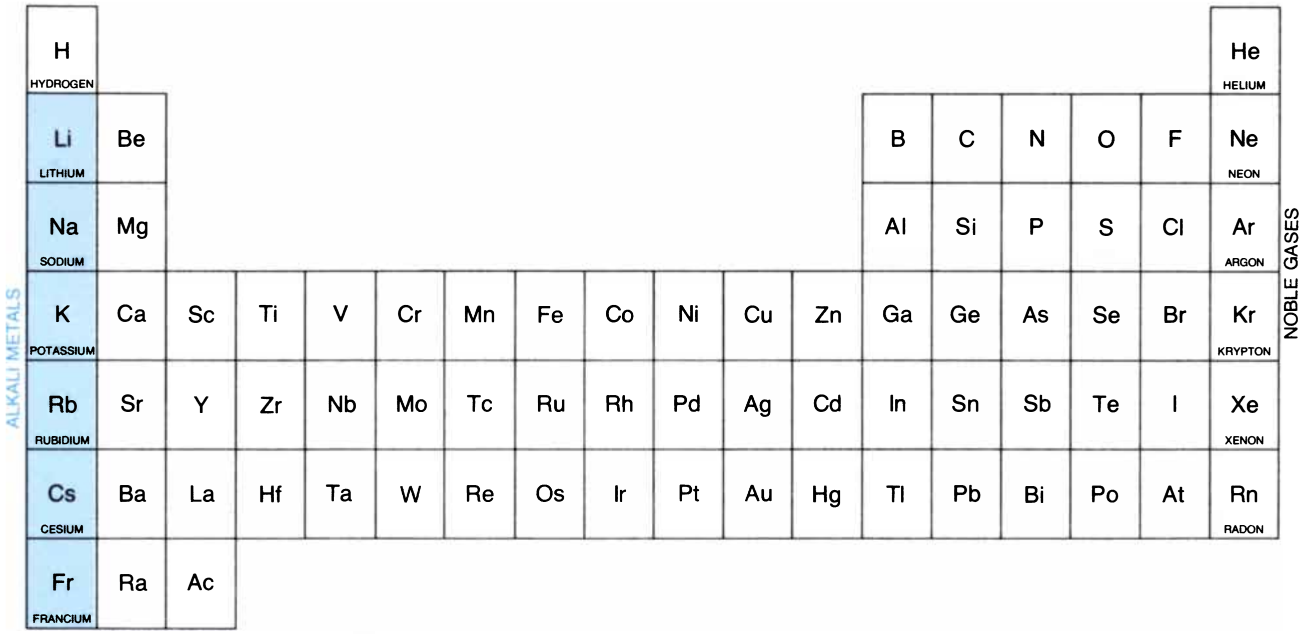
SODIUM ANIONS are an essential constituent of a blue liquid and of a bronze-colored film formed inside a glass vacuum system. In the preparation of these substances a layer of sodium metal was first deposited on the inner surface of the globe at left; some of the sodium metal was then dissolved in ethylamine. In the solution positive sodium ions (cations) are trapped by an organic molecule, called a cryptand, that has a cagelike structure; negative ions (anions) are dispersed in the solvent and give the solution the deep blue color visible at the bottom right. The film, which precipitates onto the walls of the vessel when some of the solvent is evaporated, is made up of anions and trapped cations. The plastic device at the top is a vacuum valve.

CRYSTALS OF A SALT are composed of sodium anions and a complex of sodium cations and cryptand molecules. Like any salt, this one is formed when one chemical species donates an electron to another species, creating a pair of ions; what is unusual here is that the ultimate donor of the electrons and the acceptor are forms of the same element. The crystals were precipitated from a solution by cooling. Their metallic luster is deceptive: the salt is not a metal but a semiconductor. The sodium cations are highly reactive. The salt cannot be exposed to the air, and even in a vacuum it slowly decomposes unless it is stored at low temperature.

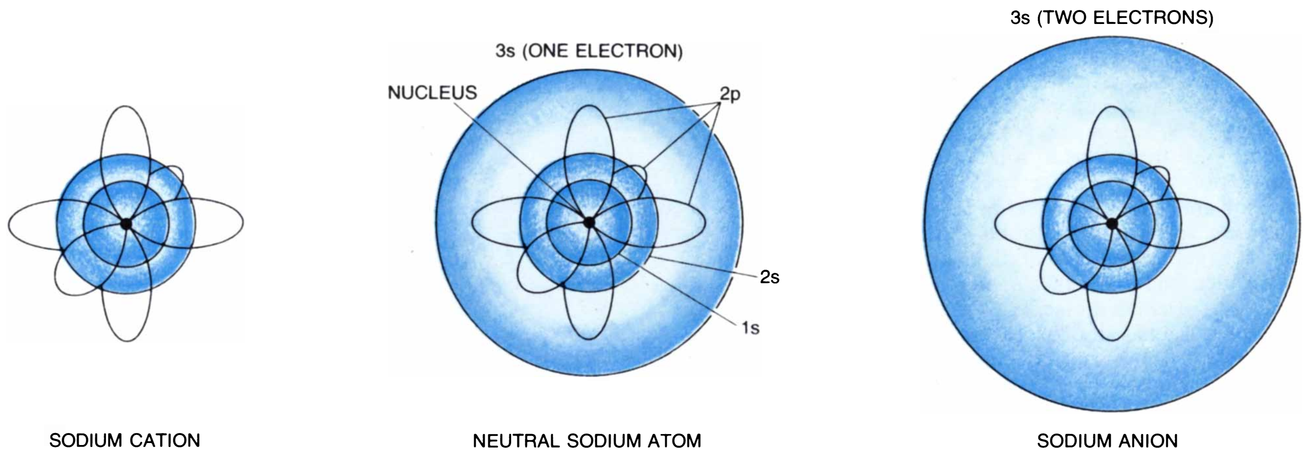
tinguishable from those of a dissolved salt, but the negative ions are quite unusual: they are simply electrons that have been released into the solution. Such electrons have been given the name solvated electrons. They behave much like any other negatively charged species, attracting and orienting the polar solvent molecules, but because they are thousands of times lighter than any other ion they are considerably more mobile [see "The Solvated Electron," by James L. Dye; SCIENTIFIC AMERICAN, February, 1967].

How can we know what chemical species are present in a solution? How can it be shown, for example, that solutions of the alkali metals in ammonia contain the solvated electron and not some other negative ion, perhaps even true alkali metal anions? A great many techniques could be applied to these problems, but we shall be particularly concerned with just two of them. These two techniques characterize ions in solution according to their magnetic properties and according to the wavelengths of radiation they absorb.

The magnetism of electrons derives ultimately from their angular momentum. The spinning of an individual electron, for example, gives rise to a small magnetic field in much the same way that a current of electrons flowing through a loop of wire does. Only in some atoms are these fields observable. In a filled atomic orbital, it will be remembered, two electrons are always paired with opposite spins; as a result they also have oppositely directed magnetic fields, and the two fields exactly cancel. In an atom that has no incomplete orbitals all the fields cancel, leaving no net magnetism. A substance made of such atoms is said to be dia-



ALKALI METALS are the elements listed under hydrogen in the first column of the periodic table. They are grouped together because of similarities in their electronic structures and consequent similarities in their chemical properties. In the configuration of its electrons, each of the alkali metals can be regarded as a noble-gas atom with one additional electron. The noble gases (listed at the far right) are the most stable of all the elements, and alkali metals readily give up an electron to form a cation with the electronic structure of a noble gas. The extra electron is so readily jettisoned that it can sometimes be donated to another alkali metal atom, which becomes an alkali anion.



GAIN OR LOSS OF AN ELECTRON profoundly alters the chemical properties of the alkali metals. A neutral sodium atom (*center*) has two electrons in each of five interior orbitals and a solitary electron in the outermost, or valence, orbital. Orbitals define the spatial distribution of the electronic charge and each can hold no more than two electrons. The lowest-energy orbital, which is designated 1*s*, has a spherical form; the next one, 2*s*, is also spherical but larger. There are three 2*p* orbitals, each with two lobes. Finally, the 3*s* orbital is again spherical. In a sodium cation (*left*) the 3*s* electron has been removed and the ion has the electronic configuration of neon, with filled 1*s*, 2*s* and 2*p* orbitals. In the sodium anion (*right*) an extra electron has been acquired, so that the 3*s* orbital is filled with two electrons.

magnetic; it has no magnetism of its own, and if it is placed in an external field, it weakly opposes it, partly expelling the lines of force.

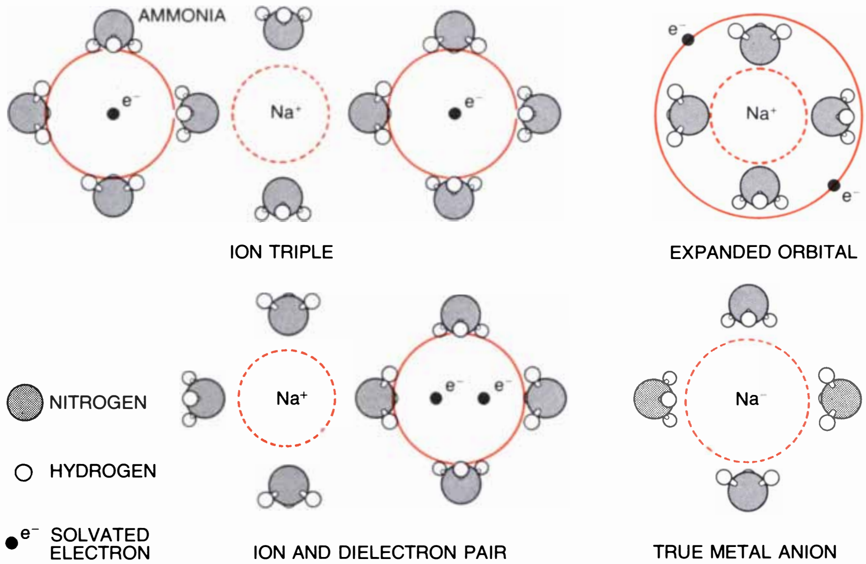
Atoms with incomplete orbitals and therefore with unpaired electron spins can have a net magnetic field. If this field is very strong, as it is in iron and some closely related metals, the material is said to be ferromagnetic. In a ferromagnet the spins of the unpaired electrons spontaneously line up, at least over microscopic domains, and the material exhibits "permanent" magnetism. In most magnetic substances, including the alkali metals, the fields associated with each atom are much smaller, too small to spontaneously align the electron spins. The fields can be made to line up, however, and the substance can be weakly magnetized, by a strong external field. Substances with these properties are called paramagnetic.

The solvated electrons in an alkali-metal-ammonia solution should behave like unpaired electrons in an atom and hence should make the solutions paramagnetic. Moreover, the strength of the paramagnetism should be proportional to the concentration of solvated electrons. Such solutions have indeed been shown to be paramagnetic, and there is no question solvated electrons are present in them. The strength of the induced field, however, is not proportional to the concentration, and so it appears that some other ionic species is also present. The nature of the deviations from proportionality indicates that some of the electrons are pairing up to form a diamagnetic species.

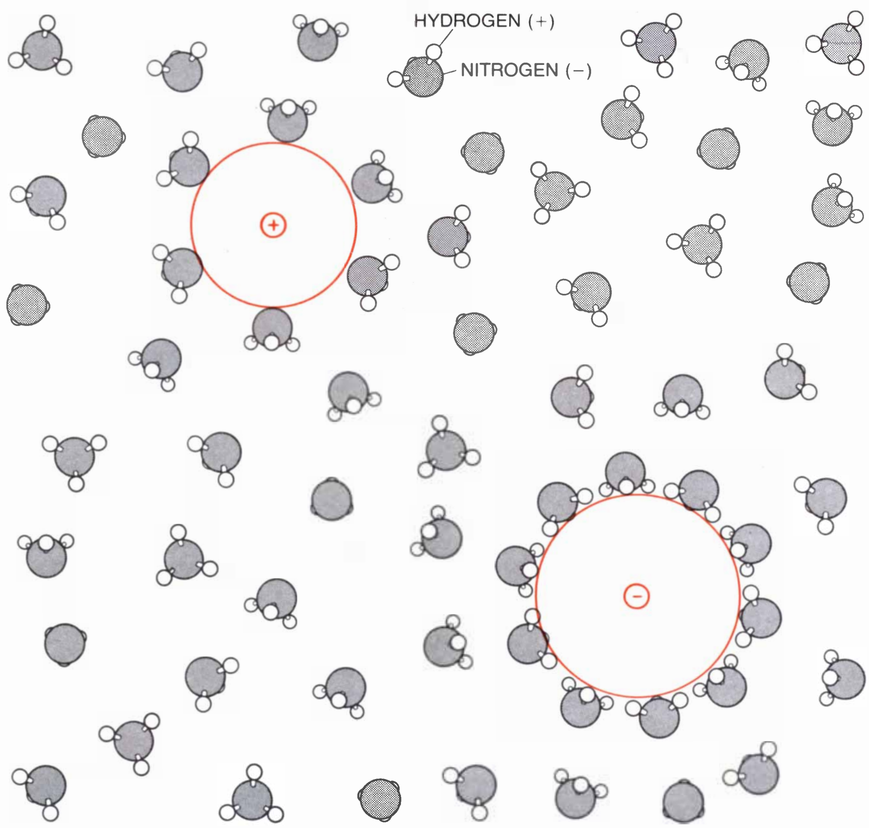
The nature of the diamagnetic ion in metal-ammonia solutions has been a subject of controversy for decades, and the matter is not yet settled. An obvious candidate is the alkali metal anion, since an alkali metal that acquires an extra electron has all its orbitals filled and must be diamagnetic. The first proposal of this kind was made in 1953 by the German investigator Werner Bingel, although the structure he had in mind was not that of a true metal anion; rather, he imagined that an ordinary metal cation might attract to its vicinity a pair of electrons bound to each other. The possibility that metal-ammonia solutions might contain true alkali metal anions was first discussed in 1965 by Thomas R. Tuttle, Jr., Sidney Golden and Charles Guttman of Brandeis University.

Metal anions may indeed exist in ammonia solutions, but it is fair to say there is no explicit evidence for their presence. Moreover, the absorption spectrum of the solutions argues, if somewhat indirectly, that the anion cannot be the only diamagnetic species.

The deep blue color of metal-ammonia solutions results from the strong absorption of light at the red extreme of



IONS IN SOLUTION are prevented from aggregating to form a solid by an electrical interaction of the ions and the solvent molecules. For the alkali metals (and for many other substances) only liquids with polar molecules are effective solvents. A polar molecule is one with regions of strong electric charge. Ammonia is an example: the nitrogen atom in liquid ammonia tends to acquire a negative charge, which is balanced by a positive charge on the hydrogen atoms. An ion in such a liquid attracts to itself a layer of oriented solvent molecules, which insulate it from the ions of opposite charge. Ammonia dissolves all the alkali metals.



POSSIBLE NEGATIVE IONS in solutions of the alkali metals include several candidate structures besides the genuine alkali metal anion. Some of these structures are depicted schematically for the case of sodium dissolved in ammonia. A sodium cation (Na^+) might attract to its vicinity two independent solvated electrons, forming an "ion triple." Or the cation might attract a "dielectron pair," an entity made up of two electrons inside a single "cage" of solvent molecules. Two electrons might also be associated with a cation in an "expanded orbital" that encompasses solvent molecules as well as the cation itself. All these configurations are distinguished by the fact that solvent molecules are interposed between a sodium cation and the two outermost electrons; in the true alkali metal anion the two electrons occupy a normal 3*s* orbital.

the visible spectrum and in the infrared region. Such long-wavelength absorption is common in materials containing electrons that are not confined to particular atoms; metals are a good example. In the ammonia solutions light and infrared radiation are absorbed by both the solvated electron and the diamagnetic species. Significantly, the absorption spectrum is not affected by the choice of alkali metal in the solution. If the diamagnetic species were an alkali metal anion, then the spectrum would be slightly different for each one; evidently the electrons interact strongly only with other electrons, not with the metal ions.

It may be that metal anions do not exist in ammonia solutions because ammonia is too good a solvent. The cation may be so strongly stabilized by its interaction with the polar ammonia molecules that no other form of the metal can survive. Any anion that might form would tend to break up into two solvated electrons and a cation.

The choice of potential solvents for the alkali metals is limited. Many liquids are not polar enough to dissolve the metals; simple hydrocarbons, such as petroleum fractions or benzene, for example, are inert to the alkali metals. Many of the polar solvents, on the other hand, react with the metals to form compounds with them. The reaction of sodium metal and water (a highly polar solvent) has already been mentioned. Similar reactions, including some violent ones, eliminate from consideration the alcohols and ketones and the chlorinated hydrocarbons.

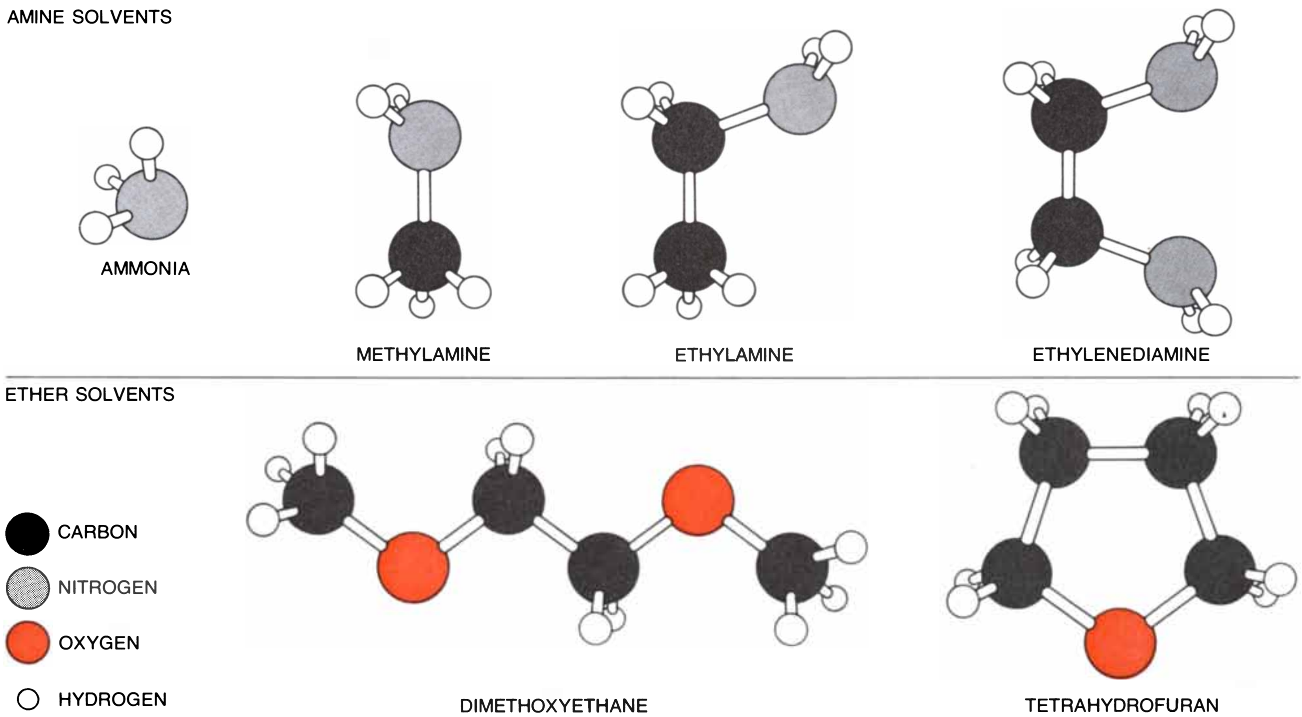
One group of solvents for the alkali metals consists of molecules related to ammonia. Methylamine, for example, has the formula CH_3NH_2; it can be regarded as an ammonia molecule in which one hydrogen atom has been replaced by a methyl group ($-CH_3$). Lithium dissolves readily in liquid methylamine, although sodium, potassium, rubidium and cesium are much less soluble. Another close relative of ammonia is ethylenediamine, with the structure $NH_2–CH_2–CH_2–NH_2$. One might think of this molecule as consisting of two methylamine molecules joined back to back. In my laboratory at Michigan State University my colleagues and I have found ethylenediamine a useful solvent for the alkali metals.

Besides the simple amines, the other main group of solvents for the alkali metals is the ethers, the class of organic molecules incorporating the structure $–CH_2–O–CH_2–$. The familiar ether employed as an anesthetic is diethyl ether ($CH_3–CH_2–O–CH_2–CH_3$), but the alkali metals do not dissolve in this liquid. The more effective ether solvents are those with more than one oxygen atom (polyethers), such as dimethoxyethane ($CH_3–O–CH_2–CH_2–O–CH_3$). Another solvent is tetrahydrofuran, an ether with a cyclic structure. It should be pointed out that none of these substances is a very good solvent of the alkali metals, and it is not possible to prepare concentrated solutions with them, but enough metal can be dissolved in them to observe the properties of the chemical species formed.

The optical properties of these solutions are significantly different from those of metal-ammonia solutions. The absorption band mainly in the infrared associated with the solvated electron can still be observed, but there is also a distinctly different absorption feature at shorter wavelengths. What is most important, this new absorption band has a different position for each of the alkali metals. In 1969 S. Matalon, Golden and Michael Ottolenghi suggested that this new absorption band is produced by metal anions. Their proposal has since been proved to be correct.

It is a simple matter to show that a solution contains some species with a net electric charge of -1, but a genuine metal anion is not the only possible structure that has this property. For example, the actual species might be a metal cation that has loosely bound to it two independent solvated electrons. Another possibility is that the two electrons might occupy an "expanded orbital" having a metal cation at its center, as in a true anion, but also encompassing several solvent molecules. Or a cation might have associated with it a pair of electrons tightly bound to each other but only loosely bound to the metal ion.

All these structures can be unequivo-



SOLVENTS FOR ALKALI METALS include two main classes of compounds: ammonia and its derivatives (the amines), which contain nitrogen, and ethers, which contain oxygen. The molecules of these liquids are sufficiently polar to dissolve metals; equally important, they do not react with the metals chemically. Except in ammonia, only modest concentrations of the metals can be attained.

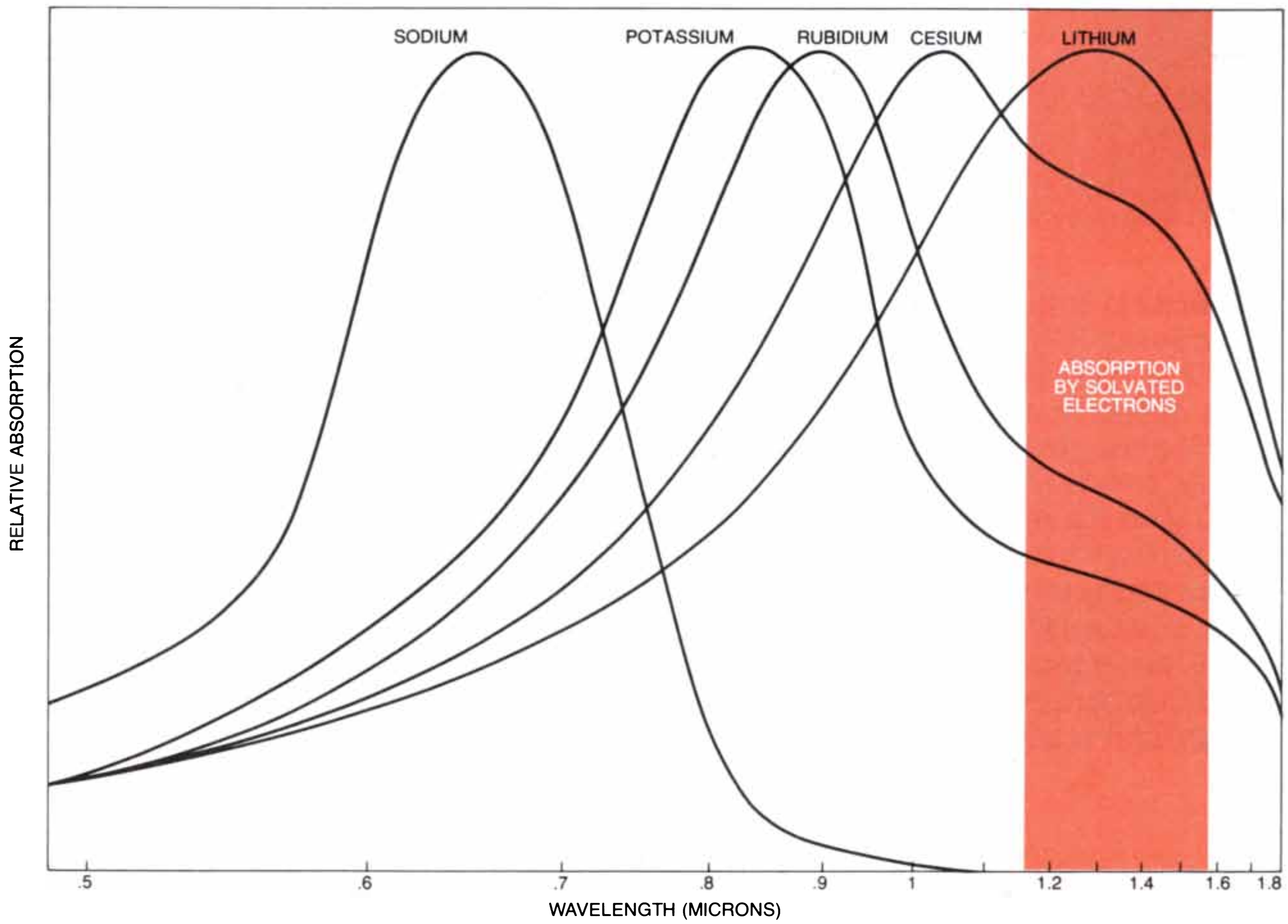
cally distinguished from the genuine metal anion. They are all characterized by the presence of a metal cation immersed in solvent molecules that have the orientation expected of molecules shielding a positive ion; the two electrons are elsewhere, outside a layer of solvent molecules. In the true metal anion the two electrons are tightly bound to the metal and the solvent interacts with the ion as a single, negatively charged unit.

Measurements of optical and magnetic effects in the amine and ether solutions of the alkali metals provide suggestive evidence for the existence of genuine metal anions. In my laboratory my colleagues and I have shown that the optical absorption band is produced by two equivalent or interchangeable electrons; that is the case for the true anion and for some other models as well. With Leon M. Dorfman of Ohio State University we measured the rate of formation of the sodium anion, Na^-, from solvated electrons and sodium cations. Other investigators showed that the negatively charged species could be broken down into a cation and two solvated electrons, which subsequently recombined. By 1970 it appeared likely that the metal anion had been discovered, but other interpretations could not be ruled out. One of the fundamental impediments to the further study of the ions was the low solubility of the metals.

In 1970 Vincent A. Nicely, working in my laboratory, suggested that the solubility of the alkali metals might be improved with the aid of a new class of compounds called crown ethers. These ethers, which had been developed by Charles J. Pedersen of E. I. du Pont de Nemours & Company, had already been shown to greatly increase the solubility of ordinary alkali salts. As suggested by the name "crown," they have the form of a crenelated ring. Cations nest in the center of the ring, which has an affinity for positive charges; the entire complex can then readily be solvated by a polar liquid such as an amine or an ether.

Since the crown ethers are chemically not too different from the polyethers we had long employed as solvents, it seemed reasonable to suppose they would not react strongly with the alkali metals. That supposition turned out to be correct. As we had hoped, solubilities increased by many orders of magnitude.

Another class of molecules that form complexes with the alkali metals was later developed by Jean-Marie Lehn of Louis Pasteur University in Strasbourg. These agents Lehn named cryptands because their characteristic structure is a three-dimensional "crypt" that encloses a cation, completely shielding it from the surrounding solvent molecules. The cryptands of interest here all have two nitrogen atoms connected by three rather long organic strands. Each strand is an ether with one, two or three oxygen atoms. Simply designating the number



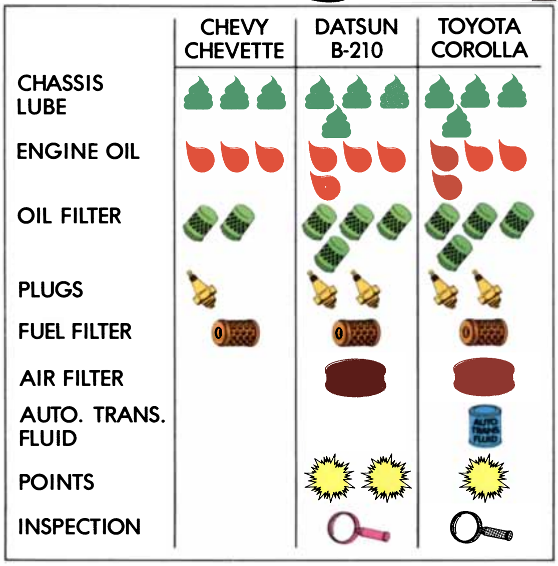
ABSORPTION OF LIGHT by solutions of alkali metals in amines and ethers suggests the presence of two light-absorbing species. At a wavelength of about 1.4 microns, in the infrared part of the spectrum, an absorption peak has been attributed to solvated electrons. The wavelength of this peak is the same no matter which metal is present in the solution. A second peak, at shorter wavelengths, is now known to result from absorption by alkali metal anions; significantly, it falls at a different wavelength for each metal. In the spectra of potassium, rubidium and cesium both of these absorption features can be observed. Lithium anions do not seem to be present in the solutions, since only the peak associated with solvated electrons can be detected in the lithium spectrum. Sodium, on the other hand, apparently forms anions so readily that no solvated electrons are present. The spectra are for solutions of the alkali metals in ethylenediamine.

Compare Chevy Chevette to the two leading imports.

In 25,000 miles of normal driving, our Chevette owner's manual recommends a lube and oil change just three times, a change of oil filter two times, spark plugs and fuel filter just once. Compare that with what Datsun and Toyota recommend in their owner's manuals.

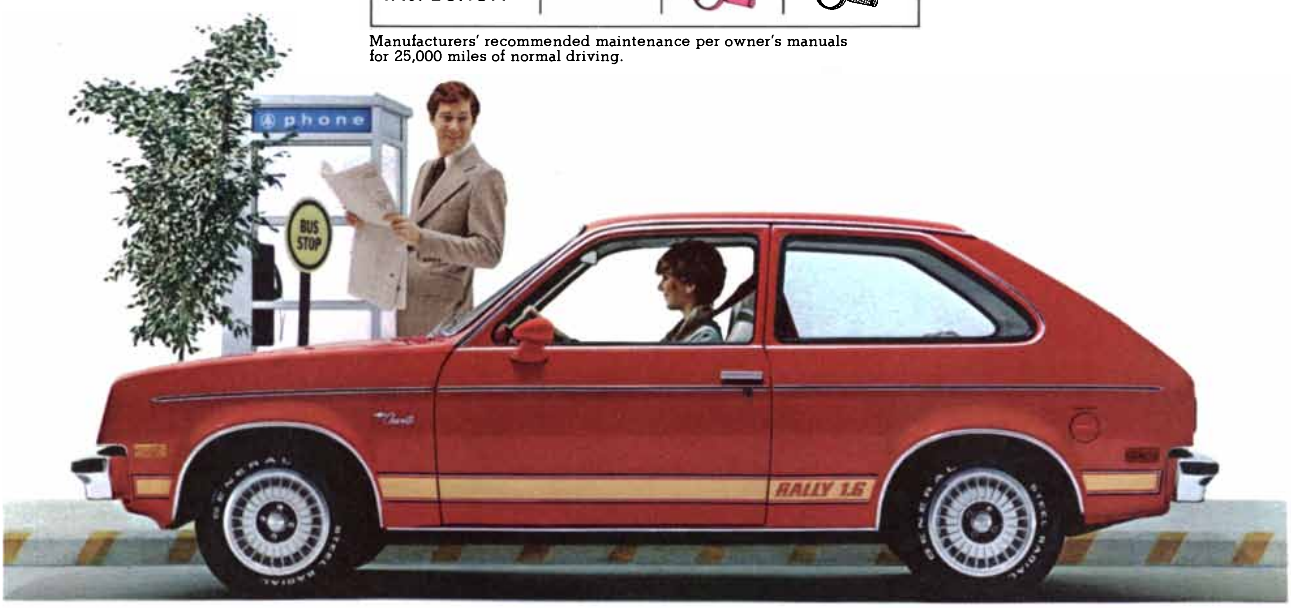
How important is it to you that Chevette offers extended service intervals? Well, it means you can expect to save time and money. But more important, it reflects the overall attitude that Chevy engineers brought to the design of Chevette, a car engineered to *give* service, not get it.

See any one of our nearly 6,000 Chevy dealers and let him go over the fine points of Chevette with you. We think you'll be impressed.



Manufacturers' recommended maintenance per owner's manuals for 25,000 miles of normal driving.

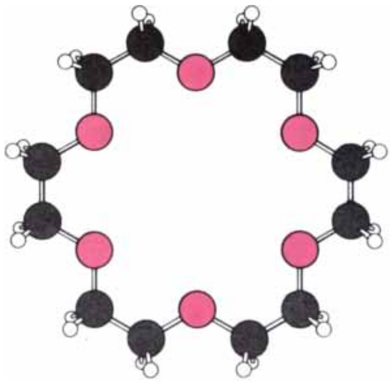
Chevy Chevette



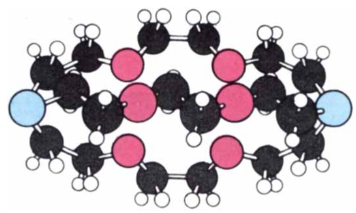
It'll drive you happy.

GM
MARK OF EXCELLENCE

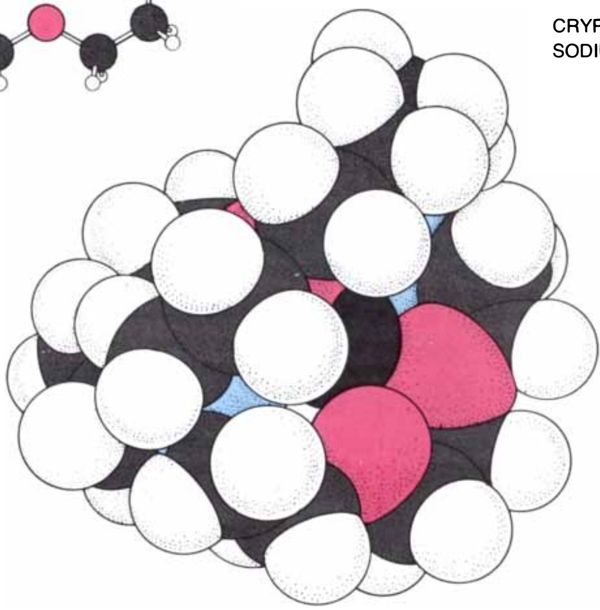
CROWN ETHER



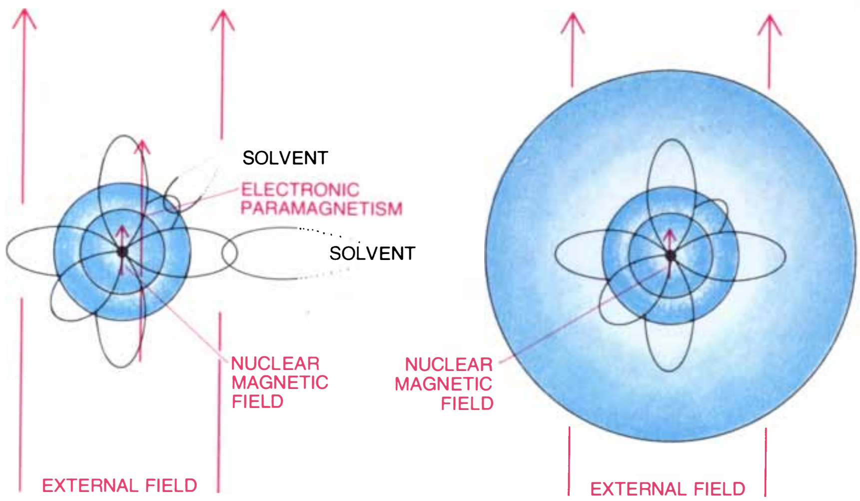
CRYPTAND



CRYPTAND WITH SODIUM CATION



CROWN ETHERS AND CRYPTANDS envelop the cations of alkali metals, improving the solubility of the metals and incidentally aiding in the formation of alkali metal anions. The structures of the two classes of molecules are shown schematically at the top; cryptands are more efficient because their three-dimensional cavity more effectively isolates the ion from the solvent. A space-filling representation of the same cryptand reveals the snug fit of a sodium cation in the cavity of the molecule. In solution and in salts cryptands confine and stabilize alkali metal cations, preventing them from recombining with the anions to precipitate metal.



MAGNETIC PROPERTIES OF AN ATOM are influenced by the atom's state of ionization and by its chemical environment. The magnetic field of the atomic nucleus can be measured (by the technique called nuclear-magnetic-resonance spectroscopy), and such measurements can provide indirect information about the state of the atomic electrons. In atoms or ions that have their outermost electrons in *p*-type orbitals (*left*), interactions with solvent molecules tend to enhance the effect that an external magnetic field has on the nuclear field; this phenomenon is called paramagnetism. If the outer electrons are in *s*-type orbitals (*right*), the paramagnetic enhancement is absent. Hence it should be possible to discriminate by nuclear-magnetic-resonance spectroscopy between cations and anions in solutions that contain alkali metals.

of oxygen atoms in the strands serves to identify a cryptand. The 1,1,1-cryptand, with one oxygen in each strand, has a small cavity that can accommodate only a hydrogen ion (a proton) or the lithium cation Li^+. The larger crypts of the 2,2,1- and 2,2,2-cryptands are optimal for the sodium ion Na^+.

Cryptands can provide enormous increases in solubility. In pure ethylamine, for example, sodium is virtually insoluble; when 2,2,2-cryptand is present, several grams of sodium per liter of solvent can be dissolved. Solubility is increased by a factor of at least 400,000.

It is the stability of the metal-cryptand complex that makes possible these high solubilities. Paradoxically, the stability of the cation within the cavity of the cryptand molecule also leads to high concentrations of metal anions. In fact, by adjusting the amounts of metal and cryptand we can control the ionic species present in the solution.

When metal atoms and cryptand molecules are present in equal numbers, virtually all the metal atoms become cations trapped in the organic cages; this rule holds for potassium, rubidium and cesium. The predominant negative species in such solutions is therefore the solvated electron, since there are simply no metal atoms available to form anions. If the metal is present in excess, the trapped cation remains the principal positive species, but now the extra metal atoms combine with electrons to form alkali metal anions. These general rules hold for all the alkali metals except lithium and sodium. (Francium has not been investigated.) In amines and ethers lithium is apparently reluctant to form an anion no matter what the conditions. Sodium, on the other hand, forms an anion so stable that solvated electrons are rarely seen.

With the concentrated solutions that can be prepared with crown ethers and cryptands it is much easier to study the properties of alkali metal anions. For example, it becomes possible to apply the technique known as nuclear-magnetic-resonance spectroscopy. With this technique we can single out for attention a particular element in a solution and deduce its chemical state.

Atomic nuclei can have spin angular momentum just as electrons do, and nuclei can therefore also generate magnetic fields. All the alkali metal nuclei are magnetic; the sodium nucleus, for example, has four possible spin states (contrasted with the two of the electron) and four possible orientations of its magnetic field with respect to any external field. The nuclear magnetism actually has no detectable influence on the chemistry of an atom, but it serves as a convenient probe of the electronic magnetic field, which *is* affected by the chemical environment.

In practice the nuclear magnetic field

NAVY NUCLEAR PROPULSION. THE FASTEST WAY UP IN NUCLEAR ENGINEERING.

If you want to get into nuclear engineering, start by getting into the Nuclear Navy.

The Navy operates more than half the reactors in America. So our nuclear training is the most comprehensive you can get. You start by earning your commission as a Navy Officer. Then we give you a year of advanced nuclear technology, training that would cost you thousands if you could get it in graduate school. During your career, you'll get practical, hands-on experience with our nuclear powered fleet. Maybe you'll work on a nuclear submarine, maybe a nuclear cruiser. But wherever you work, you'll really get to prove your worth—as a young Nuclear Propulsion Officer entrusted with the most advanced technical equipment known to man.

If that sounds like the kind of responsibility you're looking for, speak to your Navy recruiter. He can tell you if you qualify as a Nuclear Propulsion Officer Candidate. Or call toll free **800-841-8000.** (In Georgia, 800-342-5855.)

Navy Nuclear Propulsion Officer. Some men wait for the future. He lives it now.

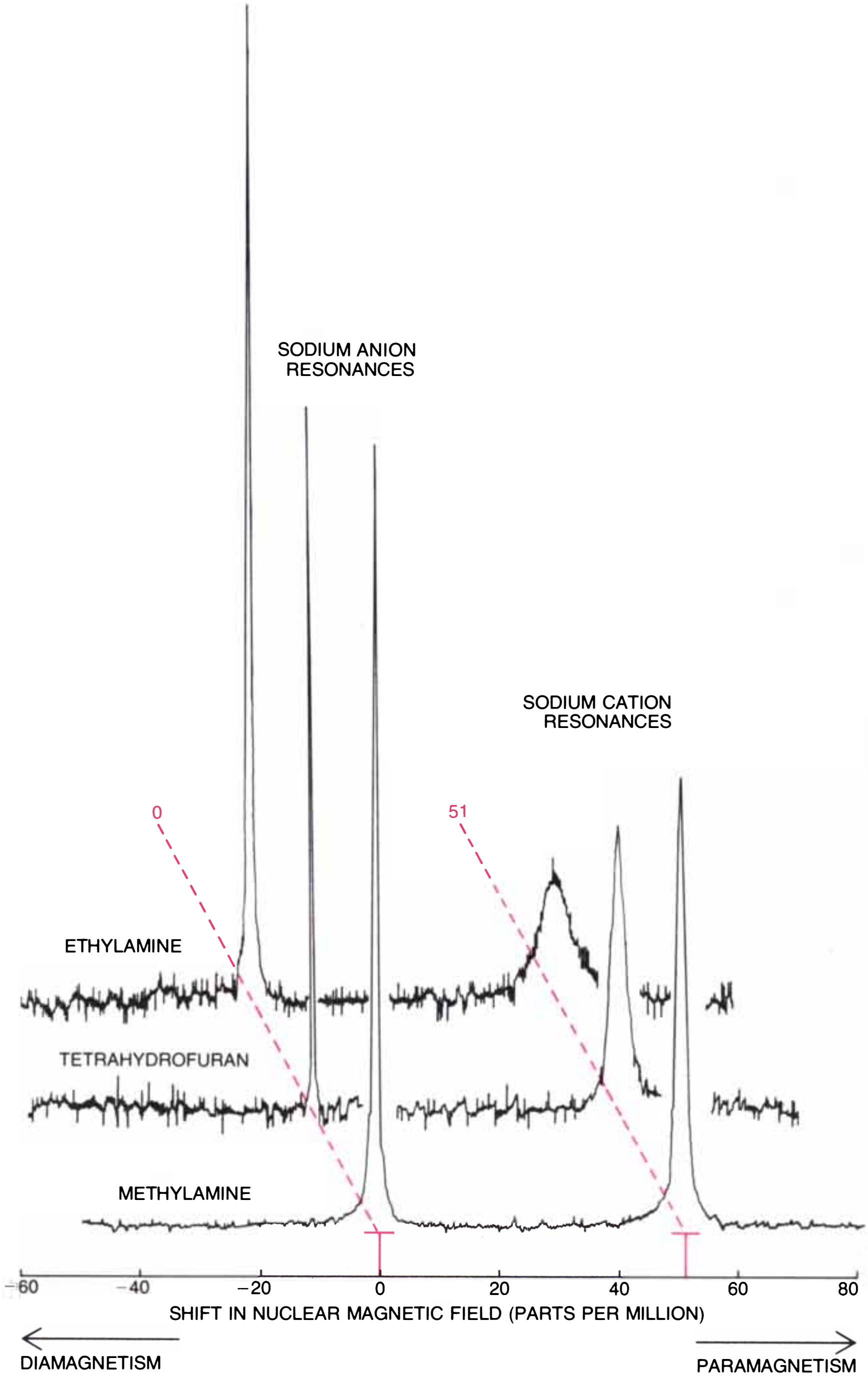
NAVY OFFICER. IT'S NOT JUST A JOB, IT'S AN ADVENTURE.

is examined by forcing the nucleus to change its state of spinning, to execute a "spin flip." Since each spin state involves a different orientation of the nuclear magnetic field with respect to an external field, each state has a different characteristic energy. The magnitude of the energy difference is proportional to the strength of the external field as that field is perceived at the nucleus. The energy required for the nucleus to shift to a state of higher energy can be supplied by a photon, or quantum of electromagnetic energy; the nucleus can fall back to a state of lower energy by emitting a photon. The photon energies required are generally those equivalent to radio frequencies of from two megahertz to 300 megahertz. When both the external magnetic field and the radio frequency have been adjusted to precisely the correct values, a resonance is observed and the radiation is strongly absorbed.

The experimenter studying nuclear magnetic resonance in solutions can control the external field in which an atom is immersed, but that is not necessarily the field present at the nucleus. The electrons surrounding the nucleus may oppose the external field (diamagnetic behavior), or they may enhance it (paramagnetic behavior). In order to compensate for diamagnetic shielding by electrons the external field must be increased, whereas a paramagnetic enhancement calls for a decrease in the external field. These "chemical shifts" therefore embody information about the environment of the nucleus.

It is the technique of nuclear-magnetic-resonance spectroscopy that has provided convincing evidence for the existence of alkali metal anions in solution. We can see how magnetic resonance is able to discriminate between genuine anions and other possible structures by considering the state of the electrons in a sodium atom in various environments.

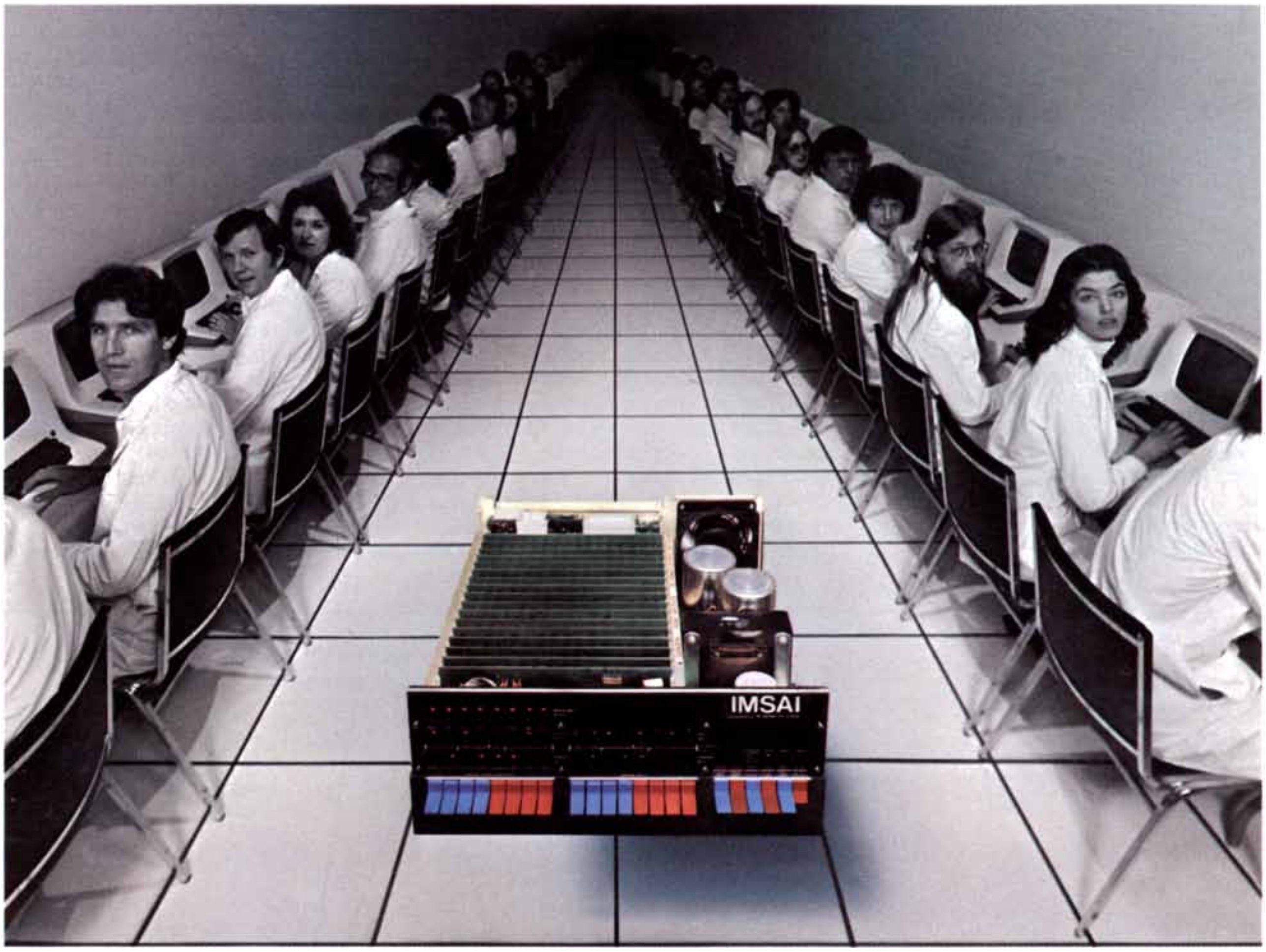
For the purposes of nuclear-magnetic-resonance spectroscopy an isolated sodium cation can be considered a reference state of zero shift, since it has the nonmagnetic electronic configuration of a noble gas. The situation is different, however, for a sodium cation in solution. The outermost electrons of the cation, it will be remembered, are in *p*-type orbitals and have orbital angular momentum in addition to their spin angular momentum. In isolation the separate orbital angular momenta of the six *p*-type electrons exactly cancel, but interactions of the electrons with solvent molecules disturb that balance. As a result of interactions with the solvent the *p*-type orbitals precess around the nucleus, creating a magnetic field that reinforces the external one. In other words, the cation in solution exhibits a paramagnetic chemical shift; the same is true of a cation confined in a cryptand molecule. The extent of the shift depends primarily on the molecules in direct contact with the cation.



NUCLEAR-MAGNETIC-RESONANCE SPECTRA provide convincing evidence for the existence of true metal anions in various solutions of sodium. The concentrated solutions needed to make the measurements were prepared with the help of cryptand molecules. The zero point of the scale is chosen to correspond to the undisturbed magnetic field of the sodium nucleus in the gaseous cation. Apparent deviations from this value in the measured nuclear field are explained by electronic fields that either reinforce (paramagnetism) or oppose (diamagnetism) the nuclear field. The comparatively broad peaks shifted paramagnetically by 51 parts per million are associated with cations confined to the interior of the cryptand. The narrower peaks at the unshifted position are the resonances derived from the sodium anions.

An alkali metal anion has only filled orbitals, and in isolation it should therefore have a nuclear-magnetic-resonance spectrum almost unshifted from that of the isolated cation considered as a reference. What is more, the spectrum should remain unshifted even in solution. The valence electrons in the anion occupy an *s*-type orbital and have no orbital angular momentum. Their inter-

POWER.



IMSAI Introduces the Megabyte Micro.™

The Megabyte Memory

Until today, the largest memory you could fit and address in a single microcomputer CPU was 65K.

Now, IMSAI presents an incredible memory system for micros 16 times more powerful than yesterday's best.

Imagine, a full megabyte of power from sixteen 65K RAM boards.

And, to control all this, the IMSAI Intelligent Memory Manager (IMM), the super control board.

You can write protect blocks throughout the full megabyte. Or, map in 16K blocks.

Plus, preset 16 mapping configurations with protect for high speed transfer or rapid change.

All interrupts are fully vectored, and there's an interrupt if an attempt is made to write into protected memory.

There's even a real "time of day" clock.

65K, 32K and 16K RAM Boards

Until today, the most memory you could plug into a single slot was 16K.

Now, IMSAI presents memory boards in astonishing multiples of sixteen: 65K, 32K and 16K low power, dynamic RAM Boards. They can be used in any S-100 bus computer individually or in combination to form conventional systems up to 65K bytes.

Every board is fast. With "hidden refresh" and *no* "wait state."

The Complete Megabyte Microcomputer System

The IMSAI Megabyte Micro™ is only part of the story. The full system can include dual floppy disks, terminals, plotters, printers and tape cassettes.

IMSAI also offers the finest high level and peripheral software available. Paper tape and Tape Cassette I/O and super Disk Operating Systems. Plus, BASIC and Disk BASIC with more high level languages coming.

Until today, the microcomputer's potential was just something you talked about.

Now, you can put it to work. Powerfully.

GENTLEMEN: SA 7

I'm power hungry!

- ☐ Send 65K RAM Board Kit $2599 ☐ Assembled $3899
- ☐ Send 32K RAM Board Kit $749 ☐ Assembled $1099
- ☐ Send 16K RAM Board Kit $449 ☐ Assembled $679
- ☐ Send IMM ROM Control Kit $299 ☐ Assembled $399
- ☐ Send IMM EROM Control Kit $499 ☐ Assembled $699
- ☐ Send full catalog $1.00

Check/MO enclosed. Amt. $ ____

Charge my: ☐ BAC ☐ M/C

____ Exp. Date: ____

Sig. ____

☐ Send name of my nearest IMSAI dealer

Name ____

Company ____ Title ____

Address ____

City ____

State/Zip ____

IMSAI®

IMSAI Manufacturing Corporation
14860 Wicks Blvd.
San Leandro, CA 94577
(415) 483-2093 TWX 910-366-7287

actions with solvent molecules do not generate a magnetic field.

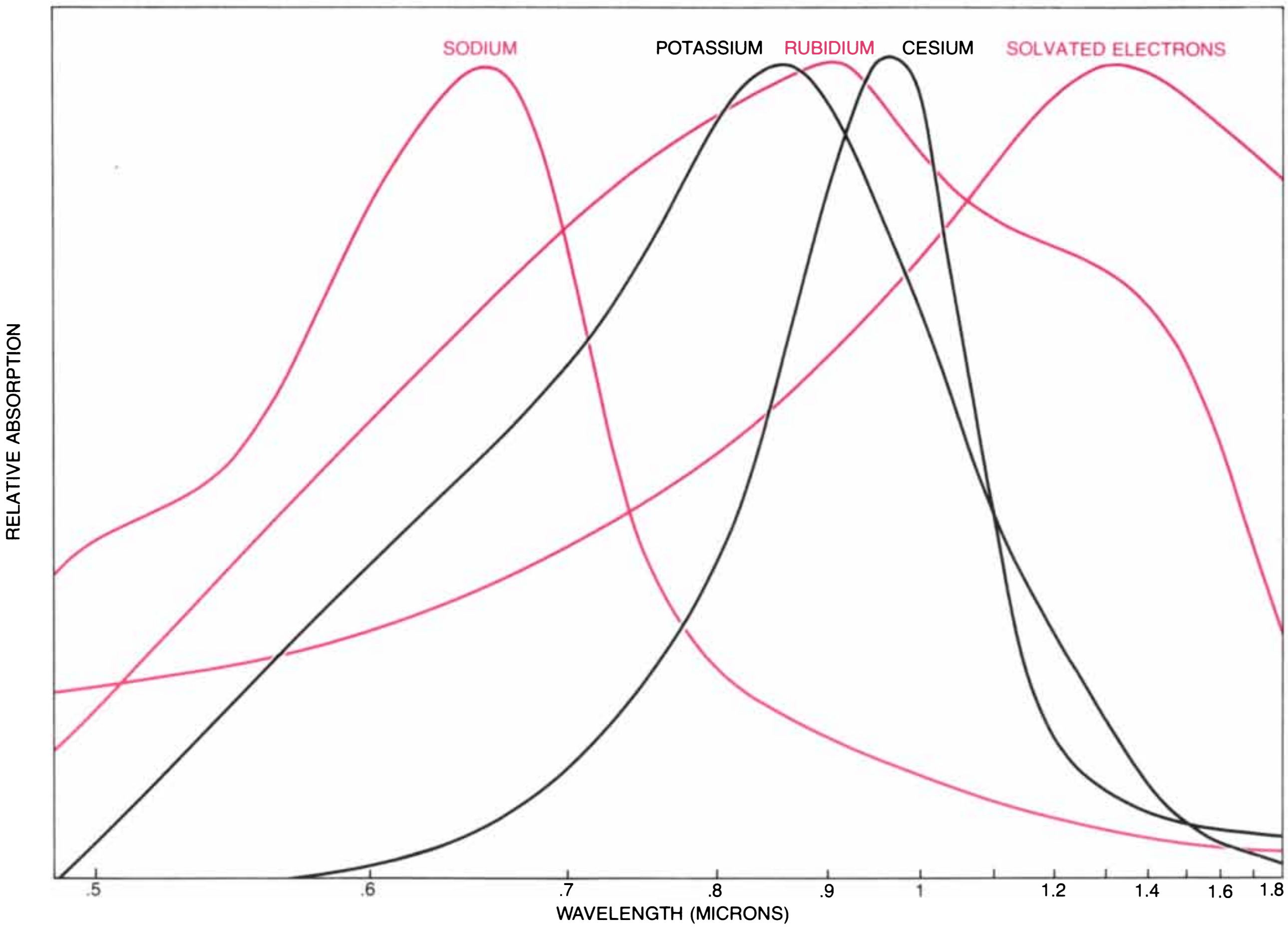
Important preliminary nuclear-magnetic-resonance studies of sodium cations in cryptands were made in 1973 by Joseph M. Ceraso in my laboratory. He showed that an individual sodium cation remains in the cavity of a cryptand molecule long enough for it to be distinguished from other sodium ions elsewhere in a solution. Later, attention was turned to solutions containing both sodium cations sequestered in cryptand molecules and the negative species suspected of being the sodium anion. Two signals were observed. One had the characteristic paramagnetic shift of the Na^+ ion in the environment of the cryptand. The other signal was at a frequency unshifted from the calculated one for the isolated sodium anion. Only in a true anion could the *s*-type orbital effectively shield the nucleus; other structures might have a net negative charge, but they would have the magnetic signature of a cation in solution. Similar results have been obtained with solutions of rubidium and cesium.

If an alkali metal in solution can behave as an anion, it might perform the same role in a salt precipitated from that solution. One can imagine, for example, a hypothetical analogue of a simple salt such as sodium chloride, but with sodium as both positive and negative ion. Perhaps the salt would be called sodium sodide. Calculations indicate that such a salt would be less stable than sodium metal, which would therefore form spontaneously, but the difference between the stability of the salt and that of the metal is only slight. We have prepared the somewhat more complex salts in which the cation is a complex of sodium and a cryptand. The cryptand increases the stability of the salt enough to prevent the formation of the metal.

Mei Tak Lok and Frederick J. Tehan in my laboratory attempted to prepare a salt from an appropriate solution of sodium and a cryptand in ethylamine. Because we suspected that the salt would tend to break down into the metal and the cryptand, we evaporated the solvent rapidly, hoping to dry the material before the cation could escape from the cryptand molecule. The results were encouraging: we obtained a gold-colored film that was in appearance quite unlike either the metal or the cryptand.

In fact, the preparation of the salt turned out to be far easier than we had expected, and the rapid drying was unnecessary. Ceraso, in preparing a sodium-and-cryptand solution for a nuclear-magnetic-resonance investigation, noticed a gold-colored precipitate that formed spontaneously as the solution was cooled to the temperature of dry ice (−78 degrees Celsius). Lok and Tehan later verified that the material has the structure of a crystalline salt with the formula $Na^+2,2,2\text{-cryptand} \cdot Na^-$.

Because of the golden sheen of the crystals we thought at first that the new compound would behave like a metal, a material with abundant mobile electrons. It proved instead to be a semiconductor, in which almost all the valence electrons remain bound, if only very weakly, to particular atoms. In solutions of ethylamine the crystals grow as thin hexagonal plates. Tehan and B. L. Barnett have determined the structure of



ABSORPTION SPECTRA of solid films indicate that both metal anions and solvated electrons can survive the evaporation of the solvent. The films, like the one shown in the upper photograph on page 93, were prepared from solutions containing both dissolved metal and cryptands. The absorption spectrum of the solvated electron, with its characteristic peak in the infrared, can be obtained with films made with any of the metals except sodium. The shorter-wavelength peaks are peculiar to each metal and represent absorption by anions.

the crystal, confirming our assumption that the compound is a salt. Layers of the sodium-cation-and-cryptand complex alternate with layers of the sodium anion. The sodium anions resemble iodide ions in their effective size and in their positions in the crystal lattice.

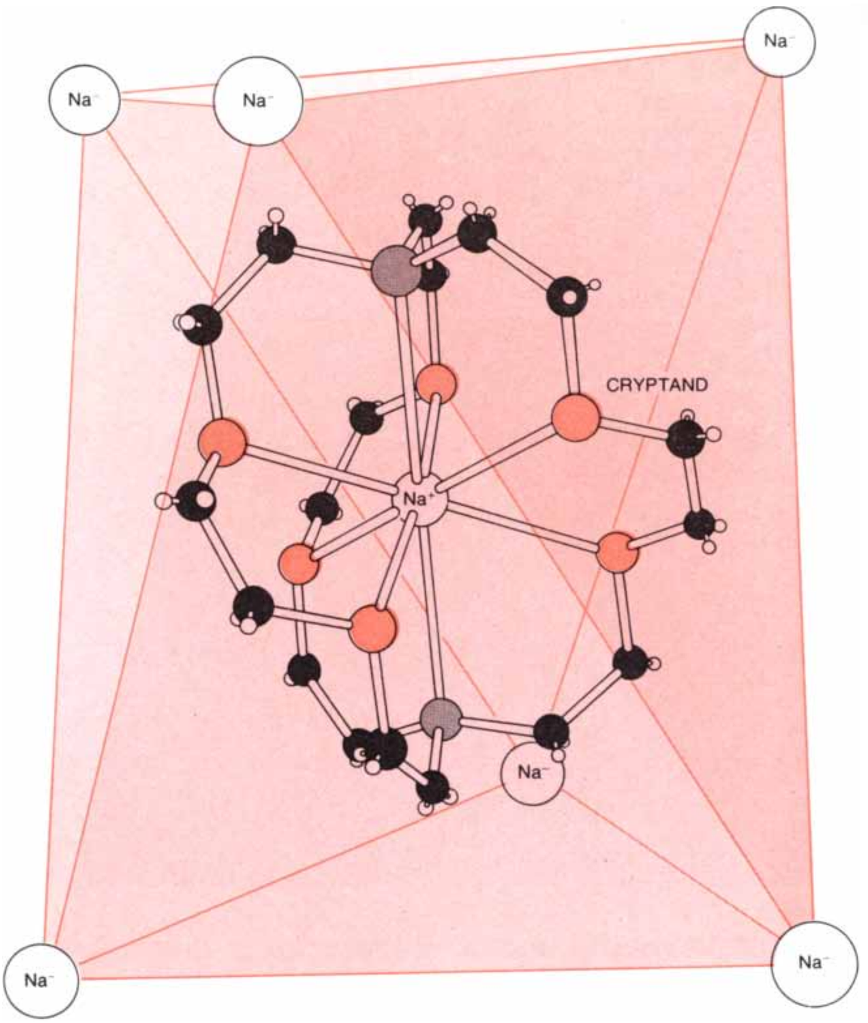
The new solid-sodium compound is highly reactive with respect to air, moisture and in general to any potential acceptor of electrons. Even when it is stored in a vacuum, it is not indefinitely stable except at low temperature. At room temperature the crystals slowly darken and lose their luster as the sodium anions attack the organic cryptand molecules. Nevertheless, by working quickly we were able to measure the melting point of the crystals. It is 83 degrees C., 15 degrees above the melting point of the pure cryptand molecules. Melting is accompanied by decomposition into sodium metal and the free cryptand.

Recently the striking metallic appearance of this nonmetallic compound has been explained by Michael G. DaGue and Michael R. Yemen and me. The luster of metals is the subjective impression given by an optical absorption spectrum characterized by steeply rising absorption as one goes from shorter to longer wavelengths. In metals that absorption profile results from absorption by free electrons distributed throughout the metallic lattice. The sodium-anion crystals happen to have a similar absorption spectrum even though they do not have abundant free electrons. Thus the metallic appearance derives not from metallic structure but from accidental mimicry. Also contributing to the illusion is the crystal structure of the sodium salt, which has large crystalline faces like those of a metal.

Since the original identification of the sodium salt, we have prepared compounds containing potassium, rubidium and cesium that appear to have an analogous structure. The definite identification of these compounds has been hindered by their comparative instability. At room temperature the sodium salt has a half-life for decomposition of several days; the corresponding potassium salt has a half-life of less than a minute. The salts of the heavier metals must therefore be maintained continuously at temperatures below about −40 degrees C., where all of them seem to be stable for long periods.

Since solutions of alkali anions and trapped cations yield solid compounds, it seems plausible that related crystalline materials might be created from trapped cations and solvated electrons. One possibility is an "expanded metal," in which the free electrons wander not among bare metal cations but through a lattice of cation-and-cryptand complexes. Efforts to prepare solids with this crystalline structure have failed so far. By rapidly evaporating the solvent, however, we have observed residual films that seem to have a different structure. The films are deep blue and strongly paramagnetic, suggesting that the electrons are confined to the vicinity of particular cations rather than wandering freely through the lattice. Preliminary measurements suggest that the solids contain exactly one unpaired electron per cation; if that is confirmed, then the films are not metallic at all but are the first example of a new class of salts: the "electride" salts.

It must now be considered established that anions exist for all the alkali metals except lithium. (Calculations indicate that the lithium anion should also be stable under appropriate circumstances, and so we continue to search for that species.) In all probability we have discovered only a small fraction of the compounds that include alkali metal anions. Because the size of the cryptand can readily be altered a variety of cations might be combined with any given anion. One obvious possibility is the preparation of heterogeneous salts containing a trapped cation of one alkali metal and an anion of another. More generally, the cryptands could also confine cations other than those of the alkali metals. For example, salts might be formed with alkaline earths, such as barium, which have a charge of +2, or with rare earths, such as lanthanum, which have a charge of +3. The salts would necessarily contain a high concentration of alkali metal anions: two or three per cation. Ultimately the cryptands might be dispensed with entirely. Certain organic cations, including for example organic derivatives of the ammonium ion, might be able to survive unprotected in the presence of alkali metal anions or even in the presence of electrons.



STRUCTURE OF CRYSTALS precipitated from solutions of sodium can be resolved into alternating layers of sodium anions and the complexes formed by a cryptand and a sodium cation. In some fundamental characteristics the crystals are much like more conventional salts, such as sodium chloride. For example, there must be equal numbers of anions and cation-and-cryptand complexes. What distinguishes these crystals is the reactivity of the metal anions.

New instrument measures distance to 3 miles, horizontal and zenith angles to 1 second.

Designed primarily to increase the speed and accuracy of a surveyor's measurements, HP's new Electronic Total Station incorporates advances that scientists and engineers in other fields will find especially useful in monitoring the angle and distance of fixed or moving objects.

For a surveyor, the impact of the new HP 3820 Electronic Total Station is clear. With a single compact instrument he can now make all the measurements he needs, more easily and more accurately than ever before:

- Measure any horizontal, vertical, or slope distance up to 5 kilometers, automatically corrected for environmental conditions, and display it in feet or meters.
- Measure zenith or horizontal angle electronically to a resolution of 1 second, automatically corrected for instrument mislevel, and display it in degrees/minutes/seconds or grads.
- Measure relative direction (the clockwise angle from the previous direction) and display it directly, thus avoiding the need to subtract degrees/minutes/seconds.

For other scientists and engineers, two other characteristics may be even more significant.

The tracking mode of the HP 2830 continuously updates any of its measured or computed displays. This capability is particularly useful for making distance and angle measurements to monitor a moving object—for example, tracking a boat while dredging, or positioning overhead cranes in a materials handling operation.

At the touch of a button, the HP 3820 outputs any of its measurements digitally to an external data collector or desk-top computer. Obviously a convenient way to record data, this feature can also be used for control. Monitoring an industrial steel beam cutting operation is one example. The HP 3820 tracks the moving beam until it reaches the desired length and outputs the measurement to a desk-top computer, which then transmits a control signal to the positioning and cutting machines.

In all its uses, the HP 3820 is a remarkable time saver when compared to traditional angle and distance measuring instruments. Automatic level compensation sharply reduces set-up time and the built-in microprocessor instantly makes all necessary computations to display the measurement in the form and the units desired, at the touch of a button.

The HP 3820 costs $17,750* including rechargeable battery pod, leveling base, and carrying case.

The HP-67

The HP-97

Almost all calculators are created =. HP is the powerful exception. ENTER↑.

As with an automobile, whose real power is *not* the engine's rated horsepower but the energy actually delivered to the wheels, so with a programmable calculator: it's not the number of available program steps that determines its power, but the actual problem-solving efficiency resident in each step.

Unlike calculators with an = key, Hewlett-Packard's programmables employ a special logic system—ENTER↑—that lets you enter complex number problems in a natural, parenthesis-free way. ENTER↑ also saves steps, particularly in longer programs. Add to this the step consolidation that fully merged keystrokes deliver, and you have *unequalled* power per program step *and the hardware part of what it takes to get solutions*.

The more important part of what it takes to get solutions is software. Programs and application pacs are the fuel on which our programmables run. And there is no shortage with HP. We offer a full program library, plus user-contributed programs for an immense variety of tasks. This working combination of calculators and comprehensive programs delivers accurate solutions painlessly.

Cases in point are the HP-67, the most powerful fully programmable hand-held we've ever built, and its briefcase-portable "twin" with a built-in thermal printer, the HP-97. These calculators are exceptionally easy to program, edit, and operate, because they use HP's special logic system, ENTER↑. They handle programs of up to 224 *powerful* steps on a single magnetic card, while larger programs—even thousands of steps—are easily handled with several program cards.

What's more, all multiple keystroke operations are fully merged into single program steps. This means that when you push the prefix keys that assign secondary functions to the function keys, it doesn't cost you an extra program step. Conservation. And such operations as adding a number to memory, requiring several program steps on some calculators, are squeezed into *one* program step. More conservation.

What does this mean in terms of programming power? Greater compression; fewer keystrokes; more competent programs; less chance of error; answers you can trust. In short, the most direct solution.

If, despite all these powerful hardware/software assists, a problem requires more computing power than the HP-67/97 has, perhaps a more productive solution resides in a computer. And because we make them too, we understand the reasonable limits of both categories.

When you use an HP programmable, you find that the ENTER↑ in HP's calculator logic, plus merged codes, are the conservators of time and cerebral energy. Ask any Professional.

HEWLETT hp PACKARD

1502 Page Mill Road, Palo Alto, California 94304

For assistance call: Washington (301) 948-6370, Chicago (312) 255-9800, Atlanta (404) 955-1500, Los Angeles (213) 877-1282.

Mail to: Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304.
Please send me further information on

( ) HP 3820 Electronic Total Station
( ) HP-67/97 programmable calculators

Name ____________________
Company ____________________
Address ____________________
City ____________ State ________ Zip ________

*Domestic U.S. prices only.

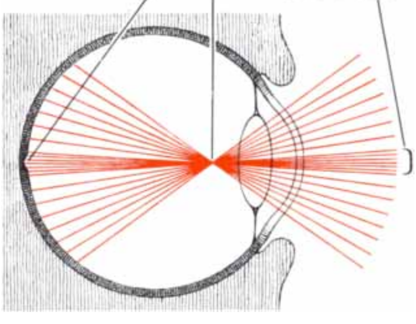
00746

The Compound Eye of Insects

It has hundreds of facets, each representing an ommatidium, or "little eye." The size, pattern and aiming of the ommatidia are determined by the requirements of the insect's way of life

by G. Adrian Horridge

Insects are legion; some are active only in bright sunlight, some in both sunlight and shade, some only at twilight. Some, such as the common housefly, travel at high speed and make rapid turns. Others, such as the dragonfly, alternately fly straight, maneuver and hover. Still others, such as the praying mantis, remain motionless for hours at a time. All these insects rely for their survival on vision, and all of them perceive the world through many-faceted compound eyes, yet their habits and their visual requirements are quite different. How does the compound eye work as an optical sampling device? To what extent does it reveal the functions for which the insect uses its eyes? To what extent does the smallness of the facets of the insect eye limit its sensitivity? What is the barrier to the insect eye's working at low light intensities with the small lenses of the facets? How do the compound eyes of insects arrive at a compromise between optical resolution and the sensitivity needed to overcome optical noise?



EYE OF A VERTEBRATE ANIMAL has a small cup-shaped area in the retina where vision is particularly acute. In that area, the fovea centralis, the visual receptors are closely packed, so that their resolution, or ability to distinguish fine detail, is high. Each line in color represents the optical axis of every *n*th receptor. The close packing of the receptors in the fovea requires that the receptors be narrow, so that the sensitivity of each foveal receptor is lower than that of the receptors elsewhere in the retina. Hence the fovea in vertebrate animals is a specialized area for seeing with high resolution only in bright light.

A compound eye is made up of ommatidia: tiny individual eyes that point in different directions. Each ommatidium consists of a lens that focuses light on several receptor cells sharing the common light-sensitive organ known as a rhabdom. In many insects each ommatidium is a single sampling element with its own optical axis; only insect eyes with this type of ommatidium will be discussed here. Light absorbed by the rhabdom transmits a signal to the insect's brain through an optic nerve. Another question is: How can we determine the directions in which the ommatidia are looking? And what does that tell us about how the insect perceives the world?

Looking at the eye of an insect, we frequently see a black spot in the center of the eye. As the insect rotates its head the black spot always points in the direction of the observer. The spot is known as the pseudopupil; the facets in it look black because they reflect less light in the direction of the observer than the facets in the rest of the eye do. In other words, the ommatidia in the center of the pseudopupil are looking directly at the observer.

To map the optical axis of each ommatidium the head of the insect (or the entire animal) is mounted at the center of a goniometer, a device for measuring angles. The angle required to move the center of the pseudopupil a distance of five facets is measured, and the measurement is repeated at intervals of five facets until the entire eye is mapped. The same data can be more easily collected by photographing the eye at intervals of five or 10 degrees. In that case the facets are marked with dust particles, and the positions of the pseudopupil with respect to the pattern of the marked facets are measured from the photographs. The position of the axis of each fifth facet is then plotted on a grid in angular coordinates. The resulting pattern is a map of the way in which the compound eye samples the visual world. So far only a few partial maps have been made, such as one prepared for the housefly by the Dutch biophysicists D. G. M. Beersma, Doekele G. Stavenga and Jan W. Kuiper. Only narrow strips can be mapped without encountering the familiar problems that are encountered in mapping a sphere on a flat piece of paper.

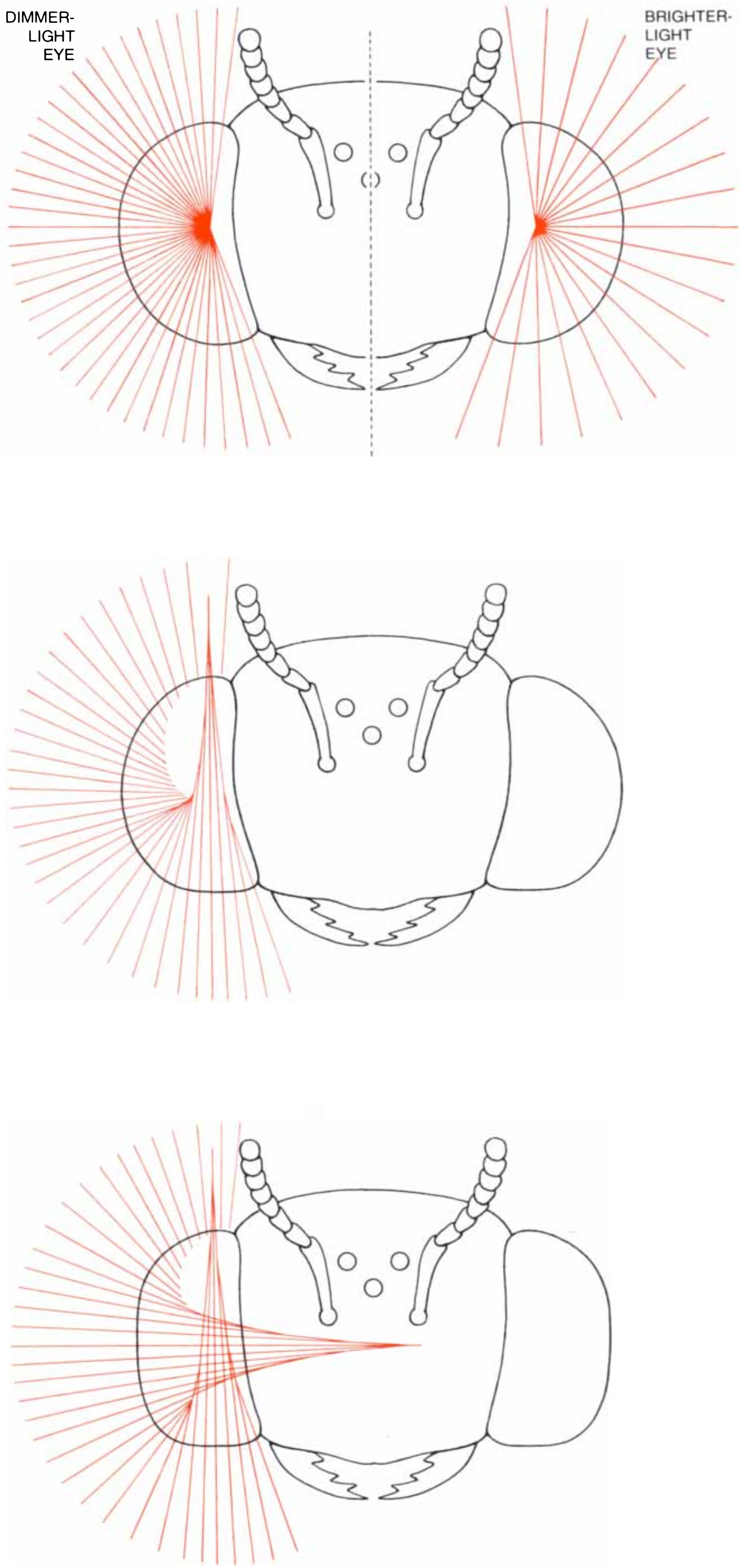
Two angles are of prime importance in understanding the function of an insect eye. The first is the angle between the optical axes of two adjacent ommatidia. This is the interommatidial angle, denoted $\Delta\phi$, which determines how densely the compound eye samples the visual world. The second is the angle defining the field of view of a single ommatidium, denoted $\Delta\rho$. The shape of the field is in effect the sensitivity of the ommatidium to a point source of light as a function of angle from the optical axis. The sensitivity of the ommatidium is at its greatest along the optical axis, and it decreases with angle from the axis. The width of the field is defined as the angle subtended by the diameter across the field where the sensitivity has dropped to 50 percent.

The eye maps reveal some surprises about the way in which insects perceive the world. First, the maps show that the pattern of the optical axes is not the same as the pattern of the hexagonal facets on the surface of the eye. Facets that are an equal distance apart on the surface of the eye do not necessarily have optical axes that are at the same interommatidial angle. In fact, in one region of the eye the optical axes of the ommatidia may be tilted toward one another so that they are nearly parallel, thus giving one segment of the visual world more than its share of the distribution of sampling points. In that region the pseudopupil is usually larger than it is elsewhere on the eye, because more facets are looking in that direction.

Such an area on the compound eye of insects has the same function as the fovea centralis in the retina of vertebrate animals (including man). In vertebrates the fovea is an area where the visual receptors are packed closer together, providing the most acute vision. Indeed, in the eyes of diurnal animals (including man) the receptors must be very narrow to be so closely packed, and the sensitivity of each receptor is correspondingly reduced. Hence in vertebrates the fovea is an adaptation for acute vision in bright light.

In insects, however, the facets in the fovea are usually larger than those over the rest of the eye. Since the interommatidial angle in the fovea is reduced, the region of the eye that includes the fovea is optically flatter than the rest of the eye; in other words, it has a larger radius of curvature than the rest of the eye. Thus for there to be room for a fovea in one region of a compound eye the eye must have a smaller optical radius of curvature elsewhere. The result is that some areas of the eye are impoverished in sampling points so that another area can be enriched.

Larger facets make it possible for their ommatidia to have a higher resolving power. For any lens the capability to resolve fine detail depends on two factors: the aperture or diameter of the lens (D), and the wavelength (λ) of the light focused through it. (Because of the difficulty in measuring the aperture of a hexagonal lens, D is taken as the distance between centers of adjacent facets.) Light from a distant point source passing through the lens is not focused to a point but is blurred into a circle in the focal plane of the lens. The intensity of the light in this blur circle is greatest in



COMPOUND EYE OF AN INSECT is built up of ommatidia. The effective size of any part of such an eye is established by its optical radius, which is defined as the radius measured from the point where the optical axes of adjacent ommatidia intersect. Here four types of compound eye are seen from above the back of the insect. If the insect eye were a hemisphere and ommatidia of equal aperture looked out perpendicularly to the surface of the eye, the entire eye would have one optical radius, and it would sample the visual world with equal density in all directions (*top*). The sensitivity of a compound eye depends partly on the size of its facets and therefore on its overall size for a given number of facets. Measurements of actual compound eyes indicate, however, that in many of them the optical radius at the front of the eye is greater than it is elsewhere on the eye (*middle*). Therefore the part of the visual world seen by the front of the eye is sampled more densely than other parts are. Such an area is called a fovea by analogy with the fovea of the vertebrate eye. At the fovea of most insect eyes the facets are larger, which is made possible by the eye's having a greater optical radius. Some insects (*bottom*) have a second fovea that looks upward or to the side; large dragonflies have foveas of all three kinds.

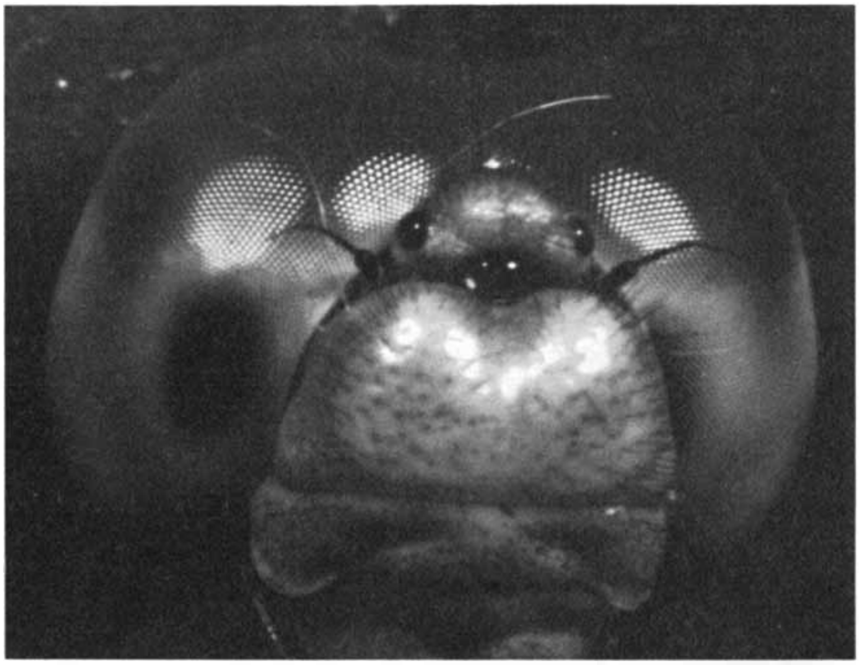
the center and falls off toward the edge. It is convenient to define the width of the blur circle as the diameter where the intensity of the light has fallen to 50 percent. That width is denoted $\Delta\alpha$. From diffraction theory it is known that the angular width of the blur circle in radians is equal to the wavelength of the light divided by the diameter of the lens ($\Delta\alpha = \lambda/D$). (One radian is a full circle divided by π, or approximately 57 degrees.) Hence for any particular wavelength of light, the larger the lens, the smaller the blur circle for each point source and the sharper the image.

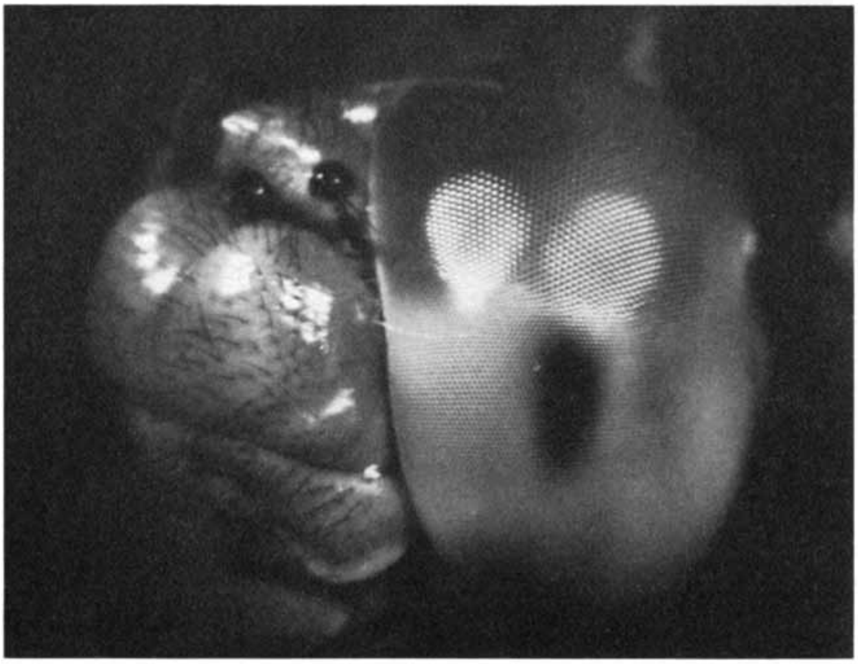
There is much evidence that in insects relying on acute vision the optical performance of the lens of each ommatidium is excellent, and that the receptor structure catching the light (which is in effect the tip of the rhabdom) lies in the focal plane of the lens. There are two fundamental limits on the resolution of the individual ommatidium. The first is the width of the blur circle: a large facet on an insect's eye can have a higher resolving power than a small one because it concentrates the light into a smaller blur circle. The second limit has to do with how narrow the receptor can be and still act as a light guide; the minimum diameter of the receptor is about .5 micrometer.

If a point source of light lies on the optical axis of the ommatidium, the most intense, or brightest, portion of the blur circle will be centered on the tip of the rhabdom. As the point source travels across the field of view, it moves away from the optical axis and the peak of the blur circle moves away from the rhabdom. Thus the rhabdom will intercept a progressively smaller and less intense segment of the blur circle. If the rhabdom is relatively narrow, it can precisely measure the variation in the intensity of the light across the blur circle, and its resolution is limited only by the diffraction of the light that gave rise to the blur circle.

By the same token, however, a narrow rhabdom will capture only a small fraction of the total light in the blur circle and its sensitivity will be low. If the rhabdom is wider, it will capture more of the light in the blur circle, but its resolution will be poorer because it will not be able to measure detail as well. Hence the optimum angular sensitivity of the rhabdom is a function both of the width $\Delta\alpha$ of the blur circle and the diameter d of the rhabdom. In every eye there is a compromise between resolution, which is limited only by the diffraction or the minimum rhabdom width, and sensitivity, which, as we shall see, is very important in eyes with such small lenses.

This compromise can be illustrated by an analogy: the way in which the profile of a spray of water from a shower head can be measured by containers of different sizes. The diameter of the spray is equivalent to the width of the blur circle, and the water drops are equivalent to the photons of light. A small bowl moved across the spray will readily measure, or resolve, the intensity profile of the spray, or will readily distinguish the boundaries of several sprays that are close together. Yet at any point the bowl will intercept only a small portion of the spray, that is, its sensitivity is low. The spray is caught much more effectively by a large tub. The tub is thus more "sensitive" than the bowl, yet by catching so much of the water it cannot distinguish detail or discriminate between sprays that are close together. When the flow is low, however, only large sprays can be detected, and the tub provides the best means for measuring the distribution of the spray.



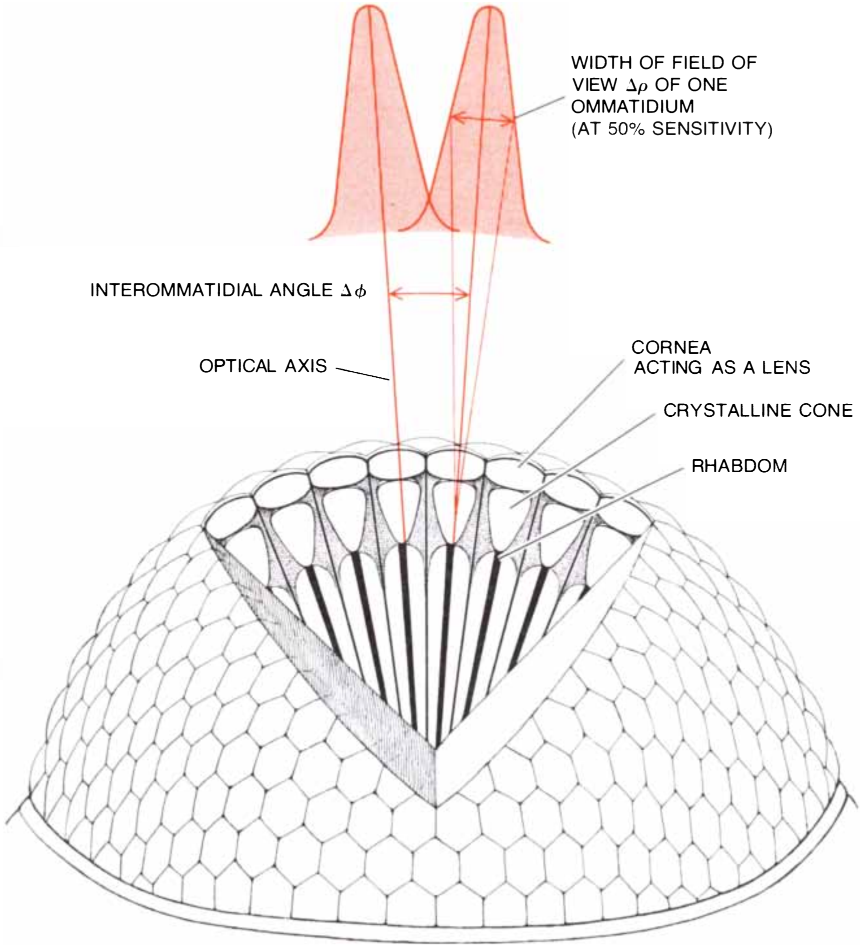


PSEUDOPUPIL OF INSECT EYE is seen as a black spot on the eye of the libellulid dragonfly *Diplacodes nematodes*. As the insect's head turns, the pseudopupil moves so that it always points toward the observer. The pseudopupil is black because the facets in that area reflect less light in the direction of the observer than the facets in other areas of the eye. The facets in the center of the pseudopupil are looking directly at the observer. In general, where the pseudopupil is larger (*top*) more ommatidia are looking in that direction. At the side of the eye (*bottom*) the pseudopupil is elongated vertically because the optical axes in vertical rows have between them about half the angle found between the visual axes of the ommatidia in horizontal rows.

If the diameter d of a circular rhabdom is narrow compared with the width of the blur circle $\Delta\alpha$, the rhabdom's field of view $\Delta\rho$ is equal to the angular width of the blur circle. If the rhabdom is wide enough to catch 95 percent of all the light in the blur circle, the rhabdom's field of view $\Delta\rho$ turns out to be twice the angular width of the blur circle at 50 percent intensity, or $2\Delta\alpha$. Increasing the diameter of the rhabdom beyond that value results in a continuing loss in resolution without a corresponding increase in sensitivity to a point source of light. In the shower analogy, having a tub larger than the diameter of the spray will not increase the amount of water the tub can catch from one spray. Making the rhabdom "tub" larger, however, is obviously a better stratagem for seeing at least something when photons are so scarce that only large sprays of them can be detected.

So far in this discussion of the origin of the insect's field of view from the blur circle it has been necessary to avoid problems presented by the refraction, or bending, of light as it passes from the air into the insect's eye. Angles such as the width of the blur circle $\Delta\alpha$ (inside the eye) and the field of view (outside the eye) are measured from the posterior nodal point, which in a one-sided lens such as this one is a point very close to the center of curvature of the lens. Rays of light passing through the posterior nodal point inside the eye are parallel to the same rays outside the eye; for that reason the posterior nodal point is a useful one for relating angles measured outside the eye to angles measured inside the eye. The diameter of a large receptor subtended at the posterior nodal point is the size of the field of view $\Delta\rho$.

Let us suppose that in a typical ommatidium of high resolving power the lens is 28.5 micrometers in aperture. If the wavelength is .5 micrometer (the wavelength of green light), the width of the blur circle $\Delta\alpha$ is then one degree. If in such an ommatidium the diameter of the rhabdom (subtended at the posterior nodal point) were equal to the width of the blur circle at 50 percent intensity $\Delta\alpha$, then from the formula $\Delta\rho^2 = \Delta\alpha^2 + (d/f)^2$ its field of view would be 1.4 degrees. This rhabdom could have a diameter of one micrometer and a focal length f (measured from the posterior nodal point to the tip of the rhabdom) of 57 micrometers. A rhabdom with a diameter of less than a micrometer is not very effective as a light guide, and such narrow rhabdoms are rarely found in insect eyes. A larger aperture of facet would generate a narrower blur circle, calling for a narrower rhabdom. Therefore if the performance of the ommatidium is limited by the diffraction of light, one would not expect to find insects with ommatidia that have facets more than about 30 micrometers across, because effective rhabdoms cannot be made thin enough to match the facets.

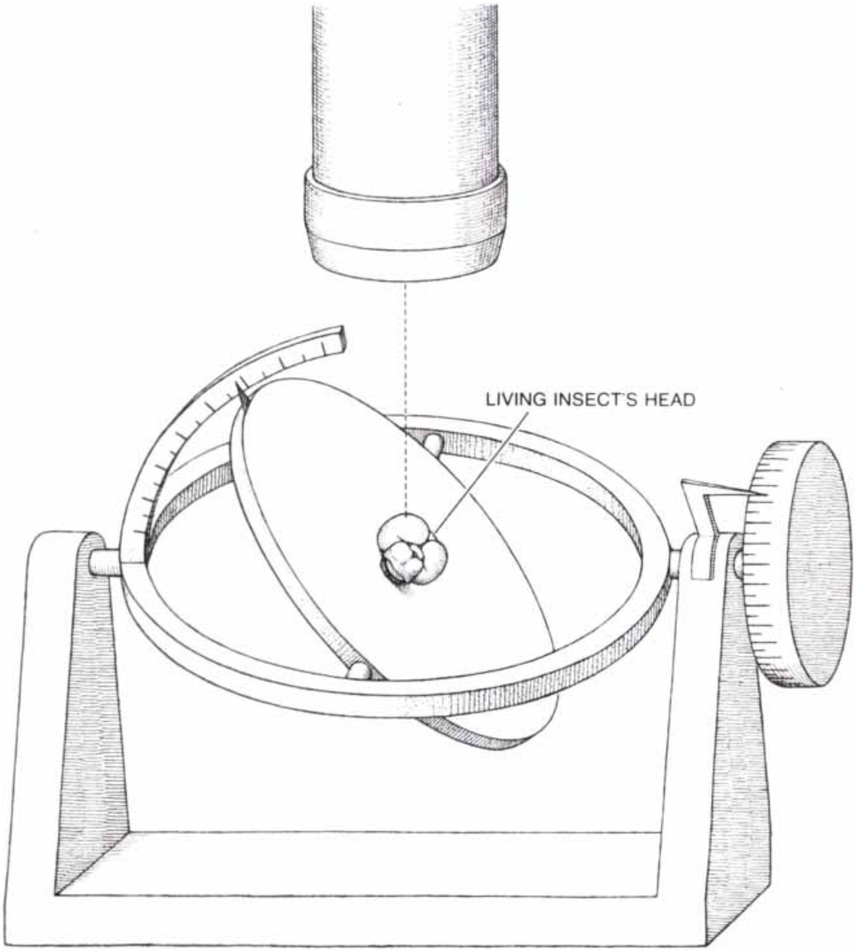


STRUCTURE OF INSECT EYE determines how much detail the eye is capable of seeing. Each ommatidium has a cornea that also serves as a lens focusing light through a transparent cone onto a light-sensitive element: the rhabdom. The optical axis of the ommatidium is the line extending through the center of the lens to the rhabdom. The interommatidial angle $\Delta\phi$ is the angle between the optical axes of the adjacent ommatidia. The curves in color above each ommatidium indicate how the sensitivity of the ommatidium decreases with angular distance from the ommatidium's optical axis. The field of view of each of the ommatidia is defined as the angle subtended where the sensitivity has fallen to 50 percent of its maximum value.

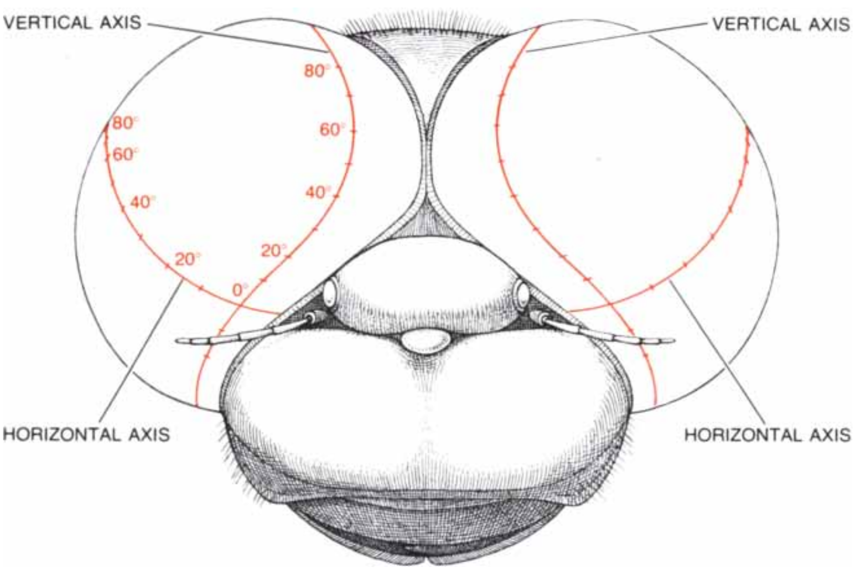
The fact is, however, that facets up to 80 micrometers in diameter are commonly found, particularly in the fovea. Some organisms other than insects, such as the stomatopod mantid shrimp *Odontodactylus* and the deep-sea amphipod *Phronima*, have a compound eye with facets up to 120 micrometers across. All compound eyes have fewer and larger facets than the diffraction theory predicts. Why should such large apertures be found? The answer is that further increasing the diameter of the lens increases the ommatidium's sensitivity in proportion to the area of the lens (D^2). Consideration of point sources and the diffraction limit can take one no further in understanding the compound eye. The individual ommatidia seem to be adapted in many ways to catching every last photon of light. In the shower analogy one must turn from the process of catching a single small shower to distinguishing low-flow showers of all sizes in different combinations. When the flow of drops (photons) is so low that only large showers can be detected, sensitivity is important.

So far I have concentrated on optical constraints and on the characteristics of a single ommatidium. A compound eye, however, is an array of hundreds of ommatidia looking out more or less equally in many directions. The advantage of such an arrangement is that it operates by parallel processing: it divides information from the external world into many parallel channels, analyzes the information in each channel and recombines some of the information in the brain. Thus an insect can be aware of its entire visual world at every moment, whereas vertebrate animals such as ourselves can gain that information only by continually moving their eyes and head in a series of scans.

What governs the spacing of the ommatidia over the surface of the eye? In 1894 A. Mallock (and later Horace B. Barlow, Hugo de Vries, Jan Kuiper and



INSECT EYE IS MAPPED by mounting the insect's head on the stage of a goniometer (calibrated gimbals) and photographing the eye over a range of angles. The angle required to move the center of the pseudopupil five facets determines the interommatidial angle between every fifth ommatidium. Position of optical axes of each fifth ommatidium is then plotted on a map.



COORDINATE SYSTEM FOR MAPPING INSECT EYE is arbitrary, but the principal functional axes of the eye are used. The zero of the vertical axis usually crosses the zero of the horizontal axis at the position of the insect's forward-looking fovea. The zero vertical axis on the maps in angular coordinates is the line of facets that look in the plane in which the forward and vertical directions lie. It is not a straight line down the eye. The zero horizontal axis on the maps is the line joining the facets that look out horizontally. Again it is not necessarily a horizontal row of facets running across the eye because optical axes can be inclined to eye surface.

many others) proposed that the angular spacing of the ommatidia might be calculated from the limits of diffraction, in that the angular spacing of adjacent lenses ($\Delta\phi$) need be no better than twice the spacing ($\Delta\theta$) of the finest pattern each can resolve. The argument was basically the following:

The world the insect perceives can be regarded as a complex pattern of contrasts between light and dark. Such a pattern can be approximated as the sum of a number of regular patterns of stripes, each with a sinusoidal distribution of intensity between light and dark and a regular period $\Delta\theta$. For an array of receptors to see a striped pattern and reconstruct it as stripes by simultaneous parallel processing, the minimum angular spacing between receptors (that is, the interommatidial angle $\Delta\phi$) must be equal to the angle subtended by half the spacing of the striped pattern (that is, $\frac{1}{2}\Delta\theta$). With that minimum spacing two peaks (bright stripes) in the pattern of the stripes' intensity, together with the trough between them (a dark stripe), fall on three adjacent receptors. If the interommatidial angle were smaller, there would be more than enough receptors to discern the pattern; if the interommatidial angle were larger, there would be too few receptors to discern it. This idea is known as the Whittaker-Shannon sampling theorem, and in two dimensions it applies to receptors arrayed in a square pattern. As has been shown recently by Allan Snyder of the Australian National University and William H. Miller of the Yale University School of Medicine, for a hexagonal pattern of receptors to attain the same visual acuity as a square pattern, the hexagonal pattern needs more receptors by a factor of 2 divided by the square root of 3. The eye maps of insects reveal that the pattern of the optical axes of the ommatidia lies between the square and the hexagonal pattern. The hexagonal pattern is warped into a pattern of oblique squares, particularly around the equator of the eye.

From diffraction theory it is also known that the spacing of the stripes $\Delta\theta$ in the finest pattern that can be resolved at any wavelength of light through a lens of aperture D is equal to the wavelength of light λ divided by the aperture of the lens ($\Delta\theta = \lambda/D$). Since the minimum interommatidial angle $\Delta\phi$ must be equal to half the spacing of the stripes, then in order for a diffraction-limited eye to see, the product of the aperture of the lens and the interommatidial angle must be equal to half the wavelength of light (that is, $D\Delta\phi = \lambda/2$). The early workers thought that this was the case.

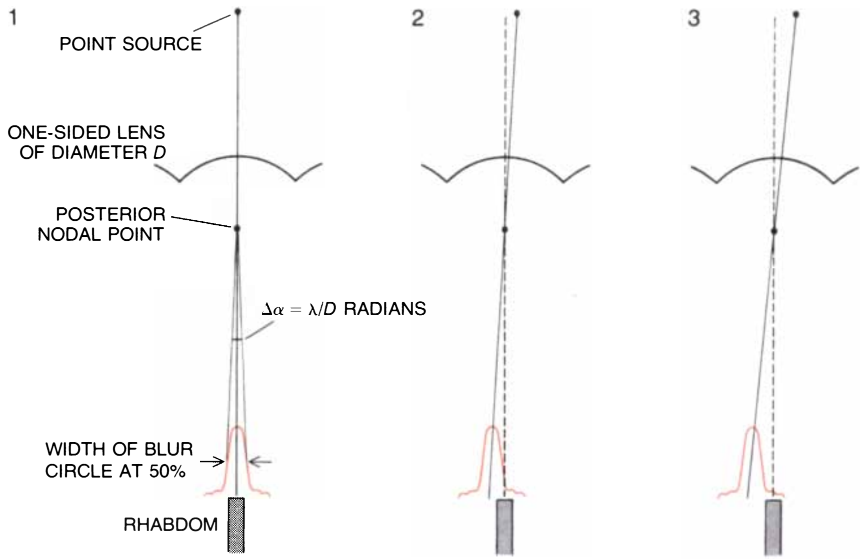
If we say, however, that the eye must operate at a higher level of light modulation to overcome photon noise, then we can say that $D\Delta\phi$ is equal to λ, so that $D\Delta\phi$ is now equal to .5 micrometer. The product $D\Delta\phi$ should therefore be constant for an eye that functions at a given modulation level, and it should be .25 micrometer for square patterns of axes working at the diffraction limit. Both D and $\Delta\phi$ can be measured from outside the eye. On the basis of simple geometry the interommatidial angle in radians is the facet diameter D divided by the optical radius R. Therefore D^2/R should be constant for eyes of different sizes that work at the same intensity. In other words, simply by knowing the aperture of the ommatidium's lens D and the interommatidial angle $\Delta\phi$ one can determine a great deal about the structure of the eye from the product of the two. That applies both for different insects and for different regions on the eye of a single insect.

If the structure of the compound eye of an insect were designed to reconstruct detail at the diffraction limit at a wavelength of .5 micrometer, then the eye parameter $D\Delta\phi$ would be fixed at a value of between .25 and .31, depending on how the pattern of the optical axes is distributed between squares and hexagons. The eye maps reveal, however, that there are no insects with compound eyes the structure of which is determined by these limits. Evidently sunlight is never bright enough to eliminate photon noise as a factor. Only the foveas of some insects that are active in bright sunlight approach the diffraction limit.

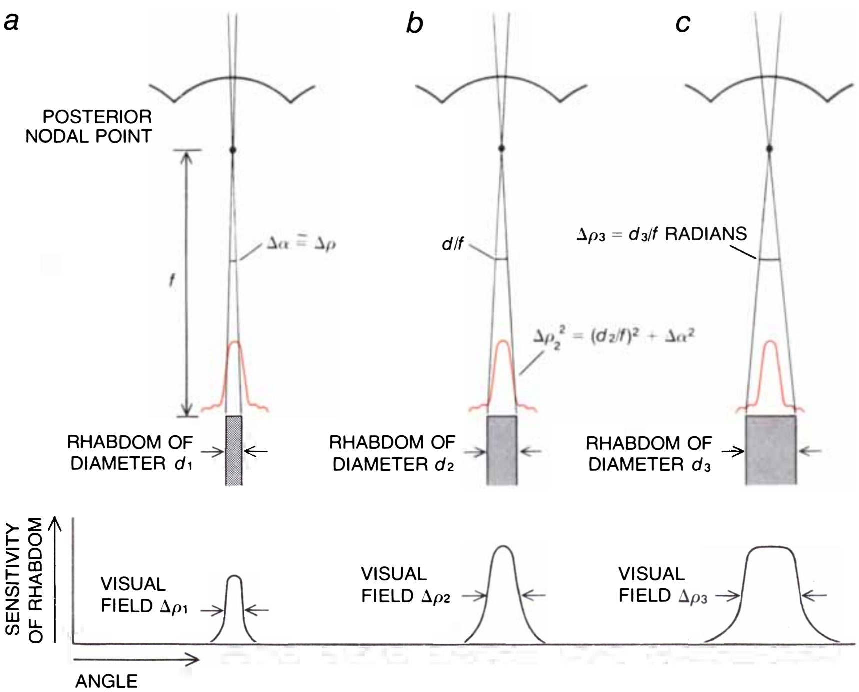
When a pattern of stripes with a spacing of $\Delta\theta$ passes across an insect's field of view, the rhabdom perceives the pattern as a modulation, or oscillation, laid over a mean intensity of light. If the spacing of the striped pattern $\Delta\theta$ is large in relation to the field width $\Delta\rho$, then the amplitude of the modulation from light to dark will also be large. If the spacing of the striped pattern is small, then the amplitude of the modulation will be small. The speed of movement is not important until levels are reached where the receptor fails to respond in time.

The mean light intensity depends on the ambient illumination of the stripes and the width of the field. The amplitude of the modulation M experienced by the receptor is in fact the amount by which the intensity of the light fluctuates from the mean intensity. Specifically, the relative modulation experienced by the receptor is equal to the difference of the maximum intensity and the minimum intensity divided by the sum of the maximum intensity and the minimum intensity. Thus a spatial modulation m in the contrast between light and dark in the visual world is converted by the motion and the receptor into a temporal modulation M, the relative amplitude of which depends on the width of the field of view, the spacing of the stripes and (at significant speeds) the angular velocity of the relative movement [*see illustration on page 115*].

A receptor with a broad field of view

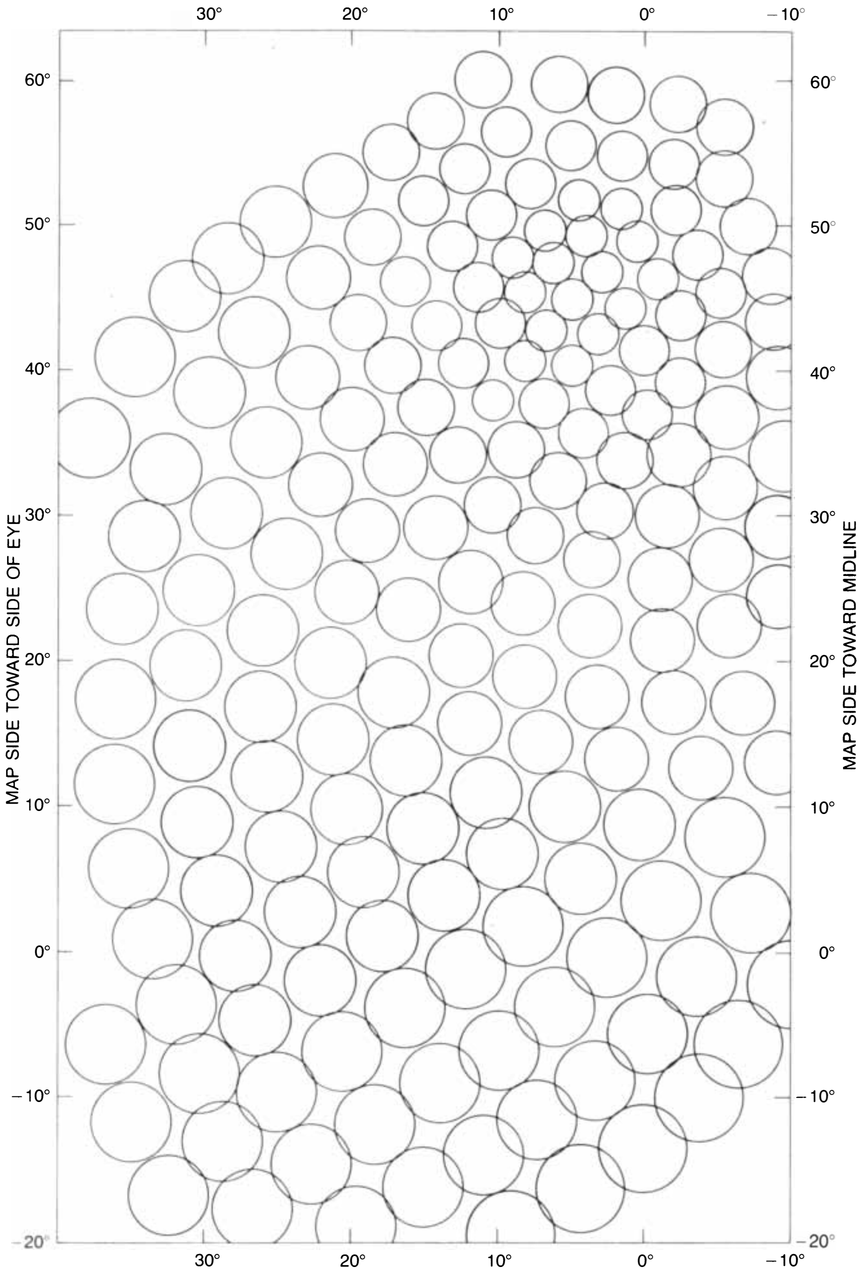


LIGHT IS DIFFRACTED as it passes through the lens of an ommatidium. As a result light from a point source is not focused to a point but forms a circular pattern, the blur circle. The intensity of the blur circle (*curves in color*) is greater in the center and falls off toward the edges. The width of the blur circle ($\Delta\alpha$) is defined as its diameter where the intensity has fallen to 50 percent. That angular width (in radians) is equal to the wavelength of light from a point source divided by the diameter D of the lens. When blur circle is on optical axis (*1*), rhabdom receives greatest amount of light. As point source moves away from optical axis (*2 and 3*), blur circle moves away from tip of the rhabdom and amount of light rhabdom receives decreases. The rays are drawn through the posterior nodal point, which is near center of curvature of lens. Rays passing through this point have a path inside eye that is parallel to their path outside it.



RHABDOM'S SENSITIVITY AND RESOLUTION are a function of its diameter and distance from the posterior nodal point. A narrow rhabdom (*a*) with a diameter d_1 less than the width of the blur circle at 50 percent intensity will receive only a fraction of the light in the blur circle, but it will respond well to the variation of light across the blur circle and thus will have high resolution but low sensitivity. An ommatidium with such a rhabdom is said to be near the diffraction limit. A wider rhabdom (*b*) with a diameter d_2 equal to twice the width of the blur circle will capture 95 percent of the light in the blur circle, but it will have a slightly wider field and therefore poorer resolution. An ommatidium (*c*) with a larger rhabdom diameter d_3 has a wide field of view that is no better than *b* for point sources, but it is better for detecting extended objects in dim light. The angular sensitivity, or shape of the field, of each rhabdom (measured across the width at 50 percent sensitivity) is shown in the curves at the bottom of the illustration. The sensitivity for large objects increases with rhabdom size in proportion to $(d/f)^2$ in the same way that photographic film is more sensitive when its grains are larger.

$\Delta\rho$ will catch more light and will have a better signal-to-noise ratio than a receptor with a narrow field of view (by a factor of $\Delta\rho^2$), but a receptor with a large field of view will also be less sensitive to the modulation produced by narrow stripes. Similarly, an ommatidium with a large facet diameter D will experience more absolute modulation by a factor D^2. An efficient transfer of the modulation into the eye is essential, particularly when the intensity of the ambient light is low. Therefore if one could determine the level of the light signal required to activate the receptors above the level of the light noise, one could arrive at the optimum relation between the diameter of the facets and the interommatidial angle for insects that are active in different levels of light.

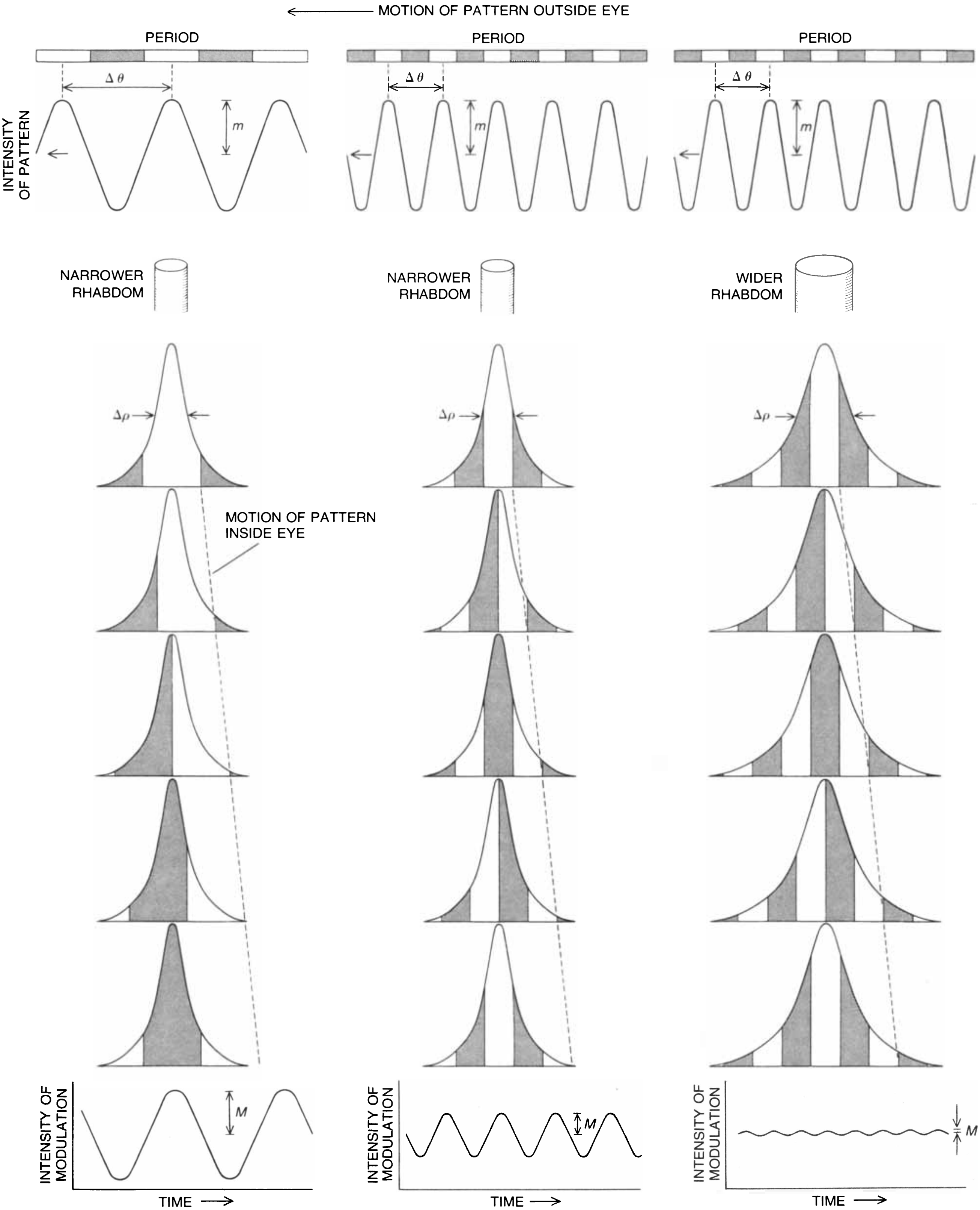


EYE MAP shows the front part of the right eye of the gomphid dragonfly *Austrogomphus*, which has one fovea looking obliquely upward and forward. Each circle represents the narrowest possible field of each fifth ommatidium. To compensate for the factor of five the diameter is equal to five times the wavelength of green light (taken to be .5 micrometer) divided by the diameter of the ommatidium's lens, or $5\lambda/D$. Smaller circles mean that the facets in that area of the insect's eye are larger and have a higher resolution than those in areas with larger circles. The center of each circle represents the position of the optical axis of each fifth ommatidium; the distance from the center of one circle to the center of the next is hence equal to five times the interommatidial angle, or $5\Delta\phi$. Amount by which the circles are separated or overlap yields the value of the eye parameter $D\Delta\phi$, which is a critical measure of how far from the diffraction limit the insect's eye is in each area. In this eye there are only about a fourth of the number of facets there would be at the diffraction limit. Center of fovea is at the center of grouping of small circles near 0 degrees, 50 degrees on axes in angular space.

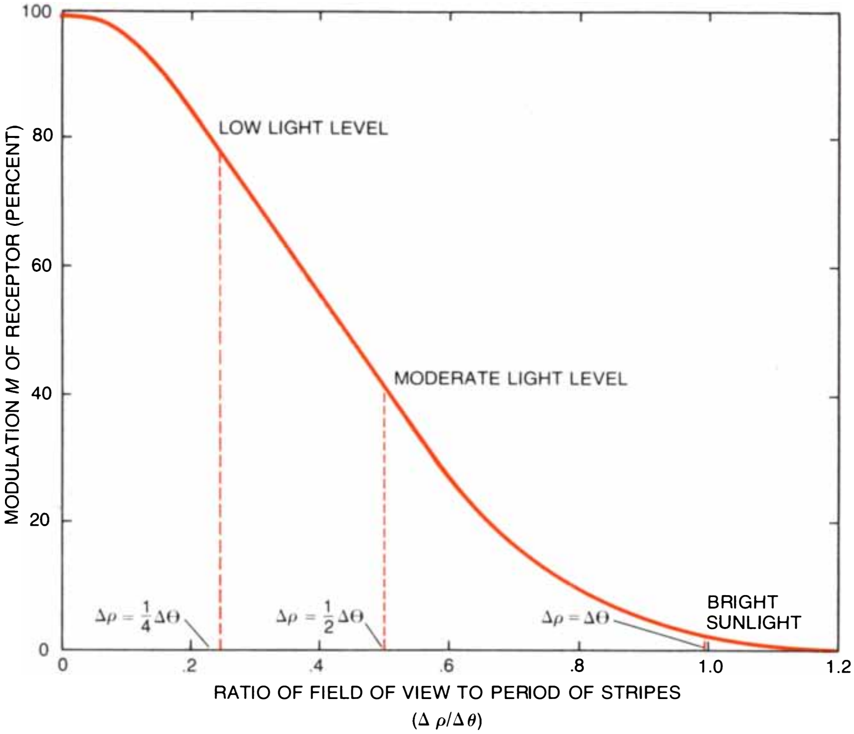
An extended light source of constant brightness is emitting an average number $\bar{N}$ of photons per unit of time. The photons nonetheless arrive at the rhabdom randomly; at any one moment the rhabdom may receive more than the average number of photons and at the next moment it may receive fewer than the average number. The variation in the actual number of photons from the average number of photons is the unavoidable noise; it is defined as the standard deviation from the average number of photons, and for light it is equal to the square root of the average number of photons, or $\sqrt{\bar{N}}$.

The absolute signal that the rhabdom actually receives is the relative amplitude of the modulation M multiplied by the average number of photons caught $M\bar{N}$. The ratio of the signal to the noise is therefore $M\bar{N}/\sqrt{\bar{N}}$, which is equal to $M\sqrt{\bar{N}}$. This ratio of the signal to the noise must be large enough for the signal to stand out above the noise, so that the ommatidium does not respond to false alarms. If $\bar{N}$ therefore decreases at the limit, M must correspondingly rise. If the rhabdom is receiving an average of, say, four photons per unit of time from a light source, the noise in that number of photons is plus or minus two photons; thus the modulation M in the average number of photons (the signal) must be greater than 50 percent of the average number of photons for the modulation to stand out above the noise. In bright light, however, the modulation can be lower as long as the product of $M\sqrt{\bar{N}}$ is still about 1. If the rhabdom is receiving an average of, say, 100 photons per unit of time, the noise is plus or minus 10 photons, so that the modulation M in the average number of photons need be only about 10 percent.

There are now enough equations to calculate the eye parameter $D\Delta\phi$ for any value of the ambient intensity of light, and the results appear as the smooth curve in the illustration on page 117 (which is from the recent work of Allan Snyder, Doekele Stavenga and Simon Laughlin). From that curve the amount by which the circles in the eye maps should overlap for optimum vision can be calculated for any intensity of ambient light. For a diffraction-limited eye, which must have an eye parameter of .3 or less, the circles of diameter λ/D on the eye maps must overlap each other almost halfway, and the actual field width $\Delta\rho$ should be twice the interommatidial angle $\Delta\phi$. For the eyes of insects that have an eye parameter of .5 the width of the field should be about 1.3 times larger than the interommatidial angle; for those with an eye parameter of 2 the width of the field $\Delta\rho$



PATTERN OF DARK AND LIGHT STRIPES (*top*) moving across an ommatidium's field of view is perceived by the rhabdom as a modulation of the intensity of light in time. The stripes have a sinusoidal distribution of intensity of amplitude m from light to dark (*curves second from top*). The amount of modulation perceived by the rhabdom (*cylindrical structures third from top*) is a function of the rhabdom's field of view $\Delta\rho$ and of the stripes' spacing $\Delta\theta$. At the left a narrow rhabdom looks at a moving pattern of wide stripes; the rhabdom's field of view is a quarter of the period of the stripes ($\Delta\rho = \frac{1}{4}\Delta\theta$). The sensitivity of the rhabdom across its field of view is represented by the shape of the peaked curves below the cylindrical rhabdom structures. As the wide stripes move across the rhabdom's narrow field of view the rhabdom perceives a modulation of large amplitude M (*curve at bottom left*). In the middle a narrow rhabdom looks at a moving pattern of narrow stripes ($\Delta\rho = \frac{1}{2}\Delta\theta$) and perceives a faster modulation of moderate amplitude. At the right a wide rhabdom looks at a moving pattern of narrow stripes ($\Delta\rho = \Delta\theta$) and sees a rapidly flickering gray blur that could be below optical noise level for all conditions except the brightest illumination or the largest eye facets. The diagonal broken lines show the motion inside the eye.



AMPLITUDE OF RELATIVE MODULATION *M* perceived by the rhabdom as a striped pattern moves through its field of view is shown as a function of the ratio of the width of the field of view to the spacing of the stripes. When the field is equal to the spacing of the stripes ($\Delta\rho = \Delta\theta$), the modulation is only a few percent, which may be sufficient for vision in bright sunlight. When the field is equal to half the spacing ($\Delta\rho = \frac{1}{2}\Delta\theta$), modulation is about 40 percent, as would be required at lower light levels to overcome photon noise. Hence receptors for working at low light levels have large field ($\Delta\rho$), for which they must have large diameter (*d*).

should be only three-quarters of the interommatidial angle $\Delta\phi$. Whether these field sizes are actually found in insect eyes must still be checked in a wide range of different kinds of insects.

The sensitivity of the insect eye can be increased by a factor D^2 by increasing the aperture D of the lens in the ommatidium or by a factor $(d/f)^2$ by decreasing the distance f from the posterior nodal point to the rhabdom. In either case the field of view is widened. In the compound eyes of large dragonflies, locusts, mantids, wasps and beetles (and also in the eyes of many crustaceans) the curvature of the lens is on the inside, so that the posterior nodal point is a little closer to the rhabdom and the eye can function equally well whether they are in water or out of it. The outside surface of the eye is then almost glassy smooth. In many insect eyes there is an absorbing pigment around the rhabdom in bright light. This pigment absorbs light in eyes that have to function over a range of intensities. In dim light the pigment moves back, increasing the rhabdom's sensitivity and sometimes widening the ommatidium's field of view. In some insects adaptation to low light levels is achieved by the movement of the receptor toward the lens. Compound eyes that adapt to darkness in this way are known as mobile-cone (or acone) eyes: they are found in hemipteran bugs, mosquitoes and some beetles.

Many nocturnal insects and some diurnal moths and beetles have superposition eyes with a quite different optical system, with the posterior nodal point at the center of curvature of the entire eye. The sensitivity of such eyes is increased by having the receptors close to the posterior nodal point. In still other insects the value of the eye parameter $D\Delta\phi$ at one position on the eye is quite different from the value at another position, suggesting that the eye is sensitive to a broad range of intensities. One can now see the value of knowing the eye parameter for any part of the compound eye before proceeding to study the eye in other ways.

For example, consider the eye of the housefly. When the eye parameter $D\Delta\phi$ measured from eye maps is compared with the ambient intensity of light at which the fly is usually active, one immediately sees what seems to be a discrepancy: the eye of the fly is apparently designed to operate at an ambient intensity some 10,000 times lower than that of full sunlight. The discrepancy is resolved when one remembers that flies not only prefer shady places but also use their eyes as they turn at high speed. High-speed maneuvers cause the environment to pass the eye at high angular velocities, reducing the effective number of photons available to a photoreceptor. High-speed turns decrease the perceived intensity of the light by a factor of about 1,000. Such considerations have to do with why hover flies hover and perching dragonflies and the praying mantids remain motionless.

With the optics of the compound eye in mind one can turn to some maps of the eyes of insects prepared from measurements of the pseudopupil. Each circle on the eye maps has a diameter of $5\lambda/D$ radians, with the wavelength λ taken to be .5 micrometer (green light). Each circle represents the width of the blur circle of each fifth ommatidium. The size of the circles is thus the key to the local resolution of individual ommatidia: the smaller the circle, the better the resolution. The distance from the center of one circle to the center of the next is five times the interommatidial angle ($5\Delta\phi$). The proximity of the circles is a measure of the density with which the insect eye samples the visual world and so is a measure of the visual resolution in that region of the eye. The overlap of the circles shows the value of the eye parameter $D\Delta\phi$ at that place. For example, where the circles just touch, their diameter λ/D is equal to their separation $\Delta\phi$ and the eye parameter has a value of .5.

I have compiled maps of the eyes of representative types of insects and have found a number of general features.

First, there do not seem to be any insects whose eyes are designed primarily to operate near the diffraction limit. Earlier values suggesting that insect eyes are limited by diffraction are wrong. Even the eyes of the Australian sand wasp *Bembix,* which is active in the brilliant sunlight of the desert, have an eye parameter of .32 only in the center of the fovea. The eyes of mantids, which catch prey that are usually stationary in bright light, have an eye parameter that is surprisingly large, ranging between .5 and 2.

Second, most insects that fly in bright sunlight, such as bees, butterflies and dragonflies, have an eye parameter of about .5. This generalization should be applied with care, because a small change in the value of the eye parameter near that value implies a large difference in the optimum intensity at which the eye would function best.

Third, the eyes of carnivorous insects seem to have a larger eye parameter than the eyes of herbivorous insects in the same habitat, but the carnivorous insects also have foveas.

Fourth, it is common to find a smooth gradient in the value of the eye parameter horizontally across the eye of a single insect, from smaller values near the front of the eye to larger values around the side. Presumably the gradients in the eye parameter are related to the fact that when the insect is flying, objects at the side of the eye have a greater angular velocity and give rise to less modulation than objects in front of the eye. The gradients are best seen by following a horizontal row of facets in the eyes of locusts, hymenopterans and dragonflies, but they are also characteristic of mantids, ants and other insects that do not have horizontal rows of facets or rapid forward motion.

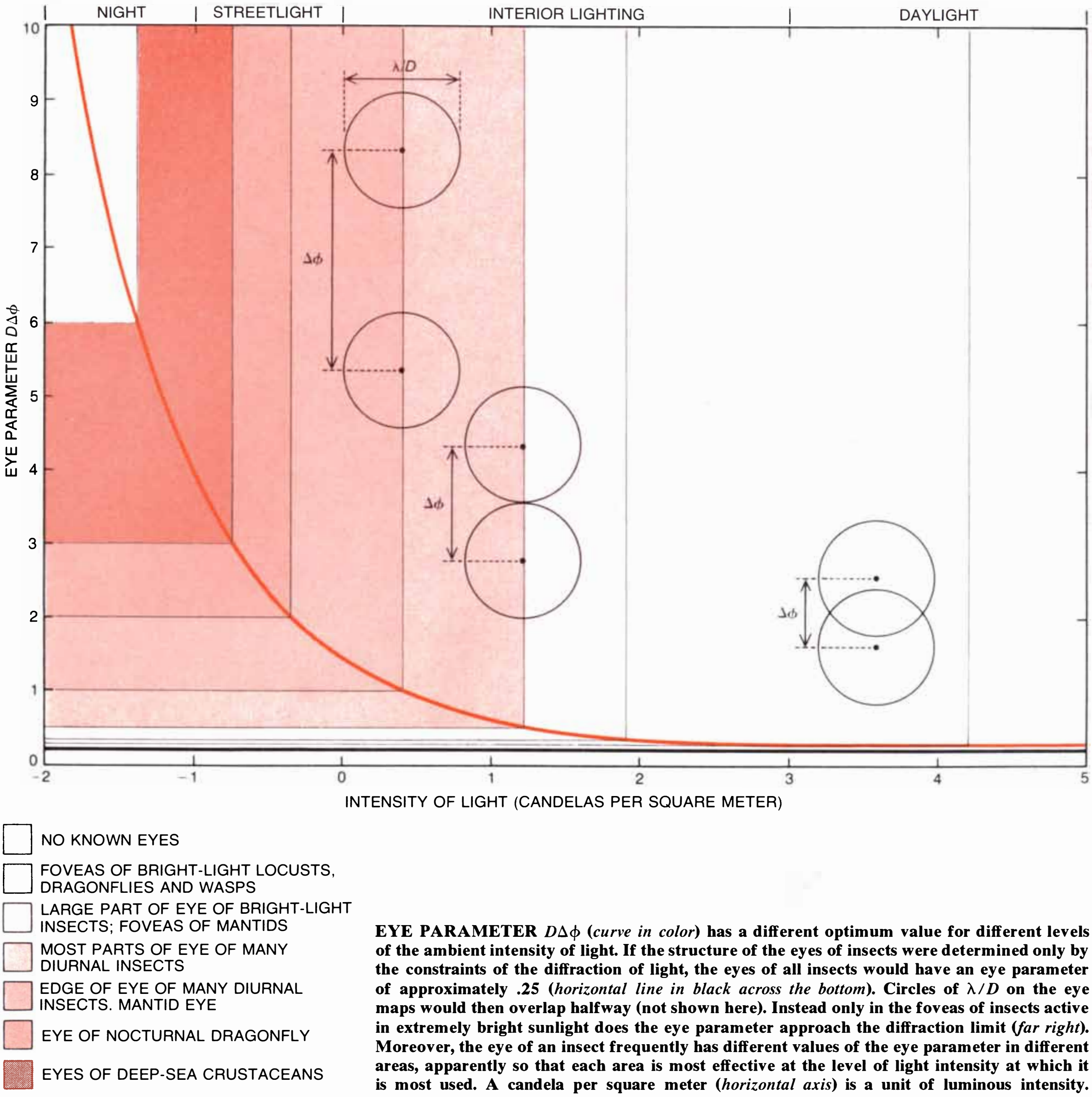
Fifth, along the equatorial region of the eye the pseudopupil is vertically elongated, because here the interommatidial angle between the facets in vertical rows is smaller than the interommatidial angle between the facets in horizontal ones. Although the apparent hexagonal pattern of the facets on the eye is little modified where the pseudopupil is elongated, the optical axes of the facets are compressed in the vertical direction to form a diamond lattice, so that the eyes look out on the world from an array of oblique squares. As we have seen, such a square pattern enables the eye to sample the visual world more efficiently than a hexagonal pattern does. These considerations suggest the intriguing possibility that the eyes might use the vertical rows of ommatidia rather than the horizontal rows for tasks requiring high resolution.

Sixth, where there is much binocular overlap the eye is stretched over a wider total angle, and the eye parameter is then always larger than expected. This is the case in mantids and damselflies.

The eye maps also reveal much information about the foveas of insects. First, a fovea is formed where rows of optical axes are gathered together, so that the density of ommatidia looking out is greater. A fovea is never formed by the interpolation of extra rows of ommatidia. The opposite is the case: in the foveal region rows of facets have to be left out to make room for the larger facets.

Second, the optical axes are gathered together by a warping of the rows of axes; the fact that the facets in a row are in a straight line does not necessarily indicate that their optical axes are lying in one plane.

Third, on the eye maps for some insect predators the circles of diameter λ/D overlap more in the fovea than they do in adjoining areas of the eye. This means that in the fovea the eye parame-



EYE PARAMETER $D\Delta\phi$ (*curve in color*) has a different optimum value for different levels of the ambient intensity of light. If the structure of the eyes of insects were determined only by the constraints of the diffraction of light, the eyes of all insects would have an eye parameter of approximately .25 (*horizontal line in black across the bottom*). Circles of λ/D on the eye maps would then overlap halfway (not shown here). Instead only in the foveas of insects active in extremely bright sunlight does the eye parameter approach the diffraction limit (*far right*). Moreover, the eye of an insect frequently has different values of the eye parameter in different areas, apparently so that each area is most effective at the level of light intensity at which it is most used. A candela per square meter (*horizontal axis*) is a unit of luminous intensity.

ter decreases, which in turn implies that in general the fovea is a region adapted for seeing in relatively bright light. In certain insects, however, such as the gomphid dragonfly *Austrogomphus*, the eye parameter remains constant across the fovea.

Fourth, the rows of facets are always more regular in the fovea than they are elsewhere on the eye, showing that regularity plays some important role in those regions of the eye where there is a need for the best possible seeing.

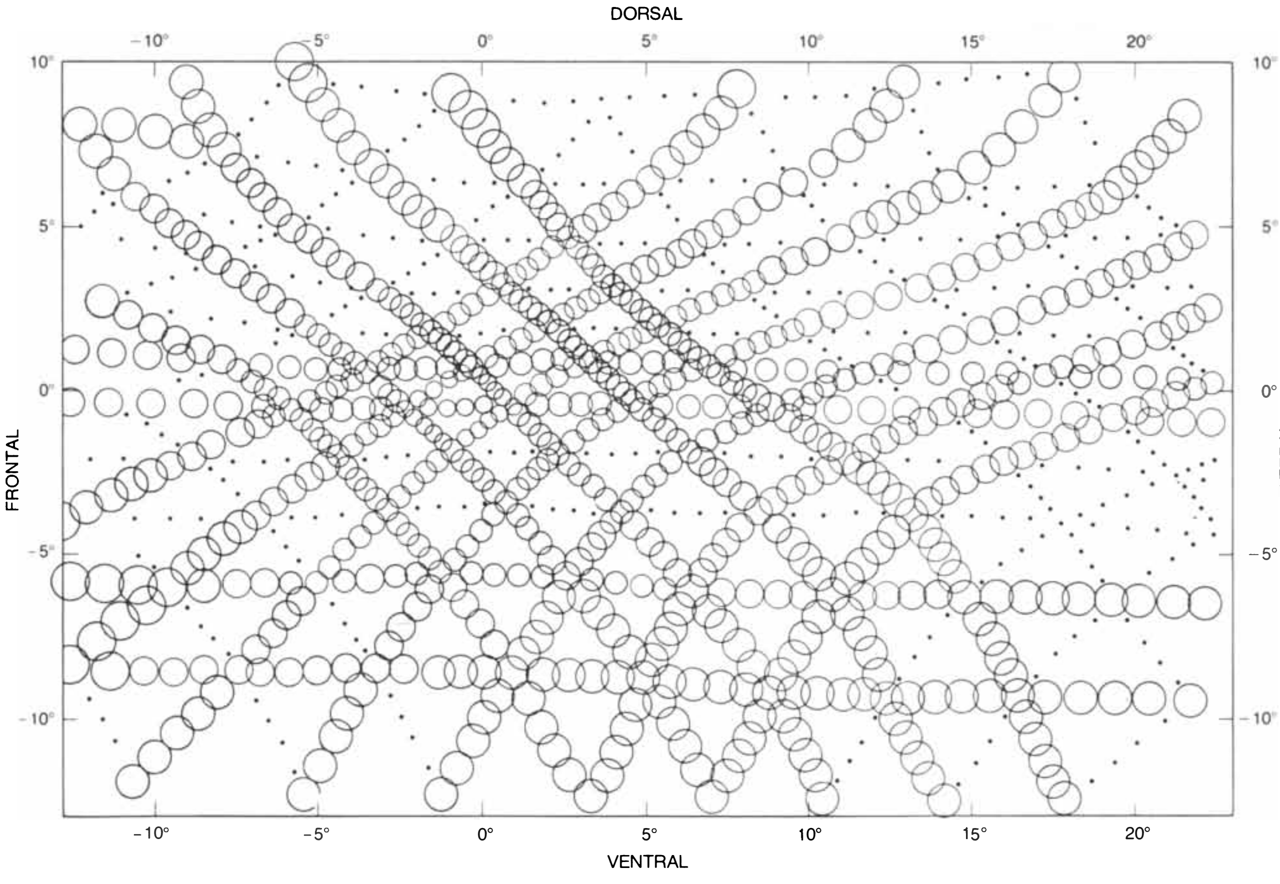
Fifth, although the facets in the fovea are usually larger than those elsewhere on the eye, that is not the case in the eyes of the locust *Locusta* and the mantid *Orthodera*. Both of those insects have a forward-looking fovea where the interommatidial angle is smaller but where the facets are the same size as those elsewhere on the eye. Such a fovea is perhaps a specialized area that enhances the visual acuity but has narrower and therefore less sensitive fields.

Sixth, in the sand wasp *Bembix*, which hovers to see in the extremely bright sunlight of the Australian desert, the eye parameter is as low as about .31 in the fovea, close to the diffraction limit. Many insects, however, have a remarkably large eye parameter even in the fovea. For example, the eye parameter reaches 1 in the fovea of the damselflies, which search for prey among reeds and under trees. Although in all cases the fovea is a region of higher visual resolution, it is clear that it can be a region of low sensitivity or can by having larger facets be designed to operate at the same intensity as the rest of the eye.

Although it must be true that the visual habits of insects are written on their eyeballs, it is difficult to read the signs correctly because we know so little about how the eyes are used. Presumably eyes that look sideways, such as those of some birds or of the rabbit, are early-warning devices. A clue is provided by locusts and flies, which are very sensitive to movement at the side of the head and are relatively insensitive to movement at the front. The large lateral fovea that looks directly sideways in many large dragonflies is puzzling; possibly it helps the animal to hover over water. The interommatidial angles between the facets of that fovea are less than .5 degree, the smallest I have found in any insect.

Insects that hunt visually would presumably need a forward-looking fovea. There is still some doubt, however, about whether most insect foveas, like those of vertebrates, work better in brighter light. If they do, then dragonflies, hunting wasps and mantids may suffer a decrease in visual acuity at dusk, much as human beings do.

The eyes of mantids are puzzling in many ways. In most mantids the fovea is quite obvious. Yet in 1971 J. C. Barros-Pita and H. Maldonado in Venezuela showed that the area immediately surrounding the fovea as it is defined here (not the center of the fovea) is the area that is essential in enabling the insect to see well enough to catch prey. Moreover, out of a large selection of mantids for which I have made eye maps none had circles of λ/D that overlapped outside the fovea region. Even more remarkable, the eyes of the largest mantids, which have facets as much as 50 micrometers across at the center of the fovea, have an eye parameter elsewhere greater than 2; thus the center-to-center separation of the circles on the eye maps is more than twice their diameter of λ/D. The only available explanation is that the mantid eye has 45 degrees of binocular overlap on each eye, so that the



EYE MAP OF SAND WASP *Bembix palmata* reveals that the optical axes of the insect's ommatidia (*dots and centers of circles*) do not follow the same hexagonal pattern that its eye facets do. Along the horizontal axis of the map they are in a pattern of oblique squares that become more oblique toward the equator of the eye. The circles have a diameter of λ/D. The large overlap between them indicates that the eye parameter is about .3. *Bembix* is one of the few insects for which the eye parameter (and then only the parameter in the fovea near 0 degrees, 0 degrees in angle coordinates) very nearly approaches the diffraction limit. *Bembix* is active in the brilliant sunlight of the Australian desert, so that photon noise is low with respect to intensity of the light and the diffraction limit can be approached.

eyes are stretched over an abnormally wide visual world.

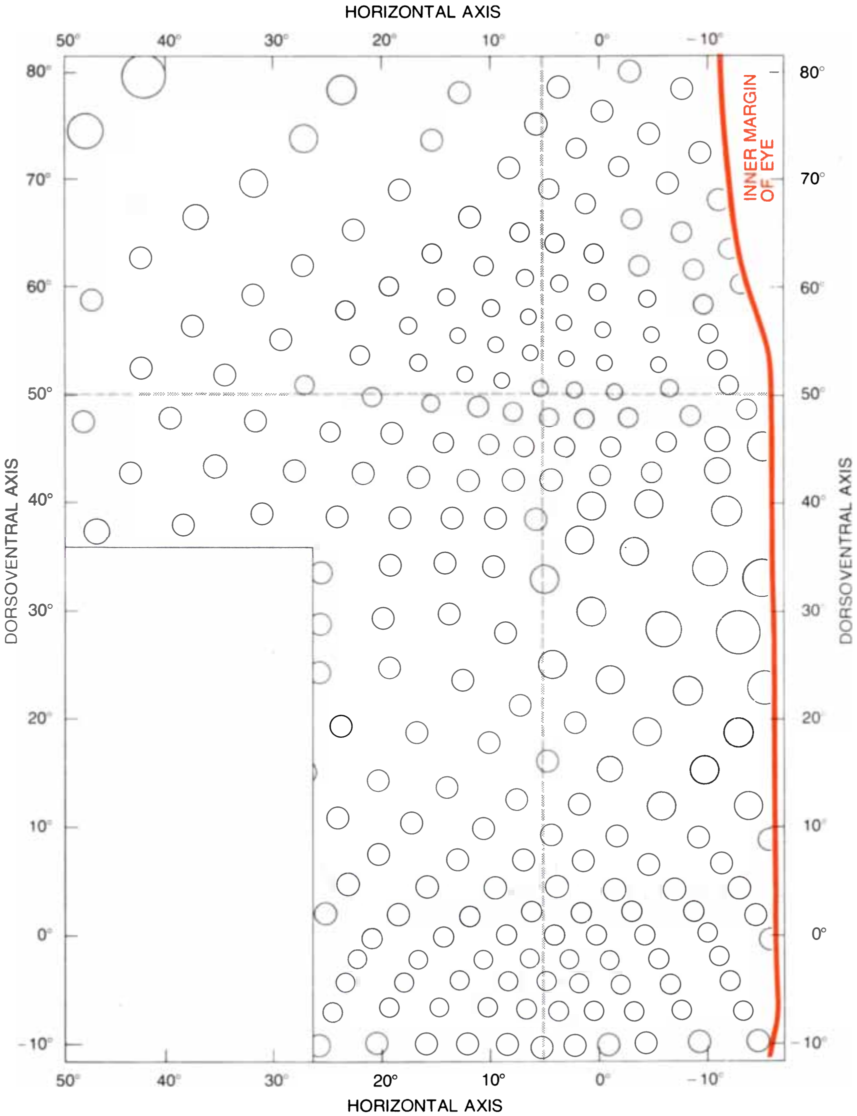
At first the fact that dragonflies have foveas that look obliquely forward and upward is also puzzling. It may be relevant, however, that a helicopter flying at full speed tilts forward so that the lift of its rotating blades is also the forward thrust. Dragonflies have some resemblance to helicopters, and it is possible that their upward-looking foveas are actually looking straight at the prey they are pursuing in the air. Moreover, an upward-looking fovea is exactly what an insect needs in order to anticipate the position of a falling object or to approach prey from below. In addition the top of the dragonfly eye is predominantly sensitive to short wavelengths of light, which suggests that the eye perceives the highest contrast and has the best resolution when it is looking at objects against the blue of the sky.

The theoretical relation between the value of the eye parameter and the effective intensity of light at which each part of the compound eye must function serves to bring together a vast amount of data. The theory is underlain by three basic assumptions: first, that the compound eye is an array of receptors that processes information simultaneously through parallel channels; second, that the facets and rhabdoms of the eye must be large enough to catch a sufficient number of photons to perceive the modulation above the noise, and third, that a striped pattern resolved by any single receptor can also be seen as a pattern by the eye as a whole.

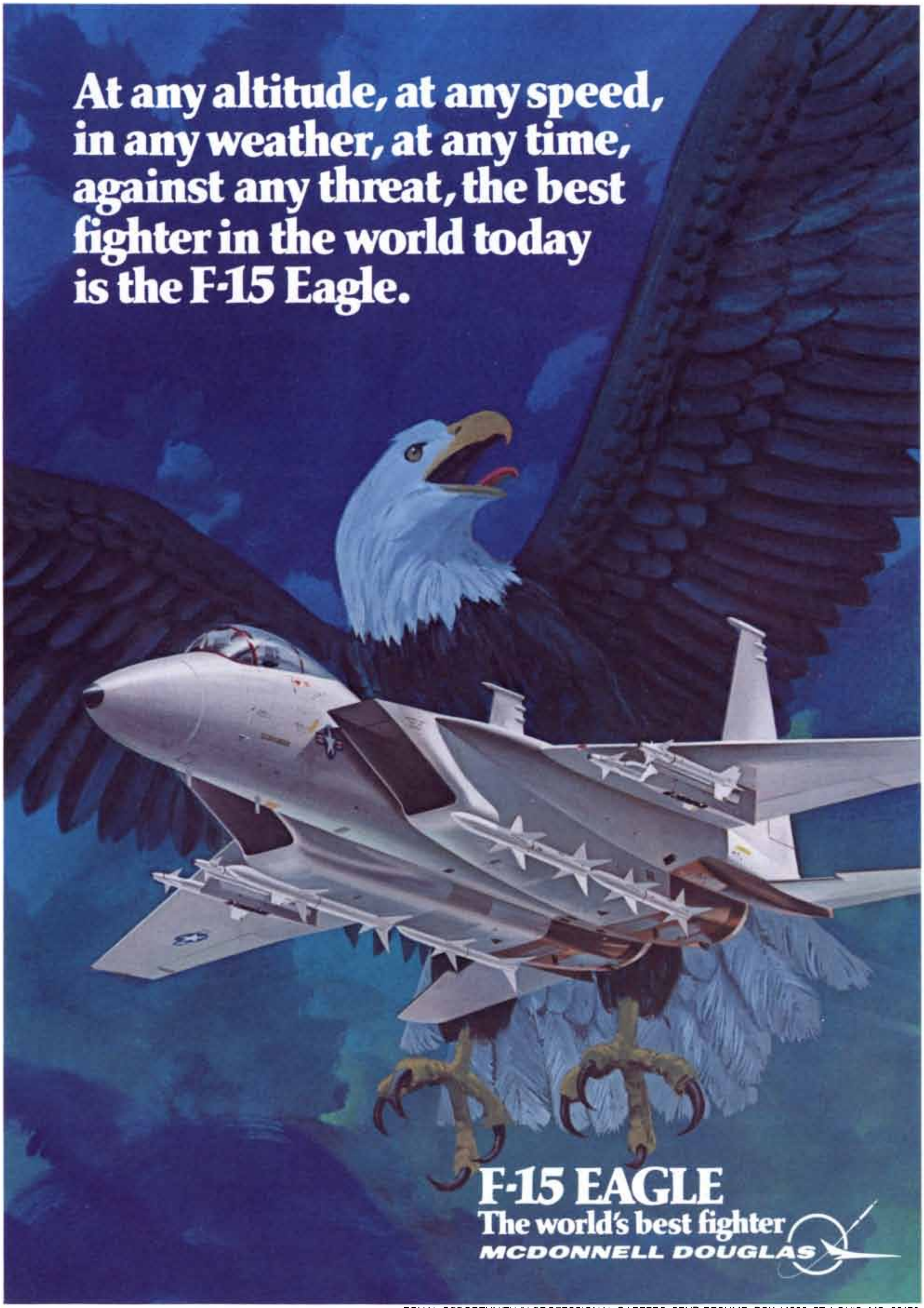
From a survey of studies of insect vision one might raise the objection that there is little evidence the last assumption is true, or that a fine pattern is in fact reconstructed by any compound eye. Such evidence does, however, exist. First, numerous tests with bees that have been trained to respond to patterns show that two aspects of the pattern are important to a bee: the amount of flicker the pattern creates as it sweeps across the compound eyes and the angles and colors of the pattern's contrasting edges. Patterns that look very different to the human eye are readily confused by the trained bee if the patterns cause similar amounts of flicker. Second, for an insect to perceive the speed and direction of a pattern moving across its eye the most important aspect of the signal is not the pattern itself but the sequential appearance of the signal at successive visual axes across the eye.

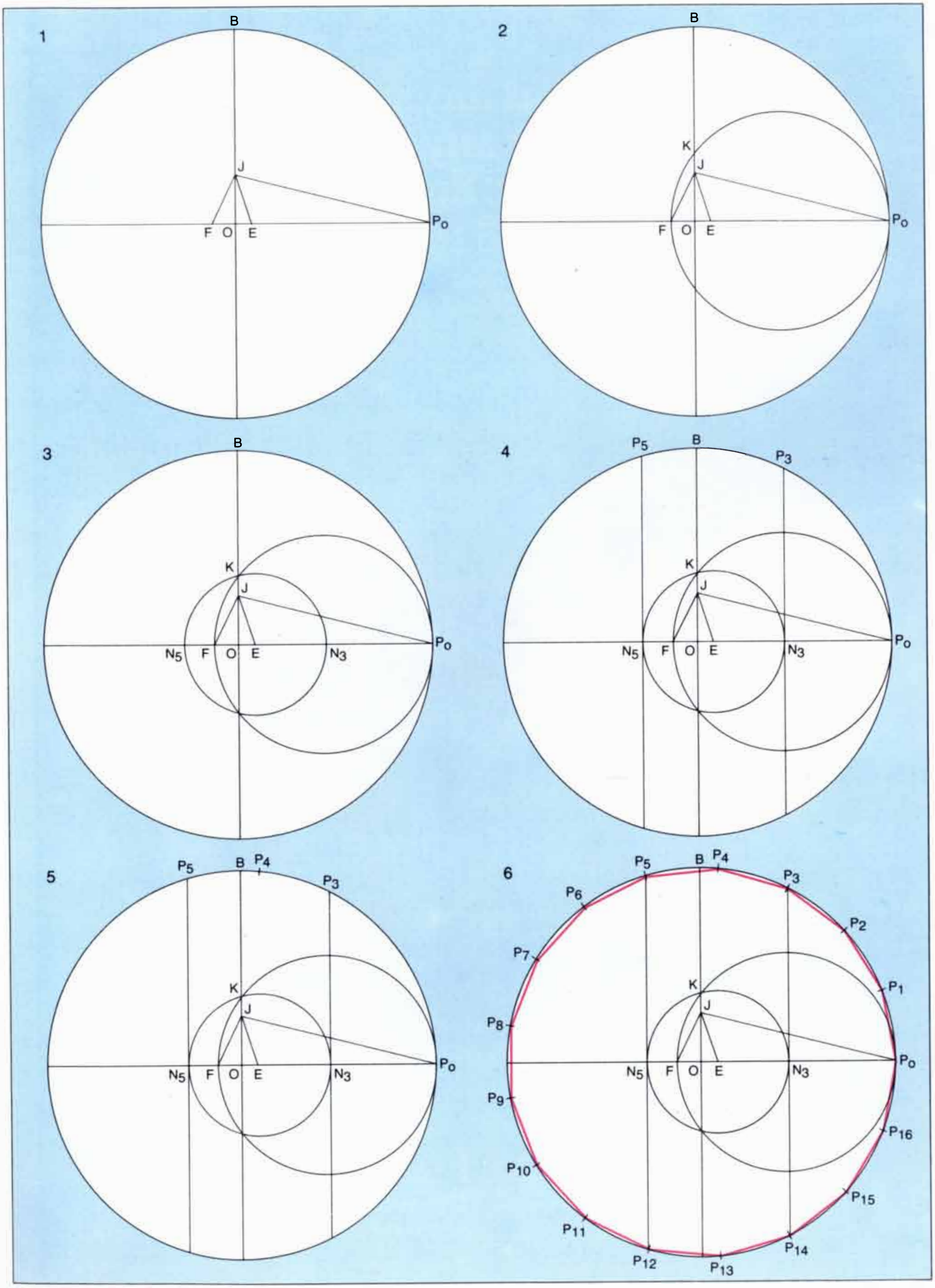
It is difficult to devise an experiment that will conclusively test whether or not an insect reconstitutes a pattern in its brain and thus perceives it as an actual pattern. It is easy to demonstrate, however, that an insect can respond to a very small movement of a large object and that it will respond particularly well to a small movement of a small object. In many small insects, which do not have room in their head for large eyes, the interommatidial angle can be as large as 10 degrees, yet a number of careful studies have shown that those insects can respond to detail subtending an angle as small as a tenth of the interommatidial angle. Part of the secret possibly lies in the fact that some insects may make use of eye tremor to scan the visual world and sequentially build up a more detailed picture of it as a pattern in time instead of in space. That would go against the assumption that sampling must be simultaneous.

To some extent the objections that can be raised against the assumptions underlying the optical theory I have discussed here cancel one another. For example, seeing motion requires that the insect be sensitive to the exact timing of small modulations at each visual axis over a region of the eye. Seeing the direction of a contrasting edge requires that the insect be sensitive to the exact phase of the sinusoidal (striped) components of the pattern. Seeing the relative movement of a small dark object in an unpredictable direction against a complex dappled background that is also moving, as a flying insect hunting for prey surely does, requires all the information the eye can get. I suspect that as we examine the visual behavior of insects more closely we shall find that there is a premium on gathering all possible information. The form of the most advanced compound eyes suggests that they are designed to collect the maximum amount of information they possibly can in a continuing battle against photon noise.



EYE MAP OF NOCTURNAL DRAGONFLY *Zyxomma* shows the optical axis of every fifth ommatidium. *Zyxomma*, an insect of Southeast Asia, comes out to feed only at twilight. The map shows the extreme separation of the circles (of diameters $5\lambda/D$ radians), the shallow fovea of the insect's eye looking obliquely forward and upward near 0 degrees, 50 degrees on the axes, and crowding of axes along horizontal zero axis (*bottom of map*). Broken lines show lines along which there are no errors due to mapping of curved surface of insect eye on a flat sheet.


At any altitude, at any speed, in any weather, at any time, against any threat, the best fighter in the world today is the F-15 Eagle.
F-15 EAGLE
The world's best fighter
MCDONNELL DOUGLAS


1
B
J
F O E
Po
2
B
K
J
F O E
Po
3
B
K
J
N5 F O E N3
Po
4
P5 B P3
K
J
N5 F O E N3
Po
5
P5 B P4 P3
K
J
N5 F O E N3
Po
6
P5 B P4 P3
P6 P2
P7 P1
K
J
P8
N5 F O E N3 Po
P9
P16
P10 P15
P11 P14
P12 P13

Gauss

A child prodigy who became the leading mathematician of his age, he was equally at home with the abstractions of number theory, the long calculations of astronomy and the practicalities of applied physics

by Ian Stewart

"Mathematics is the queen of the sciences," Carl Friedrich Gauss once said, and his own career exemplified that aphorism. Generally considered to rank with Archimedes and Newton as one of the ablest mathematicians of all time, Gauss was interested in both theory and application, and his contributions range from the purest number theory to practical problems of astronomy, magnetism and surveying. In all the branches of mathematics in which he worked he made profound discoveries, introduced new ideas and methods and laid the foundations of later investigations. It is a measure of his abilities that this year, the 200th anniversary of his birth, many of his ideas are still bearing fruit.

Gauss was in many ways an enigmatic and contradictory personality. The only son of working-class parents, he rose to become the leading mathematician of his age, yet he lived modestly and avoided public notice. His demeanor was mild, yet he was an aloof, politically reactionary and often unyielding man who asked only that he be allowed to continue his creative work undisturbed. He was prepared to recognize mathematical ability wherever he found it, in spite of contemporary prejudices, but his casual neglect of some of the best young mathematicians of his time, notably János Bolyai, one of the pioneers of non-Euclidean geometry, had unfortunate consequences.

A particularly striking aspect of Gauss's character was his refusal to present any of his work until he believed it had been polished to the point of perfection. No result, however important, was published until he deemed it to be complete. He reworked his mathematical proofs to such an extent that the path whereby he had obtained his results was all but obliterated. His published work has a quality of classical grace and elegance, austere and unapproachable. Many of his best ideas do not appear explicitly in print and have to be inferred by retracing the steps by which his discoveries must have been made. As a result important concepts did not see the light of day until they were discovered independently by others.

Gauss published some 155 titles during his lifetime, and he left behind a large quantity of unpublished work. In this brief survey I shall touch on his more important and influential discoveries and in some small measure attempt to explain how he arrived at them.

Gauss was born on April 30, 1777, in Brunswick (now in West Germany). His father worked variously as a gardener, a canal tender and a bricklayer. He was later described by his son as "an utterly honest and in many respects estimable and genuinely respectable man, but at home . . . domineering, coarse and rude." Gauss's mother, the daughter of a stonecutter, was an intelligent woman of strong character. Her brother Friedrich played an important role in the life of the young Gauss. He worked as a weaver of fine damasks, but his interests were unusually broad. He spent much time encouraging Gauss and sharpening his wits in argument.

Among the great mathematicians there are about as many who showed mathematical talents in childhood as there are those who showed none at all until they were older. Gauss was unquestionably the most precocious of them all. He joked that he had been able to count before he could talk, and many anecdotes attest to his extraordinary gifts. It is said that one day before he was three years old his father was making out the weekly payroll for the laborers in his charge, unaware that his son was observing the process intently. Finishing his calculations, the elder Gauss was startled to hear a tiny voice saying: "Father, the reckoning is wrong, it should be. . . ." When the computation was checked, the child's figure was found to be the right one. The remarkable thing about the story is that no one had taught him arithmetic.

Other tales concern Gauss's continued precocity at school. When he was 10, he was admitted to the arithmetic class. The master handed out a problem of the following type: What is the sum of $1 + 2 + 3 + \ldots + 100$, where there are 100 numbers and the difference between each number and the next is always one. There is a simple trick for doing such sums that was known to the master but not to his pupils.

The custom was for the first boy who solved a problem to lay his slate on the master's table with the answer written on it, for the next boy to lay his slate on top of that one and so on. The master had barely finished stating the problem when Gauss put his slate on the table. "There it lies," he said. For the next hour he sat with folded arms, getting an occasional skeptical look from the master, as the other boys toiled at the addition. At the end of the hour the master examined the slates. On Gauss's was a single number. Even in his old age Gauss loved to tell how of all the answers his was the only correct one.

To the master's credit he was so im-

GEOMETRIC CONSTRUCTION of a regular 17-sided polygon using only a straightedge and a compass, the first such discovery since Euclid, started Gauss on his mathematical career in 1796, at the age of 18. A simplified version of his construction, devised by H. W. Richmond in 1893, is shown in the illustration on the opposite page. The instructions are as follows: (1) Construct a circle with center at point O and radius OP_0 of arbitrary length. Draw OB perpendicular to OP_0. Find point J a quarter of the way up OB. Find point E such that the angle OJE is a quarter of the angle OJP_0. (This can be done by bisecting the angle twice.) Find point F such that the angle FJE is 45 degrees. (It is obtained by bisecting a right angle.) (2) Construct a circle with diameter FP_0. This circle intersects OB at point K. (3) Construct another circle with center at E and radius EK. This circle defines points N_5 and N_3. (4) Draw the lines N_3P_3 and N_5P_5 perpendicular to OP_0. (5) Bisect the arc P_3P_5 to obtain point P_4. (6) Using the chord P_4P_5, lay down that length sequentially starting at P_0. Connect points to obtain polygon.

pressed that he bought an arithmetic book with his own money and gave it to Gauss, who devoured it quickly. Gauss was also fortunate in that the mathematics master's assistant, a 17-year-old named Johann Martin Bartels, had a passion for mathematics, so that the two of them spent many hours studying together. On encountering the binomial theorem, which states that for all numbers n the expression $(1 + x)^n$ is a series, and that when n is not a positive integer, the series is infinite, Gauss was dissatisfied with the lack of rigor in the book the master had given him and constructed a proof. Although he was still a schoolboy, he was the first mathematician to pay serious attention to the problems of infinities. Few mathematical prodigies possess any ability beyond a facility for computation, but Gauss's abilities clearly extended into higher realms of thought.

At the age of 14 Gauss was introduced to the Duke of Brunswick, who had heard of his reputation and became his patron. The following year Gauss entered the Collegium Carolinum in Brunswick, where he studied and soon mastered the works of Newton, Leonhard Euler and Joseph Louis Lagrange. By the age of 19 he had discovered for himself and proved a remarkable theorem in number theory known as the law of quadratic reciprocity (to which I shall return). To appreciate how remarkable this was one must realize that although Euler had discovered the theorem earlier, both he and Adrien Marie Legendre had failed in their efforts to prove it.

When Gauss left the Collegium Carolinum in October, 1795, to study at the University of Göttingen, he was torn between mathematics and his other great love: the study of ancient languages, at which he was equally brilliant. On March 30, 1796, however, his mind was made up by one of the most surprising discoveries in the history of mathematics.

PORTRAITS of Gauss (*foreground*) and the physicist Wilhelm Weber (1804–91), with whom he collaborated on many practical experiments in magnetism and telegraphy, are shown in this elaborate etching. The Greek inscription on the ribbon reads: "God is an arithmetician." The Latin one at right reads: "The end crowns the work." The Greek quotation below, from the beginning of Plato's *Republic*, is translated: "Those who have torches will pass them to others."

To provide some background for this discovery, let us go back two millenniums to Classical Greece. The main Greek contribution to mathematics was the flourishing school of geometry associated with the names of Pythagoras, Eudoxus, Euclid, Apollonius and Archimedes. The Greeks were probably the first to recognize the importance of rigor in proofs, and in search of such rigor they had imposed a number of restrictions. One of them was that in geometric constructions only a straightedge and a compass were to be used. In effect the only curves allowed were the straight line and the circle.

Euclid showed that it was possible to construct regular polygons with three, four, five and 15 sides, together with the polygons derived by repeated bisection of these sides, with a straightedge and a compass. Such polygons, however, were the only ones the Greeks could construct; they knew of no way to make polygons with, for example, seven, nine, 11, 13, 14 and 17 sides. For the next 2,000 years no one appears to have suspected that it might be possible to construct any of these other polygons. Gauss's achievement was to find a construction for a regular polygon of 17 sides, which he inscribed within a circle using only a straightedge and a compass. Moreover, he explained exactly which polygons could be so constructed: the number of sides must be any power of 2 (2^n) or a power of 2 multiplied by one or more different odd prime numbers of the type known as Fermat primes (after Pierre de Fermat, who discovered them). A prime number is by definition a number that cannot be divided evenly by any number except itself and 1; a Fermat prime has the additional requirement of being one greater than 2 to a power of 2, or $2^{2^n} + 1$. The only known Fermat primes are 3, 5, 17, 257 and 65,537. Thus we have the remarkable result that although regular 17-sided polygons can be constructed with a straightedge and a compass, regular polygons with seven, nine, 11, 13 and 14 sides cannot be.

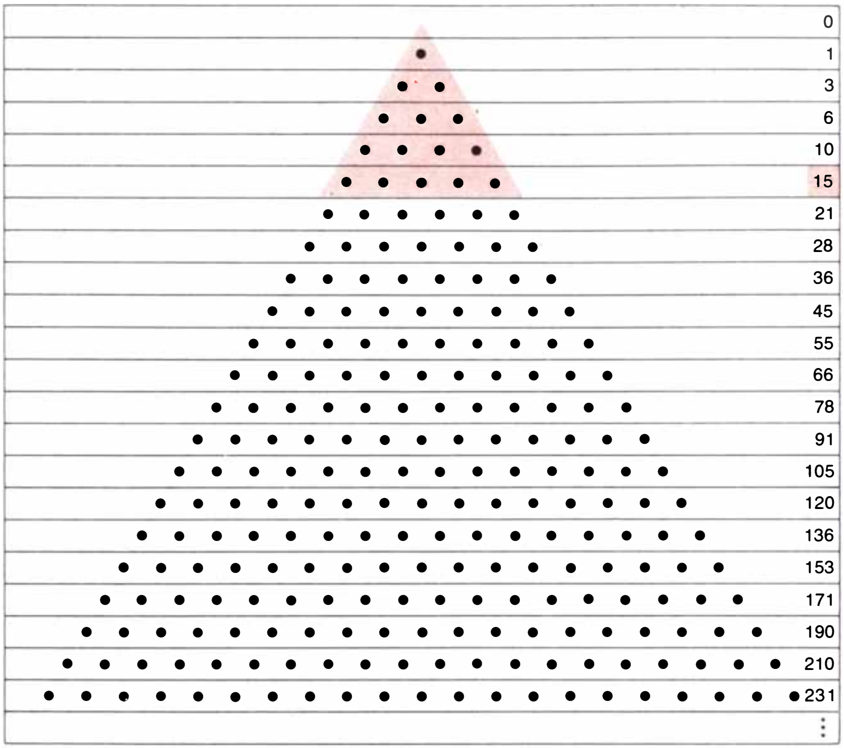
Gauss proved this theorem (at the age of 18) by combining an algebraic argument with a geometric one. He showed that constructing a 17-sided polygon is tantamount to solving the equation $x^{16} + x^{15} + \ldots + x + 1 = 0$. Because 17 is prime and 16 is a power of 2 it turns out that this equation can be reduced to a series of quadratic equations (expressions of the form $ax^2 + bx + c = 0$, where a, b and c are given numbers and x is to be found). Since it had already been proved that quadratic equations can be solved with a straightedge-and-compass construction, the proof was complete. Apart from the proof's importance in inducing Gauss to take up mathematics as a career, it is the first real instance (beyond Descartes's introduction of coordinates) of a technique that has since become one of the most useful in mathematics: moving a problem from one domain (in this case geometry) to another (algebra) and solving it there.

By the time he was 20, Gauss once wrote, he was so overwhelmed with mathematical ideas that there was not enough time to record more than a small fraction of them. Many of those he managed to pursue appeared in his *Disquisitiones Arithmeticae*, published in 1801, when he was 24. This work may be said to have done for number theory what Euclid did for geometry: it organized scattered discoveries about integers and supplemented them with some deep ideas of Gauss's own. Gauss based his theory on the concept of congruent numbers, which are defined as two integers, a and b, that have the same remainder when they are divided by a given number m. Any two numbers that satisfy this requirement are said to be congruent "modulo" the number m, which is known as the modulus. For example, 16 and 23 both have a remainder of 2 when they are divided by 7, and they are therefore congruent modulo 7. Seven and 9 are congruent modulo 2, since there is a remainder of 1 when each number is divided by 2. Clearly any two even numbers and any two odd numbers will be congruent modulo 2.

Gauss also pointed out the possibility of doing arithmetic with congruent numbers. He showed that, for integers a, b, c and d, if, modulo a number m, a is congruent to b and c is congruent to d, then $a + c$ is congruent to $b + d$, and ac is congruent to bd. Hence we can replace integers by congruent numbers in our arithmetic without running into contradictions. We do, however, get some surprises, such as the fact that, modulo 3, $1 + 1 + 1$ is congruent to zero.

This arithmetic of congruent numbers is taught in many "new math" courses under the name modular arithmetic. I wonder how many teachers realize where it came from and why Gauss developed it? He wanted to use it as a tool to prove deep and difficult theorems. Perhaps the greatest of these, and certainly Gauss's personal favorite (he called it the golden theorem), is the law of quadratic reciprocity. Here some additional terminology is necessary.

Gauss first defined a "quadratic residue" by stating that if m is a positive integer and a is an integer that has no factors in common with m, then a is a quadratic residue of m if it is congruent, modulo m, to a perfect square. Another way of saying this is: If a is a quadratic residue of m, it is possible to find at least one x whose square divided by m leaves the remainder a. Thus 13 is a quadratic residue of 17 because the statement "x^2 is congruent to 13 modulo 17" is satisfied if x takes the value (among others) 8. Gauss proved that if p and q are different odd primes, then p is a quadratic residue of q if, and only if, q is a quadratic residue of p. There is a single exception to that rule: if both p and q are of the form $4n + 3$, then one is a quadratic residue and the other is not. This result may at first seem strangely specialized, but it enables us to decide whether a given odd prime is a quadratic residue of another prime by asking a question that in many cases is simpler: Is the second prime a quadratic residue of the first? This theorem has inspired some deep ideas of modern algebra and is of great importance throughout number theory and in other branches of mathematics. Gauss himself thought it so significant that during his lifetime he proved it eight different ways.

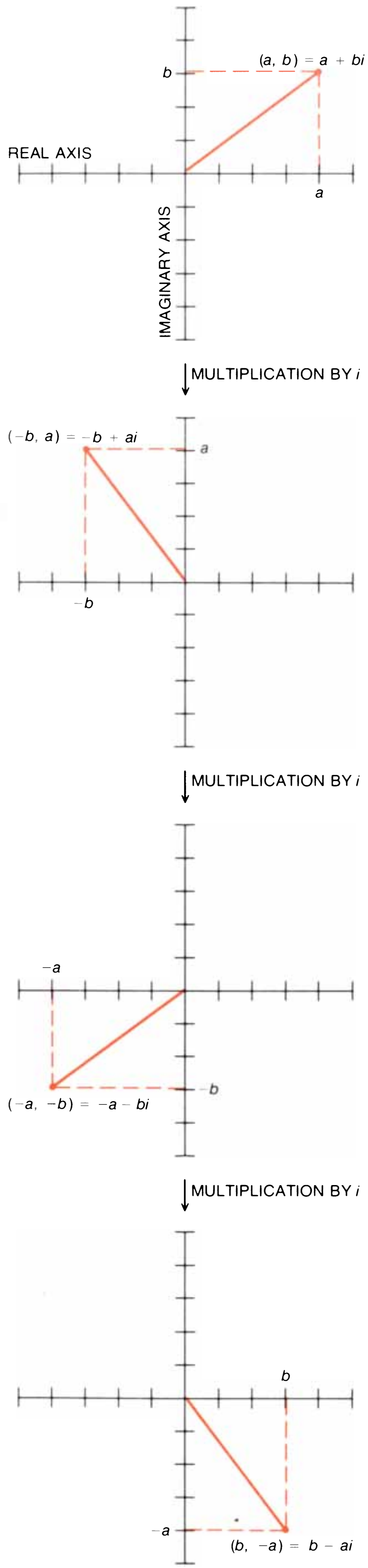


TRIANGULAR NUMBERS are numbers of form $n(n + 1)/2$, where n is any positive integer. They can also be represented as a triangular array of dots. In his *Disquisitiones Arithmeticae* Gauss proved that every positive integer is the sum of three such triangular numbers. He noted his discovery in his diary for July 10, 1796, with the cryptic entry: "Eureka! num = Δ + Δ + Δ."

The *Disquisitiones* displayed a tendency that for Gauss became a way of life. Proofs were polished until they shone, with every trace of the process by which they had been reached removed, if it was at all possible, so that only the finished structure remained. Gauss once said: "When a fine building is finished, the scaffolding should no longer be visible." Later generations, faced with the problem of understanding Gauss's methods as well as his results, may be forgiven if they accuse him of not only removing the scaffolding but also throwing away the plans. In mathematical investigation ideas and methods are often of greater importance than the theorems for which they were developed. A genuinely good idea can be extended into new fields to yield results that could not have been contemplated in advance. Here two aspects of mathematics are in conflict: mathematics as an art form and mathematics as a living discipline.

This view is not entirely a modern one. Gauss's contemporary Karl Gustav Jacobi said of him: "His proofs are stark and frozen ... so that one must first thaw them out." Another contemporary, Niels Henrik Abel, remarked: "He is like the fox, who erases his tracks in the sand with his tail."

Why did Gauss conceal his methods, preferring to give only a synthesis and suppressing the analysis? He adopted the motto *Pauca sed matura* (Few but ripe), which reflects his dissatisfaction with the incomplete theorems of his colleagues. One cannot be sure, but perhaps the near-poverty of his childhood led him to be parsimonious with the distribution of his ideas. Perhaps too he did not want to display half-finished work for fear of being ridiculed if it should turn out to be wrong. This fear is common among the great mathematicians;



COMPLEX NUMBERS of the form $a + bi$ (where a and b are integers and i is the square root of -1) can be represented as number pairs (a, b), or points of a plane, in the same way that real numbers are represented as points of a line. The complex number pairs can then be manipulated in a geometric fashion. For example, the rotation through 90 degrees of a line joining the origin and the point (a, b) is equivalent to multiplying the original complex number by i. (Three such rotations are shown in the illustration.) Gauss was the first mathematician to notice that this geometric interpretation could be used to obtain purely algebraic definitions of the sum and product.

Newton, for example, had to be persuaded to publish his *Principia*. It may even be justified; witness Georg Cantor, whose epochal work on set theory and transfinite numbers was ridiculed by leading mathematicians of his time. The experience seems to have led to his mental breakdown.

Much of Gauss's best work on number theory was related to the problem of complex numbers, which are defined as numbers of the form $a + bi$, where a and b are integers and i is the square root of -1 (so that $i^2 = -1$). Complex numbers were introduced by the Renaissance algebraists, but they were liberally endowed with mystical properties and issue-begging descriptions such as "real" and "imaginary." Even a man as intelligent as Leibniz was terribly confused by the subject, producing reams of nonsense about these mysterious numbers that are neither positive nor negative. "The Divine Spirit," he wrote, "found a sublime outlet in that wonder of analysis, that portent of the ideal world, that amphibian between being and not-being, which we call the imaginary root of negative unity."

Gauss was more prosaic, preferring to use a geometric representation of complex numbers as points in a plane. Although this approach had first been published in 1797, when a Norwegian surveyor, Caspar Wessel, devised an analytic representation of the geometry of the plane that was tantamount to complex numbers, his achievement was not noticed until 1897. A Swiss bookkeeper, Jean Robert Argand, developed a similar description in 1806, and he receives the credit in textbooks to this day. Gauss's contribution was to go beyond a purely geometric definition of complex numbers. In a letter he wrote in 1837 he says that in 1831 he had realized that a geometric interpretation can be avoided by using couples of the form (a, b) in place of $a + bi$, with purely algebraic definitions of the sum and the product. Using number pairs, he showed that the operations of arithmetic with complex numbers are defined by the rules: $(a, b) + (c, d)$ equals $(a + c, b + d)$ and $(a, b)(c, d)$ equals $(ac - bd, ad + bc)$. It is easy to verify that the couple $(a, 0)$ behaves just like the real number a, and that $(0, 1)^2$ equals $(-1, 0)$. Thus the number pair $(0, 1)$ is our mysterious square root of -1. The true value of this approach, which is the one generally adopted today, was not appreciated until William Rowan Hamilton published it in 1837. Gauss was the first mathematician to make free and extensive use of complex numbers and to give them full acceptance as genuine mathematical concepts.

In his doctoral dissertation for the University of Helmstedt, awarded in 1799, Gauss gave the first proof of the "fundamental theorem of algebra" (today more naturally proved as a theorem in topology): that every polynomial equation has a complex root. Here again Gauss thought the theorem was so important that during his lifetime he proved it four different ways. The third proof is particularly characteristic of his impenetrable style and original turn of mind. He constructs from a polynomial equation a complicated expression in the form of a double integral. If the polynomial has no root, this integral should take the same value whether it is integrated with respect to first one variable and then the second, or the other way around. Gauss shows that this is not the case, that the two processes give different results; therefore the assumption of no root is false, and hence a root exists.

To see where the proof came from we have to know that Gauss possessed the basic theorems of complex-number analysis but had not published them. His proof is a translation of an argument in complex-function theory into the more standard terms of real-number analysis. The results of this translation are logically rigorous, but they involve a certain perversity of formulation that obscures the motivating principles. Apparently Gauss felt that his theorems of complex-number analysis were not sufficiently complete for publication, and he recast his arguments accordingly.

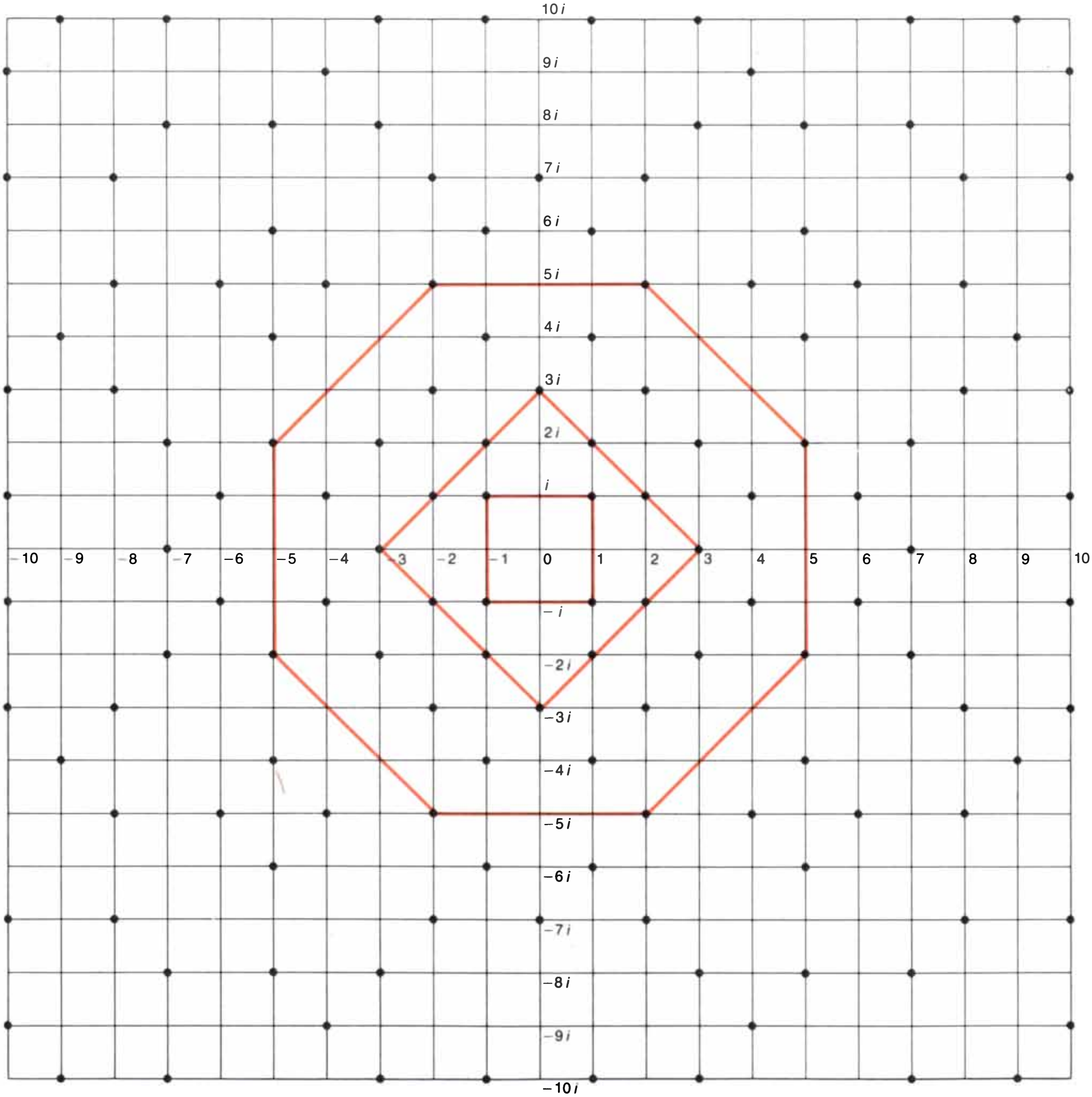
Gauss did much more with complex numbers. In 1811 he discovered what is now called Cauchy's theorem: The integral of a complex analytical function around a closed curve that encloses no singularities is zero. Augustin Louis Cauchy made it the foundation of complex-number analysis, and it still is. We do not call it Gauss's theorem because Gauss never published it. It seems likely that he intended to produce a definitive work on complex-number analysis but never found the time to work out the ideas to his own satisfaction.

Gauss also developed a method of factoring primes with complex numbers, with some intriguing results. The prime number 2, for example, can be factored in the form $(1 + i)(1 - i)$. In the same way, 5 can be factored as

$(2 + i)(2 - i)$, 29 as $(5 + 2i)(5 - 2i)$, and so on. Certain prime numbers, however (among them 7, 11 and 19), cannot be factored and remain prime. Gauss discovered that only primes of the form $4n + 1$ can be factored with complex numbers (the number 2 is a special case) and that each of these primes can be factored in exactly one way. Such methods were later used to solve problems that on the face of it did not involve complex numbers at all.

In particular Gauss used complex numbers of the form $a + bi$ (now called Gaussian integers) to formulate and prove a version of the law of quadratic reciprocity for biquadratic residues. The number k is said to be a biquadratic residue of another number m if k is congruent modulo m to the fourth power of an integer. Thus the biquadratic residues of 10 are 0, 1, 5 and 6. The law of biquadratic reciprocity states that for two odd prime numbers p and q there are connections between the statement "p is a biquadratic residue of q" and the statement "q is a biquadratic residue of p," with a host of conditions on the form of p and q. This theorem is analogous to the law of quadratic reciprocity, but it is much more cumbersome to state mathematically (and hence is very hard to conjecture, let alone to prove). If the theorem is extended to the case where p and q are Gaussian integers of the form $a + bi$, there are remarkable simplifications both in the statement of the result and in its proof. Thus going over to the case of complex numbers makes the problem easier and is more natural than employing the case of real numbers.

Gauss's proof of biquadratic reciprocity using Gaussian integers provided an archetypal procedure for solving many problems in number theory: First,



GAUSSIAN PRIMES, complex numbers of the form $a + bi$ having no factors of the same type, are distributed irregularly over the complex plane. Gauss discovered three classes: (1) $\pm 1 \pm i$, forming the vertexes of a square, (2) $\pm p$ and $\pm pi$ (where p is a real prime number equal to $4n + 3$), forming a diamond, and (3) $\pm a \pm bi$ and $\pm b \pm ai$, forming a truncated square. Gaussian primes always occur in one of these patterns.

extend the theorem to a domain of suitably chosen complex numbers called a number field, where it may be dealt with more naturally, solve it there and return to the case of ordinary integers at the end of the proof. This powerful method opened the door to what is now called algebraic number theory.

In 1801 Gauss's mathematical interests changed direction sharply when he became involved in astronomy. His love of calculation was clearly a contributory factor. Throughout his most erudite works are long calculations, and some of his deepest theorems in number theory were inferred by examining long lists of figures. Moreover, Gauss carried many of his calculations out to 21 decimals long before the availability of any kind of computer.

Gauss's interest in astronomy can be traced back to a discovery by Johann Titius, who in 1776 stated an empirical rule for the distances between the sun and the planets. Titius started with the series 0, 3, 6, 12, 24, 48 and 96, in which each term after the first is double the preceding one; then he added 4 to each term to get 4, 7, 10, 16, 28, 52 and 100. As it turned out, these numbers were very closely proportional to the distances from the sun of Mercury, Venus, the earth, Mars, Jupiter and Saturn, except that there was no planet at distance 28. This rule, which came to be known as Bode's law (because Johann Bode appropriated it without acknowledgment), remained a curiosity until 1781, when William Herschel discovered Uranus at a distance of approximately 196 units. Since the next term in the Titius-Bode sequence was $2(96) + 4 = 196$, interest now focused on the gap at distance 28.

On New Year's Eve, 1800–1801, Giuseppe Piazzi discovered what he thought was the missing planet. It was Ceres, which we now know as one of the thousands of small bodies in the asteroid belt between Mars and Jupiter. Once Ceres was sighted it became important to calculate the elliptical orbit of the new body before observers lost track of it. The difficulty of observing such a small object made the available data scanty and the calculation of the orbit all the harder; Newton himself had remarked how difficult it was to compute orbits from meager data. It was a golden opportunity for Gauss to follow in the footsteps of the man he admired most.

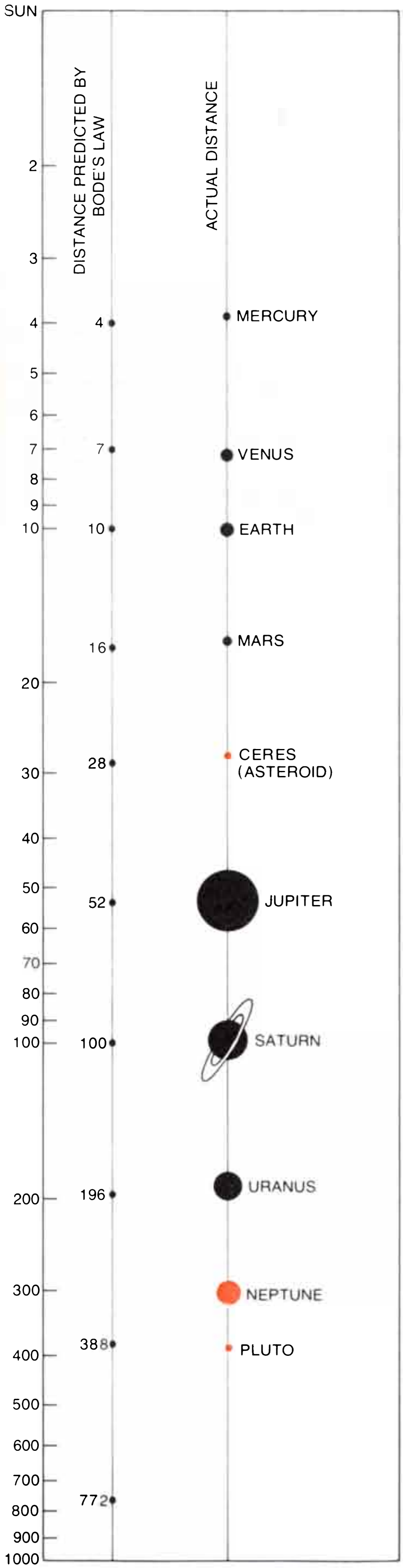
After only three observations Gauss developed a technique for calculating orbital components with such accuracy that late in 1801 and early in 1802 several astronomers were able to locate Ceres again without difficulty. As part of his technique he showed how the variation inherent in experimentally derived information could be represented by a bell-shaped curve (best known today as the Gaussian distribution). He also developed the method of least squares, by which the best estimated value is derived from the minimum sums of squared differences in a particular computation. His methods, which he described in an 1809 paper titled "Theory of the motion of the heavenly bodies revolving around the sun in conic sections," are still in service today. Only a few modifications were required to adapt them to modern computers.

Gauss had similar success in determining the orbit of the asteroid Pallas, for which he refined his calculations to take into account the perturbations caused by the other planets in the solar system. In 1807 he became professor of astronomy and director of the new observatory at the University of Göttingen, where he remained for the rest of his life. His first wife died in 1809, soon after the birth of their third child. From a second marriage were born two sons and a daughter.

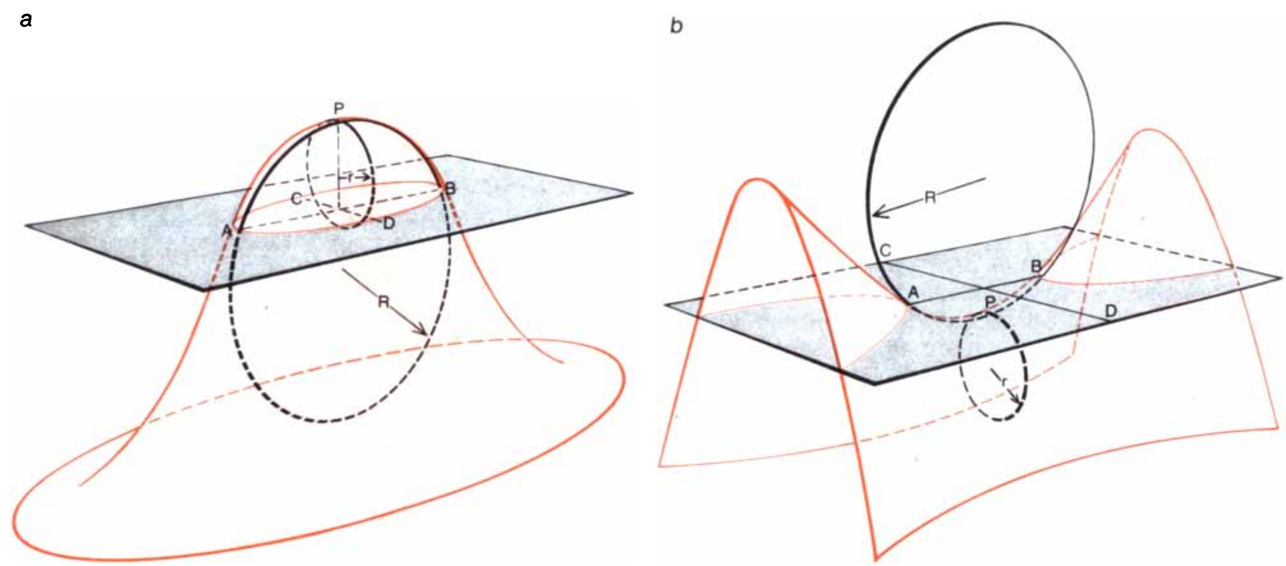
In about 1820 Gauss turned his attention to geodesy: the mathematical determination of the size and shape of the earth. To it he devoted much of the next eight years in theoretical studies and fieldwork. In 1821 he became scientific adviser to the governments of Hannover and Denmark, which charged him with making a geodetic survey of Hannover by means of the technique of triangulation. To this end he developed the heliotrope, a device that reflected the rays of the sun in a precisely specified direction, thus making it possible to accurately align surveying instruments over long distances.

Gauss's efforts to determine the shape of the earth by actual geodetic measurements led him back into pure theory. Working with data from his surveys, he developed a theory of curved surfaces in which the characteristics of a surface can be determined solely by measuring the length of the curves that lie on it. This intrinsic-surface theory inspired one of his students, Bernhard Riemann, to develop a general intrinsic geometry of spaces with three or more dimensions. Some 60 years later Riemann's ideas formed the mathematical basis for Albert Einstein's general theory of relativity.

Beginning in 1831, when the physicist Wilhelm Weber arrived in Göttingen, Gauss worked closely with him on ex-



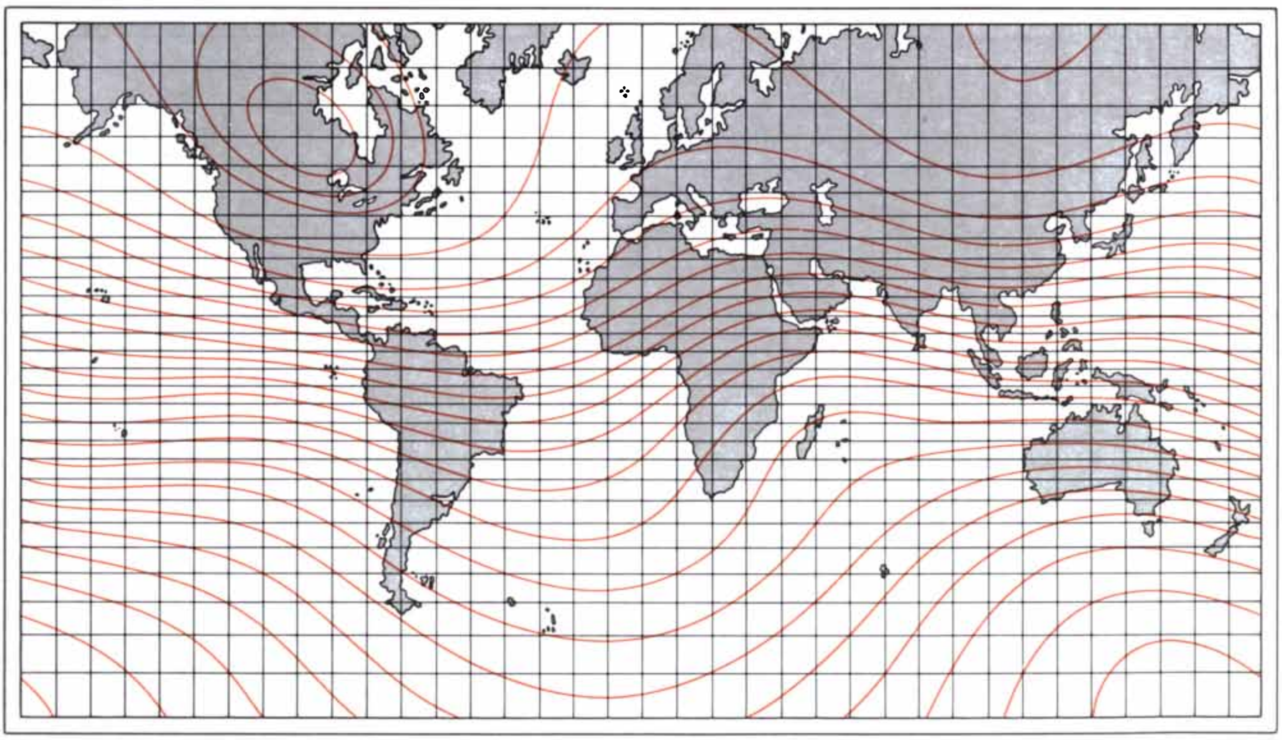
DISTANCES OF THE PLANETS from the sun are approximated by Bode's law, originally discovered by Johann Titius in 1766. (One unit equals 9.3 million miles.) The gap in the series known in 1800 precipitated the search for the "missing planet," culminating in the discovery of the asteroid Ceres in 1801. Gauss carried out the difficult task of computing the orbit of Ceres from the scanty data available. It reappeared where he predicted. (Planets discovered after 1800 are in color. Note that Bode's law poorly approximates the locations of the planets beyond Uranus.)



INTRINSIC SURFACE THEORY developed by Gauss enables one to compute the curvature of a surface solely by measuring the length of the curves that lie on that surface. The curvature of a convex surface (*a*) is found as follows: First, using a plane parallel to the tangent plane at point *P*, slice the surface near *P* along an ellipse (*dark color*). Draw the major and minor axes of the ellipse, designated *AB* and *CD*, and project these axes onto the surface to obtain the curves *APB* and *CPD*. Next find two circles having the closest fit at point *P* to the projected curves. If the radii of the two circles are *R* and *r* respectively, then as the plane approaches *P* and the ellipse gets very small the curvature will approach $1/Rr$. In *b* the surface is saddle-shaped and the plane cuts it at two hyperbolas. By convention one radius (and hence the curvature) is negative. For ordinary flat space the two radii *R* and *r* are infinite in length and the curvature is zero.

perimental and theoretical investigations of magnetism. They invented a magnetometer, and as a result of their interest in the magnetism of the earth they organized a network of observers across Europe to measure variations in the terrestrial magnetic field. Gauss was able to show by a theoretical analysis that the field arose within the solid earth, a result of considerable importance because it limited the field's possible origins and focused attention on geophysical mechanisms of its generation. His contribution is recognized in the gauss, the unit of the density of magnetic flux.

Gauss and Weber were also among the first to point out the possibility of sending messages by electricity. The history of telegraphy is long, but before 1800 or so only nonelectric methods were in use. Then in 1827 an electric



MAGNETIC-FIELD MAP of the earth is based on a drawing that appeared in 1840 in *Atlas des Erdmagnetismus*, written jointly by Gauss and Weber. The two men obtained the experimental measurements for the map by organizing a worldwide network of observers.

impulse was transmitted over a distance of a sixth of a mile, which immediately suggested that electricity could serve for telegraphy. Various electric telegraphs were designed, but none was implemented until 1832, when the Czar's summer and winter palaces in St. Petersburg were linked. A year later Gauss and Weber had a telegraph running over the rooftops of Göttingen for a distance of 2.3 kilometers. The signals were transmitted as a sequence of five deflections of a needle either to the right or to the left (a total of 32 possibilities), and the device worked so well that the two men used it routinely to communicate. The Gauss-Weber telegraph was probably the first to work in any useful sense, and it antedated Samuel F. B. Morse's famous patent by seven years.

Gauss's great reputation was still further enhanced after his death by the discovery of unpublished results anticipating many of the major advances of the 19th century. In addition to Cauchy's theorem he had discovered the double periodicity of elliptic functions, which in the hands of Jacobi and Abel became the center of 19th-century function theory. The elliptic functions are certain special functions $f(z)$ of a complex number z. Double periodicity means that there are two distinct complex constants, say a and b, such that for any z, $f(z)$ equals $f(z + a)$, which also equals $f(z + b)$. This is analogous to the single periodicity of the trigonometric functions, so that $\sin(z)$ has the property $\sin(z + 2\pi) = \sin(z)$. Gauss's discovery had wide implications for the connections between number theory and complex-function theory.

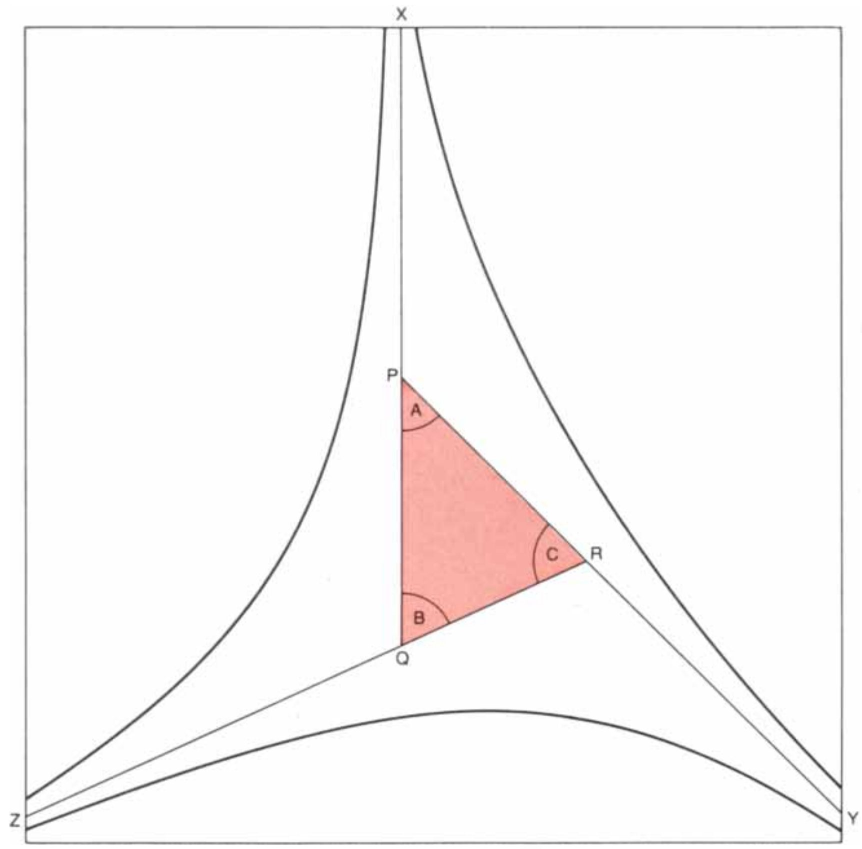
Gauss was also among the first to doubt that Euclidean geometry was inherent in nature and human thought. Euclid's systematic geometry had been based on certain axioms, or fundamental propositions, that were regarded as self-evident truths. On these unproved foundations the entire system was constructed through pure logic. The parallel axiom states that only one line can be drawn parallel to a given line through any point not on that line. Throughout history the parallel axiom was intensively investigated because, unlike the other axioms, it was not immediately persuasive and required the concept of infinity. Even in antiquity mathematicians tried in vain to replace it with more obvious axioms.

In the 18th century there was a renewed interest in this unsolved problem, and many mathematicians and amateurs attempted to prove that the parallel axiom followed logically from Euclid's other axioms. All these proofs turned out to be fallacious. Gauss became aware of the controversy as a student at Göttingen, and in 1804 he wrote a letter to the Hungarian mathematician Farkas Bolyai, showing that Bolyai's proof of the parallel axiom was false because he had replaced an infinite argument by a finite one. Gauss accompanied his refutation with a comment to the effect that he had been troubled by the same difficulty. In 1815, however, he wrote certain book reviews in which he clearly implied that a geometry might exist where the parallel axiom did not hold, yet which would still be internally consistent and free of contradiction. Given Gauss's usual caution about revealing his ideas, this statement strongly suggests that he had some proof to back it up. Perhaps because his ideas ran counter to contemporary views he chose to suppress them; he may have felt, probably correctly, that they would be misunderstood.

In 1820 Bolyai's son János, who had been infected by his father's fanatical preoccupation with proving the parallel axiom, concluded that a proof was impossible and began to develop a new geometry that did not depend on Euclid's axiom. Three years later he completed a paper proposing a consistent system of non-Euclidean geometry, which was published as an appendix to his father's book *Essay on the Elements of Mathematics for Studious Youths*. Gauss read that paper in 1832 and wrote the elder Bolyai that he could not praise it because doing so would be tantamount to praising work he had done himself 30 years before. The younger Bolyai was profoundly disappointed by Gauss's rejection, and he died essentially unrecognized for his solution to an enormously important and long-standing problem (solved independently at about the same time and in much the same way by Nikolai Ivanovich Lobachevsky). Gauss's attitude was quite unfair in view of the fact that he had never felt confident enough in his own work to publicize it. Perhaps he was a bit jealous of Bolyai's success.

In many ways Gauss stood at a crossroads. He can be viewed equally well as either the first of the modern mathematicians or the last of the great classical ones. The paradox can easily be resolved: his methods were modern in spirit but his choice of problems was classical.

A trademark of Gauss's work, particularly in pure mathematics, is his reasoning with the particular as if it were the general. The success of this technique depends on using only those prop-



HYPERBOLIC GEOMETRY is a non-Euclidean system devised by Gauss. It enables one to find the area of a triangle from its angles, a result that fails in Euclidean geometry. The diagram shown here is part of Gauss's proof that the area of the triangle is proportional to the difference between 180 degrees and the sum of its angles *A*, *B* and *C*. Because the diagram is drawn on a Euclidean plane the lines appear curved, but in hyperbolic space they are "straight." Moreover, the lines occurring in triplets at the edges of the diagram are all parallel, so that the apparently curved line *XY* is parallel to both *XQ* and *PY*, a situation that cannot occur in Euclidean geometry. Although the triangle *XYZ* has all its vertexes at infinity, its area is finite.

erties of a special case that have general counterparts. Gauss's approach combines the breadth of general methods with the intensity and simplicity of particular examples. Thus his work on complex numbers has within it the seeds of a general theory of algebraic numbers. His writings always leave one with the feeling that he knows more than he is telling, that when he explains a result, he already has an idea of more general questions surrounding it and of how to make a start on solving them.

A measure of the depth and fertility of Gauss's ideas can be obtained from recent investigations inspired by them. For example, in 1947 André Weil, starting from some theorems of Gauss's on the number of solutions to polynomial equations modulo a prime number, was led to formulate three far-reaching conjectures about algebraic varieties over finite fields. A finite field is a set of algebraic elements, finite in number, that, in addition to meeting certain other requirements, can be added, multiplied, subtracted and divided to give quantities of the same type. Thus the integers modulo a prime number p form such a field with p elements.

The Weil conjectures give formulas for the number of solutions to an algebraic equation in a finite field. In particular they allow one to deduce that a given equation does or does not have solutions; this information can be transferred to analogous equations involving integers or algebraic numbers. Of course, the actual formulation of the Weil conjectures is very technical. They have recently been proved by Pierre Deligne, and they are of fundamental importance in algebraic geometry.

Elsewhere Gauss conjectured that prime factorization was unique for numbers of the form $p + q\sqrt{-D}$, where D is a positive integer, only when D is equal to one of the numbers 1, 2, 3, 7, 11, 19, 43, 67 and 163. This result, which he inferred from numerical evidence, has recently been proved by Harold M. Stark and Alan Baker and leads to important new theorems in number theory.

Gauss's importance stems from this mixture of the general and the specific. He forms a bridge between old and new, and his ideas contain within them the seeds of spacious theories and important results. He is the prime instance of a mathematician who could extract the maximum amount of juice from a ripe example by inductive reasoning (working from a particular case to a general principle) rather than deductive reasoning (drawing a specific conclusion from general principles). Of Gauss's influence on his successors Eric Temple Bell wrote in 1937: "He lives everywhere in mathematics." If anything, that is truer today than it was when Bell said it 40 years ago.

MATHEMATICAL GAMES

Cutting things into equal parts leads into significant areas of mathematics

by Martin Gardner

"So the Wizard lost no more time, but leaping forward he raised the sharp sword, whirled it once or twice around his head, and then gave a mighty stroke that cut the body of the sorcerer exactly in two.

Dorothy screamed. . . ."

—L. FRANK BAUM,
Dorothy and the Wizard in Oz

A popular type of puzzle, often found in old puzzle books, is to divide a given shape into two, three or more equal parts. Sometimes "equal" means congruent; sometimes it means no more than equal in area. Longtime readers of this department may remember many problems of this type: enumerating the ways a chessboard can be cut along lattice lines into two or four congruent pieces, bisecting the yin-yang symbol (with one straight line) into four parts of equal area, dividing a square cake into n slices of equal volume, "rep-tiles" that can be cut into congruent copies of themselves and numerous others.

This month we look at a variety of problems of dividing shapes into equal parts that have not been discussed before. Some of them lead into significant areas of modern mathematics.

Let us begin with the simplest of the problems: cutting a plane shape into two congruent parts. Pieces are considered to be congruent if they are mirror images. You might think that all such problems would be easy to solve, but they can be annoyingly difficult. As far as I know there is no algorithm for deciding in general whether a shape can be divided into two or more congruent parts, and interesting theorems about such divisions are curiously scarce.

The reader is invited to try his skill at cutting each of the 12 shapes in the illustration on the opposite page into two identical pieces. There are no catches. Each half is connected simply (there are no holes or parts joined at single points). Some of these figures are from a French magazine column on recreational mathematics by Pierre Berloquin. (Scribner's has recently published translations of three of Berloquin's popular puzzle books.) The shape with the hole is difficult and suggests an unexplored region of puzzledom: shapes with holes that are to be cut into congruent pieces.

Note that the only kind of symmetry possessed by the fourth shape is bilateral, the axis of symmetry passing vertically down the middle. Any plane or solid figure with bilateral symmetry obviously can be sliced into congruent halves by cutting along its axis or plane of symmetry. The illustration on page 134 shows how the Wizard of Oz bisected an evil Mangaboo into congruent parts while Eureka, Dorothy's pet cat, watched. (The scene occurs in the Glass City, below the earth's surface, where the inhabitants are vegetables.) Is it always necessary to cut along an axis or a plane of symmetry to divide a bilaterally symmetric figure into congruent parts? The answer is no, and the fourth shape proves it. You are asked to solve this figure in a way other than the obvious one.

Dividing a shape into n congruent parts usually becomes more difficult as n increases, particularly if the shape of the part is specified. The 12 pentominoes in the top illustration on page 135 pose an interesting quartering problem. (I am obliged to mention that "pentominoes" was registered as a trademark in 1975 by Solomon W. Golomb, who coined the term.) How many of the pentominoes can be dissected into four congruent pieces? All except three. I was surprised to discover that the nine with solutions can all be solved by using the same component shape—a smaller pentomino—and that this shape is unique. Can the reader discover it and identify the three impossible pentominoes?

Let us drop the requirement of congruence, demanding only that the n parts have equal area. It seems intuitively clear that any plane shape can be cut into halves of equal area by a straight line. It is perhaps not so immediately obvious that the line can be parallel to any given line outside the figure. To prove this relation we make use of a famous theorem: If there is a continuous one-variable function in a closed interval from A to B, then the function has a minimum value and a maximum value and all real values in between.

Let me give an example. You are walking up a mountain along a crooked path from A to B. Your altitude at any moment is a continuous function of your position on the path. The theorem tells us that on the path there is at least one point (there may be more than one if the path goes up and down) of minimum altitude, at least one point of maximum altitude and at least one point for every real value in between. The theorem seems obvious to the point of triviality, and yet it has fantastic power in proving theorems that are not at all obvious.

Consider the shaded region at the left in the middle illustration on page 135. Outside it is an arbitrary line x. We want to prove that a line parallel to x can be drawn through the figure that will exactly bisect its area. Imagine a line moving slowly, always parallel to x, in the perpendicular direction indicated by the broken arrow. The line touches the region at A and leaves it at B. As the line moves over the region the area below it is a continuous function of the distance it has moved from A. The area of side A is zero at A and maximum at B. According to our theorem, somewhere in between is a point where the area is just half of the maximum. That is the point at which the line bisects the area.

The proof is so general that it applies not only to any connected shape, including one with holes, but also to disconnected regions. A glance at the right-hand part of the middle illustration on page 135 should convince you that a line parallel to a given line can be drawn through any number of regions in such a way that the total area on one side of the line equals the total area on the other side.

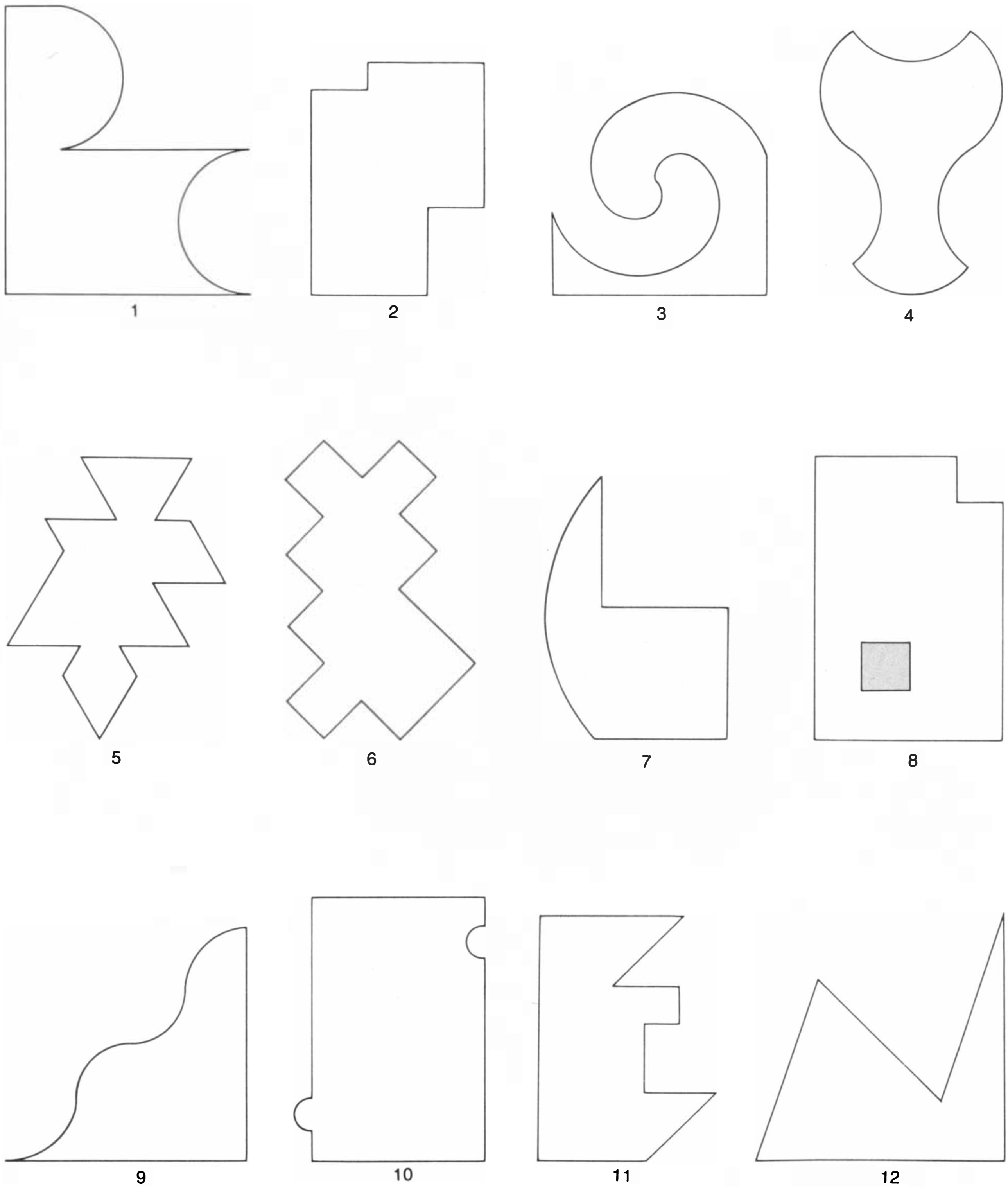
Suppose you have two regions of any kind in the plane, as is shown in the bottom illustration on page 135. Is it always possible to draw a straight line that will simultaneously divide each region into parts of equal area? We can prove that it is. First bisect the white region with a line that misses the gray one. Rotate the line, always preserving the bisection of the white region. (We know we can do this from the above theorem.) As the illustration shows, the turning line will touch the gray region at point A and leave it at point B. As the line sweeps from A to B the area on side A will vary continuously from zero to the maximum. Hence there is a spot in between at which the area on side A is half of the total area.

We have proved for the plane a famous theorem that generalizes to all higher spaces. In 3-space the volumes of any three solids can be bisected by a plane; in 4-space the hypervolumes of any four figures can be bisected by a hyperplane of $N-1$ dimensions. The 3-

space theorem is sometimes called the "ham-sandwich theorem" because it applies to a generalized ham sandwich consisting of two pieces of bread and one piece of ham. No matter how the pieces are shaped or how they are placed in space, there is a plane that simultaneously halves all three.

The generalization to *N*-space requires high-powered mathematics, but in 2- and 3-space the proofs follow easily from our fundamental theorem. Both proofs provide in a sense a way of finding a bisecting line or plane, but in practice it is not easy to find such a bisection. Connecting centroids (centers of gravity) will not do it because a straight line (or a plane) through the centroid of a plane (or a solid) figure does not necessarily bisect it into halves of equal area (or volume).

Given a simply connected figure, not necessarily convex, is there always a straight line that simultaneously bisects both the area and the boundary? There is, and the proof is much like the last one. Draw a line that bisects the bound-



Shapes to be cut into congruent halves ary at P and Q, as is shown in the top illustration on page 136. If this line, extended, also bisects the area, there is no more to do.

Assume that the line does not bisect the area. Attach to the line an arrow that points into the side of smaller area. Now move points Q and P clockwise around the boundary, always preserving the boundary's bisection. This maneuver will cause the line going through P and Q to alter its orientation in a continuous way. As it does so the ratio of the areas on each side will vary continuously. After the line has rotated 180 degrees it will coincide with its original position, but now the arrow will point into the larger of the two regions.

Consider the difference obtained by subtracting the area on the arrow side of the line from the other area. At the start it is positive. At the finish it is negative. This value clearly is a continuous function of the line's angle, and so there must be an angle at which the value is zero and the two areas are equal. The theorem also generalizes to higher spaces. A solid can be bisected in volume by a plane that bisects the surface area, and in general an N-space solid can be bisected in hypervolume by an $N - 1$ hyperplane that bisects its hypersurface.

Another beautiful theorem, equally far from obvious, states that any region—again it need not be connected simply or even connected—can always be exactly quartered by two perpendicular lines. The standard proof is both subtle and marvelous.

Imagine you have a sheet of transparent paper larger than the region. The sheet is divided into four quadrants by perpendicular lines X and Y. The quadrant at the upper left is gray and the one at the upper right is in color. Place the sheet over the region with the horizontal line below the region and the vertical line to the right. Slide the sheet up until line X bisects the region, then slide the sheet to the left (letting X slide along itself) until Y likewise bisects the region. Label (on the region, not on the transparent sheet) the four parts of the figure A, B, C and D. The arrangement is shown in the top part of the bottom illustration on page 136.

By construction A plus B equals C plus D and A plus C equals B plus D. Subtracting the second equation from the first one results in $B - C = C - B$, or $2B = 2C$. Therefore B equals C and A equals D.

If A equals B, the region is quartered. Assume that it is not quartered and that region B is larger than A. Therefore the area of the region's gray part, subtracted from the region's colored part, is a positive number.

Rotate the sheet 90 degrees counterclockwise, always maintaining the double bisection of the region by X and Y. Their intersection point may wander here and there. After the rotation is completed the two lines necessarily return to their former position, as is shown in the bottom part of the illustration, except that now X and Y have exchanged places.

The colored area of the region equals A. The gray area equals C, and because C equals B, the gray area also equals B. Thus the areas of the gray and colored regions have also exchanged places. Consequently if we subtract the larger gray area from the smaller colored area, we now get a minus number.

It is easy to see how our fundamental theorem applies. The value obtained by taking the area of the gray region from the area of the colored region is a continuous function of the angle of rotation as it varies from 0 to 90 degrees. Since the former value goes from plus to minus, there must be a value in between at which the difference is zero and the areas are equal. When that value is attained, the perpendicular lines precisely quarter the region.

This theorem also generalizes to higher spaces. Any solid can be divided into eight equal parts by three mutually perpendicular planes. In general any N-dimensional solid can be cut into 2^n parts of equal "volume" by n perpendicular "planes."

As in the preceding examples, the theorem is not much help in the practical problem of quartering nonsymmetric figures. We know, for instance, that a triangle of sides 3,4,5 can be quartered by two perpendicular lines, but the task of constructing those lines is an altogether different matter. Indeed, I know of no easy way to do it.

The illustration on page 137 presents four equal-division problems for which the constructions are much less difficult than quartering the 3,4,5 triangle is. They are nonetheless tricky enough to call for a considerable amount of ingenuity.

The first problem is to divide the entire region of nine squares into two parts of equal area with a straight line that goes through corner P.

The second problem is to halve the total area of the five circles by drawing a

The Wizard of Oz bisects a Mangaboo

straight line through *P*, which is the point at the center of the circle farthest to the left.

The third problem is to trisect the area of a regular hexagon with two straight lines that go through *P*.

The fourth problem is to bisect an equilateral triangle with a curve of minimum length. The illustration of the triangle shows the shortest bisecting straight line, but there is a bisecting curve that is shorter.

All four problems will be answered next month, along with the preceding ones.

February's column on wordplay brought such a flood of remarkable letters that I can touch on only a few highlights. Dozens of readers sent poems, each line of which was an anagram of ETAOIN SHRDLU. (One of the best, by Walter G. Leight, appeared in the letters department of the April issue.) W. E. Bullard and Weston Hare each found a single-word anagram, OUTLANDISHER (a variant of outlander), that actually appears in a dictionary. Dolores Kozielski came close with ANTISHOULDER. John S. Rigney wrote that if a man in Spain wanted to get to SOUTH IRELAND, he would SAIL DUE NORTH. T. J. Doyle pointed out that Etaoin Shrdlu is a character in Elmer Rice's play *The Adding Machine*.

Many readers scrambled Oscar Wilde's four lines in other ways (see the April letters department), some including the two extra lines of Wilde's original stanza. Several readers suggested new titles for Mary Youngquist's poem that would make it possible for the last letter of her name to join the first letter of the title.

Carolyn Weyant sent her favorite perverb:

In one ear and gone tomorrow.

Leonard Morgenstern recalled:

Many are cold but few are frozen.

One man's Mede is another man's Persian.

A bun is the lowest form of wheat.

John F. Gummere sent:

Virtue is its only reward.

Honor is without profit in your own country.

All jack and no work makes a dull play boy.

Many hands make light work harder.

Too many crooks spoil the brothel.

Del Ratzsch praised Walt Kelly, the author of the comic strip *Pogo*, as the master of this kind of thing, citing as being among his best:

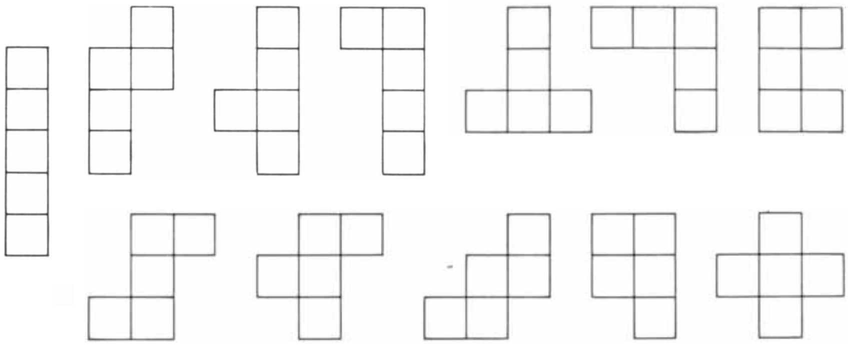
Now is the time for all good men to come to.

A bird in the hand is worth two in the bush leagues.

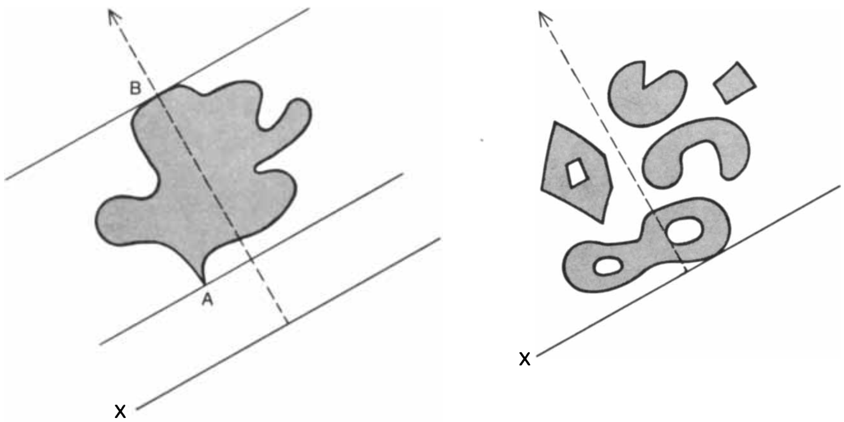
Romeo wasn't bilked in a day.

Pride goeth over the falls.

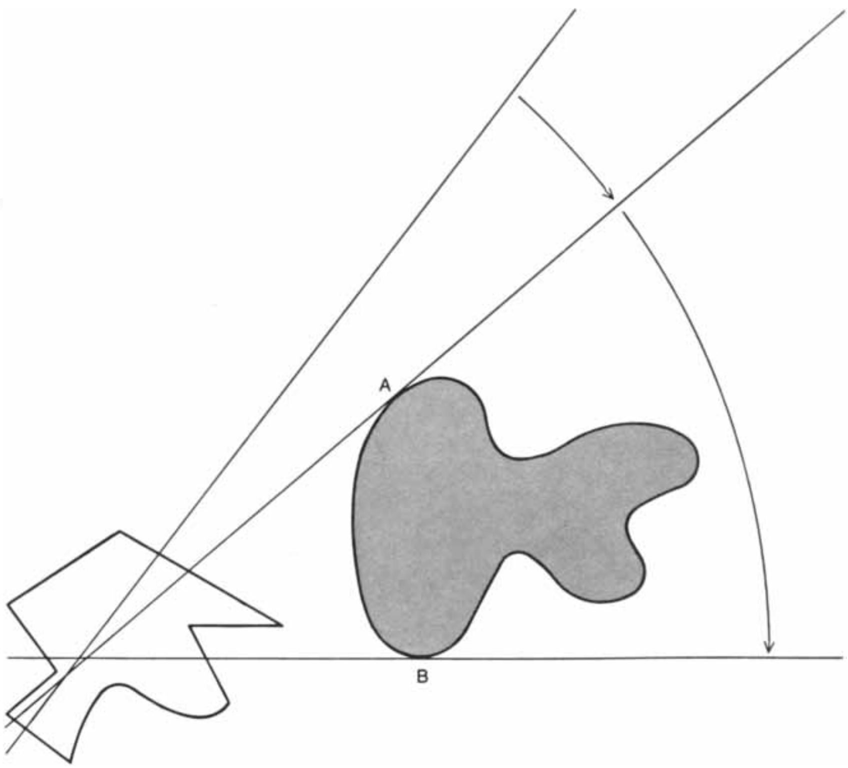
Albert L. Ely, Jr., interwove the quatrains by Elinor Wylie and Edna St. Vincent Millay with A. E. Housman's "A



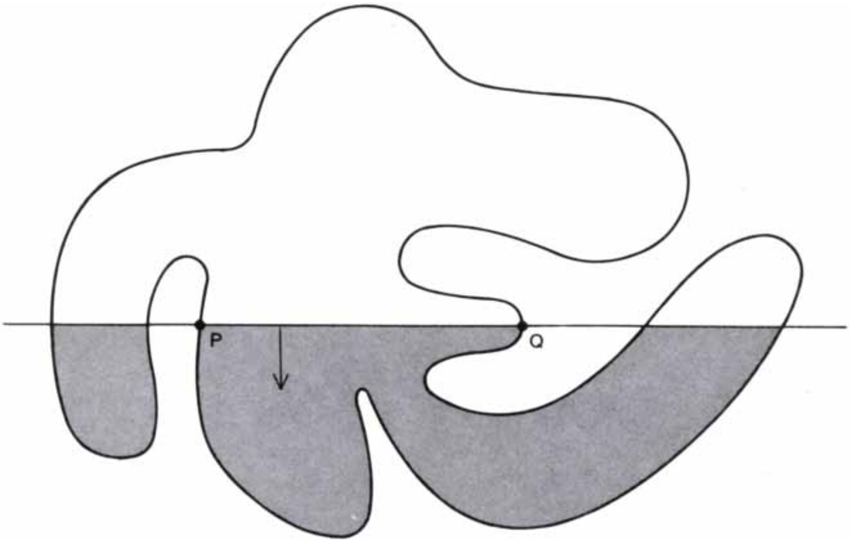
The 12 polyominoes of order 5



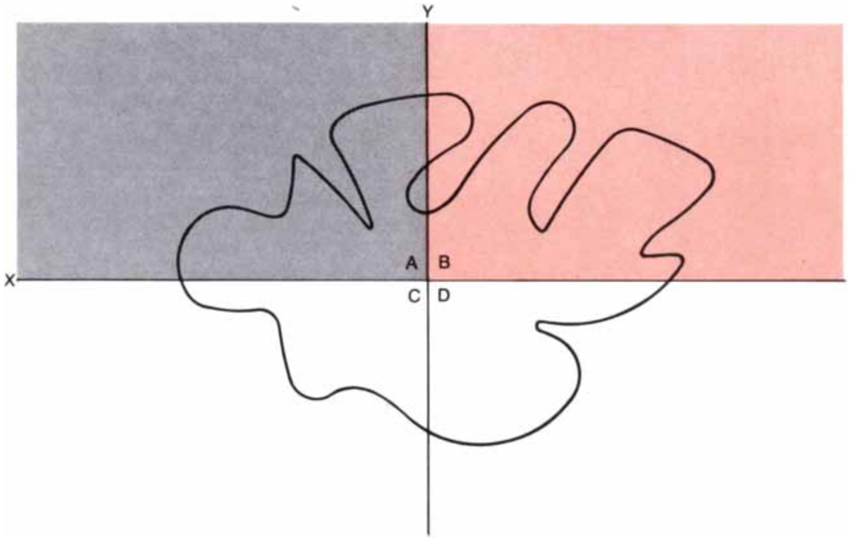
Proof of a theorem on bisecting an area

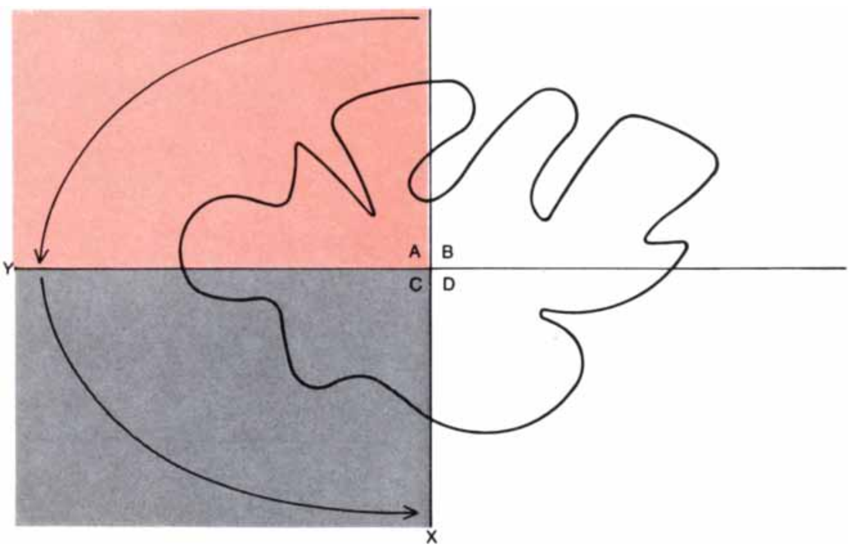


Proving the ham-sandwich theorem in the plane



Proof of a theorem on bisecting a boundary and area





How any region can be quartered by two perpendicular lines

Shropshire Lad" to obtain the following stanzas:

With rue my heart is laden—
My candle burns at both ends
For golden friends I had.
It will not last the night
For many a rose-lipt maiden,
But ah, my foes, and oh, my friends,
And many a lightfoot lad,
It gives a lovely light.

By brooks too broad for leaping
Sleep falls with limpid drops of rain.
The lightfoot boys are laid
Upon the steep cliffs of the town;
The rose-lipt girls are sleeping;
Sleep falls; men are at peace again
In fields where roses fade
While the small drops fall softly down.

Dick Ringler, inspired by J. A. Lindon's palindromic dialogue between Adam and Eve, composed the following pair of quatrains, each a word reversal of the other:

DIPTYCH

To Theseus: Finding No Minotaur

Thread the chaos, pattern the
despair.
Shadows loom and worry you:
Dead hope, and empty heaven, and
now bare
Meadows—fearful! but all
perspective true.

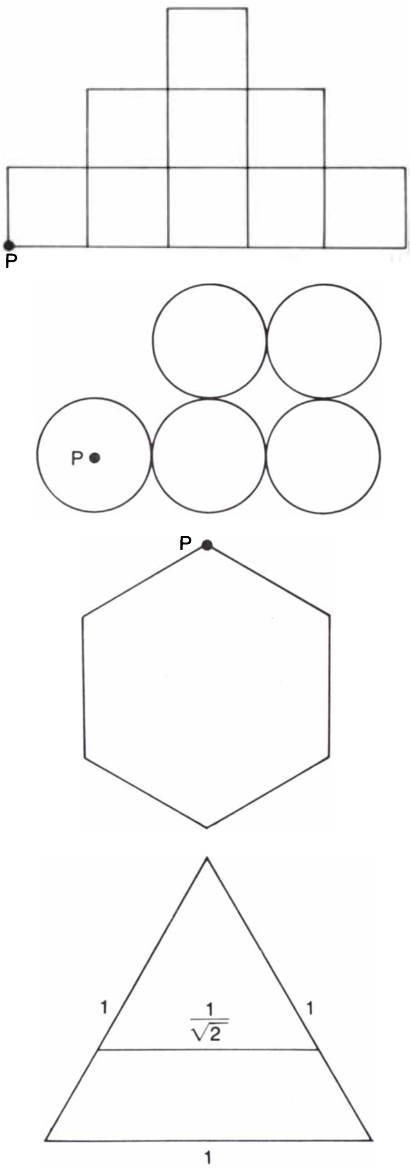
To Penelope: Weaving in Autumn

True perspective all, but fearful!
meadows
Bare now, and heaven empty, and
hope dead,
You worry and loom shadows,
Despair the pattern, chaos the
thread.

Last month, answering a problem about colored hats, I presented a variant: There are n men, $n - 1$ black hats and one red hat. The men are "progressively blind" because they are seated in a row so that each man sees only the hats in front of him. *A*, at the rear, sees all the other hats, *B* sees all but *A*'s hat and so on to the front man, who sees no hats. The n hats are put on the men's heads in any order, and the men are asked (in the sequence *A, B, C*...) if they know the color of their hat. Will the man with the red hat always be able to say yes?

The solution is curious. If *A* has the red hat, he will of course say yes because he sees $n - 1$ black hats. If *B* has the red hat, he will say no. Why? Because *A* must always answer yes, therefore his yes gives no information. If *C* has the red hat, he will say yes because he hears a yes from both *A* and *B*. If *D* has the red hat, he must say no because *C*'s yes could result from a deduction that he, *C*, has the red hat. This process generalizes by mathematical induction to the conclusion that the man with the red hat says yes if and only if he is in an odd position from the back. If he is in an even position, the yes of the man directly behind him fails to give sufficient information for a valid deduction.

Last December's discussion of "monster" curves, such as snowflakes and flowsnakes and space-filling Peano curves, introduced Benoît B. Mandelbrot's term "fractal" for curves that have a Hausdorff dimension greater than their topological space dimension. *Fractals: Form, Chance and Dimension,* in which Mandelbrot writes about these curves and their many applications in science, has just been published by W. H. Freeman and Company. It is a beautiful book, lavishly illustrated with fantastic computer graphics.



Four problems in equal division

THE AMATEUR SCIENTIST

How to create and observe a dozen rainbows in a single drop of water

by Jearl Walker

The rainbow, at once grand and delicate, appeared in the sky as I sat swaying with my grandmother on her porch glider. Aledo, Tex., is a small town, and there is not much to do there but swing on my grandmother's front porch. As the arc of color took form she turned with a sly look to her grandson, the physicist, and asked me why the rainbow colors appear only in that one arc. Why is the entire sky not filled with colors?

I offered the conventional explanation, comparing the separation of colors in the rainbow to the dispersion of white light by a prism. A particular geometry of scattering, I explained, is needed in order for each of the colors to reach our eyes. Hence the colors are confined to a single arc, a fixed set of directions.

Feeling satisfied with my explanation, and feeling rather smart, I turned back to the sky just as a second rainbow appeared, somewhat higher than the first. One look at my grandmother and I knew I was in trouble. Almost in rhythm with our swinging, she asked question after question. Why is the higher bow wider? Why is the sequence of colors reversed in the higher bow? Why is the space between the bows darker than the surrounding sky? What are the faint, narrow arcs just below the first rainbow and just above the second one? And, again, if there can be two rainbows, why not more? Why is the entire sky not covered with rainbows?

Orange ray from the first-order rainbow

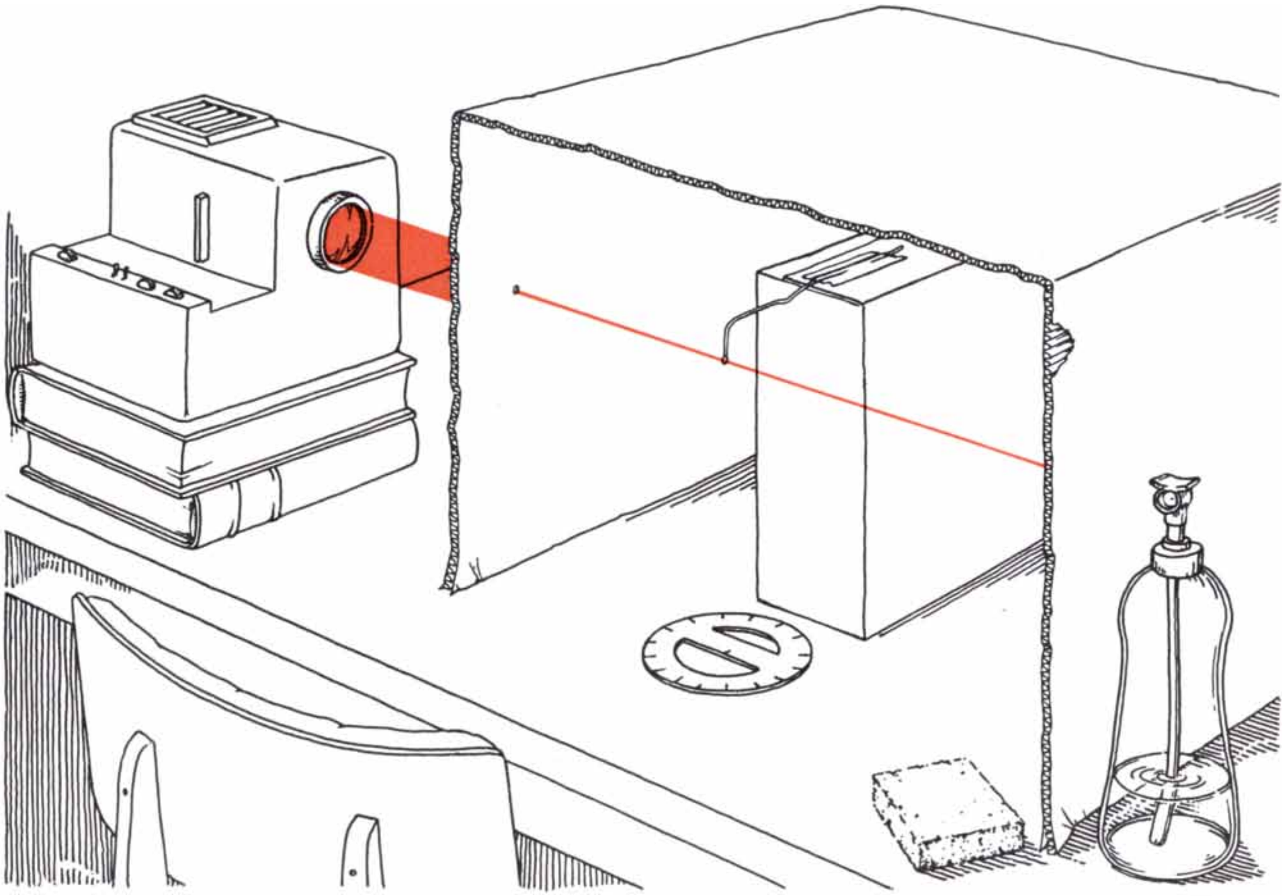
When my grandmother starts asking me how the world works, it often seems that my academic training comes out second best to her intuition. The question of why the sky is not covered with rainbows is a good example. In fact, the sky should be filled with rainbows, but reports of more than two are rare. The primary rainbow, the one most often seen, is formed by light rays that are reflected once inside raindrops and then scattered toward the observer. The occasional secondary rainbow requires two internal reflections of the rays. Additional internal reflections should give rise to higher-order rainbows, but apparently they are too faint compared with the background glare to be readily distinguished. They are there in the sky, hanging with the same delicate colors as the familiar rainbows, but they are not ordinarily seen.

The theory of the rainbow was discussed by H. Moysés Nussenzveig of the University of São Paulo in the April issue of *Scientific American.* He described attempts made since the time of Newton to explain the rainbow's appearance, culminating in recent mathematical theories. Here I shall be concerned mainly with the observation of rainbows, and in particular with an experiment in which many of the higher-order rainbows can be observed. With care, all the rainbows through the 13th should be observable. They appear not in the sky but in a single drop of water.

Only a handful of people have seen the higher-order rainbows. The French investigator Felix Billet was one of them; in 1868 he reported seeing 19 rainbows in a thin stream of falling water. Instead of a stream we shall employ a droplet suspended from a thin wire. The experiment is simple enough to be done in the kitchen, and yet it enables one to join that small club of people who have seen the higher-order rainbows.

With an atomizer or a plant sprayer, gently spray water onto the end of a piece of wire that has been mounted vertically with its free end pointing down. Spray until a drop forms and hangs from the end of the wire. Obtaining such a drop may take practice and patience. It may help to coat the wire with black wax, which makes the water bead up and also reduces glare when the drop is later illuminated.

Once a drop has formed, cover the apparatus with a large box from which one side has been removed. Make a small hole in the box in a side adjacent to the open one and at the height of the water drop; the drop will be illuminated through the small hole and viewed from the open side of the box. The source of illumination can be a slide projector or a movie projector. It should be positioned



Apparatus for observing rainbows in a drop of water

so that the beam entering the box through the small hole strikes the drop horizontally and traverses a circular cross section. Ideally only the drop and not the wire will be lighted.

The main reason higher-order rainbows are not visible in the sky is the glare of sunlight directly transmitted or reflected from raindrops. In our experiment too glare tends to obscure the rainbows, but it is more easily controlled in the kitchen. Only the light falling on a part of the drop contributes to a given rainbow, and the bright glare spots often result from light striking another part. It is therefore possible to eliminate or at least reduce the glare by masking a part of the incident beam. That is most easily accomplished by adjusting an index card to block off part of the hole in the carton.

Center under the suspended droplet a full-circle protractor, with the zero-degree mark at the side of the drop opposite the incident beam. With the protractor aligned in this way, it measures the total angle through which the light rays are deflected in the drop. Rays passing straight through emerge at zero degrees; those reflected back into the light source emerge at 180 degrees. The first-order rainbow can be found between 137.6 degrees (red) and 139.4 degrees (blue), the colors appearing near the right-hand edge of the drop. The second-order bow comes into view at about 129 degrees, on the left side of the drop. Near the center in both views are glare spots.

The first-order rainbow is made up of rays reflected once inside the drop, but only some of those rays contribute; the rest emerge at the wrong angles. For the sake of simplicity suppose only half the drop is illuminated, so that incident rays are uniformly distributed from the center to one edge, where they graze it tangentially. All these rays are refracted as they enter the drop; they are then reflected by the interior surface, and finally they are refracted again as they emerge from the drop. The scattered light is spread over a large range of angles, the deviation of any particular ray being determined by where on the surface of the drop that ray first made contact. Rays incident at the very center pass through the drop, are reflected and are doubled back on their incident path, for a total deviation of 180 degrees. Rays incident farther from the center of the drop have smaller deviation angles, but only up to a point. As the incident ray nears the edge of the drop a minimum deviation is reached. The rainbow is simply the enhancement in the brightness of the scattered light at the angle where the largest number of rays emerge. That angle is the angle of minimum deviation.

The index of refraction of water, which determines how much the rays are bent on entering and leaving the drop, is different for each color, or wavelength, of light. (Blue light is refracted more than red light.) As a result each color has a different angle of least deviation, where the scattered light is most intense. The rainbow would still exist even if water did not disperse white light into its component colors, but it would be only a bright white band; dispersion creates a series of monochromatic arcs, each at a slightly different angle.

Even for the first-order rainbow from a single drop, the white glare spot near the rainbow colors is comparatively bright. The spot is made up of rays reflected from the outside surface of the drop. If you were more than about a meter from the drop, as you surely would be from raindrops forming a natural rainbow, the glare spot and the rainbow spot would be indistinguishable, and the effective brightness of the rainbow colors would be reduced. Natural higher-order rainbows are lost entirely in the glare (except, perhaps, under unusual circumstances).

In the kitchen you can move close enough to the drop to distinguish the separate bundles of glare and rainbow light; moreover, you can block some of the glare. Slowly move the vertical edge of an index card across the incident beam. If the card is inserted from the side of the beam opposite you, it soon blocks the incident rays that contribute to the rainbow spot. If it is started on the near side, then when its shadow is about a quarter of the way across the drop, it intercepts the rays that produce the glare spot. The selective blocking is easiest with a large drop, say a drop two millimeters in diameter. As the drop evaporates and shrinks, positioning the index card becomes difficult. The color separation is also clearest with a large drop and fades to white as the drop evaporates. Natural bows with washed-out colors are seen when sunlight is scattered by very small raindrops.

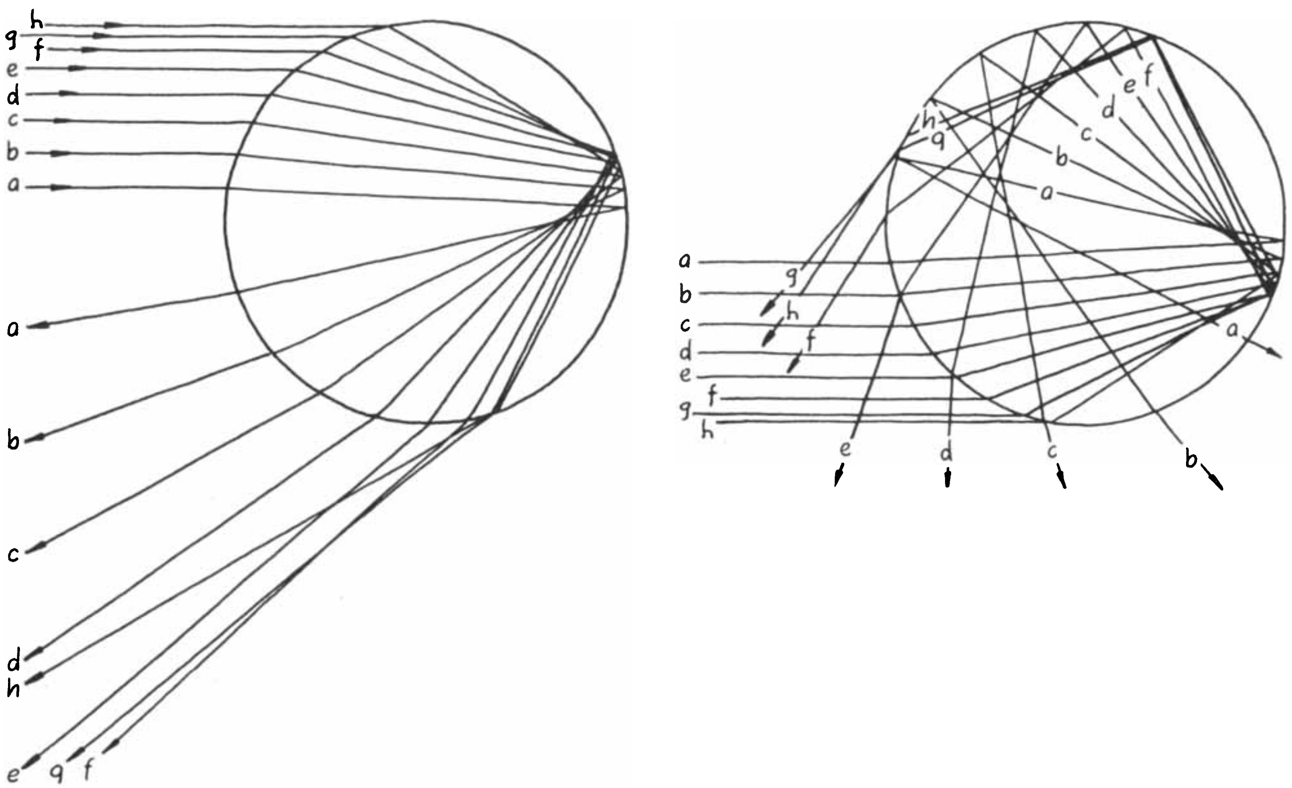
Rays reflected twice inside the drop, contributing to the second-order rainbow, are also brightest at the angle of least deviation; again, that angle is different for each color. If the second-order rainbow is observed on the same side of the water drop as the first-order one, then blue falls at 126.5 degrees and red at 129.6. As my grandmother pointed out, the color sequence is opposite to that of the first-order rainbow, and the second-order rainbow is wider. The angular width is greater because the rays giving rise to the second-order rainbow enter the drop closer to the edge and undergo greater dispersion.

Between the first- and second-order rainbows relatively little light is scattered from the drop. Rays reflected once inside the drop all emerge at angles equal to or greater than the angle of the first-order rainbow. Rays internally reflected twice emerge at angles equal to or less than the angle of the second-order bow. The angular region between the two rainbows is therefore dark. The only light emerging in that region is light that has been reflected more than twice and as a result is relatively dim. This dark area between the first two rainbows is called Alexander's dark band, after the Greek philosopher Alexander of Aphrodisias.

All the higher-order rainbows are formed by the same mechanism as the first two, but after additional internal reflections. Many of them are faint and are obscured by glare, but they can be found if you know where to look. By examining the drop at the angles calculated for the higher-order rainbows I was able to observe all of them through the 13th order. In several cases the incident rays contributing glare had to be blocked with an index card. The most difficult rainbows to see are the ones near the axis of the incident beam. If you must look toward the light source, as with the fourth, eighth and 12th rainbows, then you must contend not only with the glare reflected from the surface of the drop but also with light transmitted through the drop with no internal reflections. Both are white and relatively bright. When looking almost directly away from the light source, as in the case of the 10th-order rainbow, it is difficult to avoid blocking the incident beam with your head.

The 11th-order rainbow is surprisingly prominent, if part of the incident beam is blocked in order to eliminate the glare spot. The 12th and 13th rainbows require considerable patience in order to block just enough of the glare rays but not too many of the rainbow rays. If Billet saw the 19th-order rainbow in his falling-stream apparatus, either his patience or his eyesight was better than mine.

Colored light can appear on the suspended drop through one other phenomenon in addition to the rainbow. Some of the incident light is nearly tangent to the drop and enters almost parallel to the surface. For these rays, as for all others, blue light is refracted slightly more than red light. If the corresponding emerging rays are observed at the appropriate angle, only the blue light will be visible; the drop will have a blue edge. The correct angle is one slightly too large for the red rays to reach but within the range of the more strongly refracted blue light. The blue edges are too faint to be seen naturally because



***Ray paths in primary rainbow* (left) *and secondary rainbow* (right)**

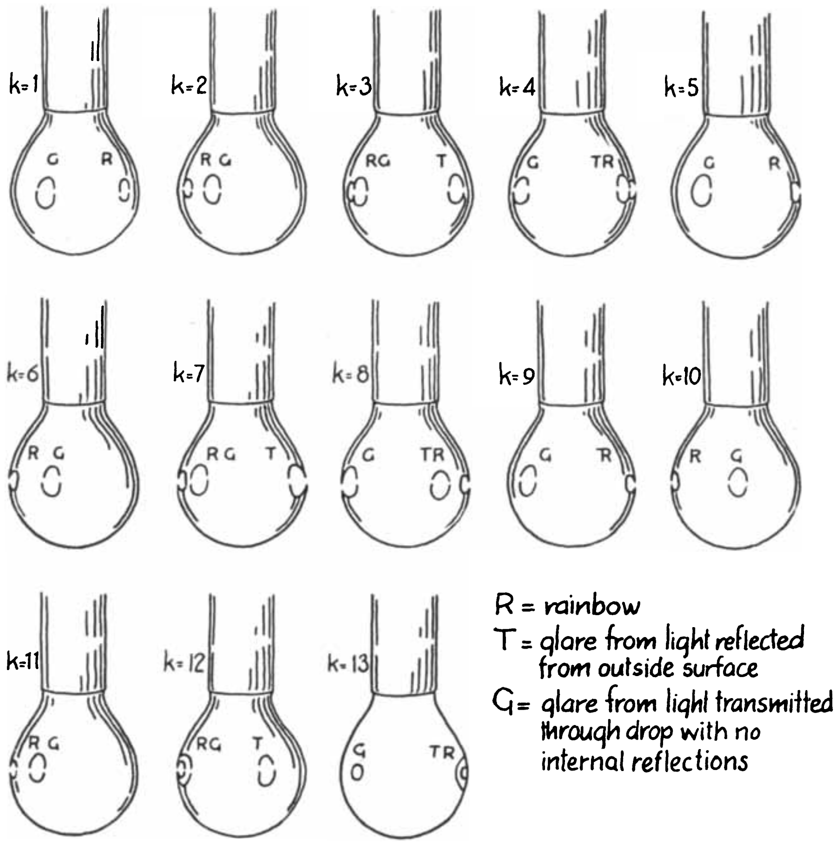
little of the incident light enters the raindrops tangentially. Each of the rainbows has its own associated blue-edge angle, but beyond the first two rainbows the edges are difficult to find. The first-order blue edge appears on the right side of the drop at a scattering angle of 165 degrees.

At some angles near rainbow angles white spots appear that are not glare. They are internally reflected rays that emerge at angles other than the minimum-deviation angle. They are white because they are composed of approximately equal intensities of the colors. Two such spots can be seen on the drop from an angle of 145 degrees. Both result from light reflected once inside the drop, but with different points of incidence, one on each side of the point that gives rise to the primary rainbow. As you move your view from an angle of 145 degrees toward the rainbow angle of 138 degrees the two white spots move toward each other. At the rainbow angle they merge and take on color.

The natural rainbow is polarized parallel to its arc. You can determine the sense of polarization of the rainbow rays emerging from the suspended water drop by viewing them through a polarizing filter, such as one eyepiece of polarizing sunglasses (which are designed to exclude light with a horizontal plane of polarization). The incident light can be regarded as an equal mixture of two polarizations: one in the plane that includes the light source, the drop and the observer, the other perpendicular to that plane. More light with the second polarization is transmitted through the drop without reflection. The emerging rainbow is therefore polarized perpendicular to the defined plane and parallel to the rainbow arc.

The angular positions of the rainbows depend on the index of refraction of water, which is about 1.33 (at a wavelength of 5,890 angstroms, the wavelength of the yellow light from a sodium-vapor lamp). I searched my grandmother's kitchen for some other transparent liquid with a different index of refraction. The most convenient one I found was light corn syrup. By carefully applying small amounts of the syrup to the wire mount you can build up a suspended drop. You must work quickly, however, or the drop will harden in a distorted shape. Fresh corn syrup has an index of refraction of 1.47 or 1.48. As a result the angular positions of all the rainbows are shifted clockwise from their positions in water, although not by a uniform amount. As the drop dries over a period of several days and the syrup becomes more concentrated the refractive index increases. The rainbows shift slowly around the drop.

I have also seen rainbows formed by drops of diiodomethane, a transparent fluid with the relatively high index of

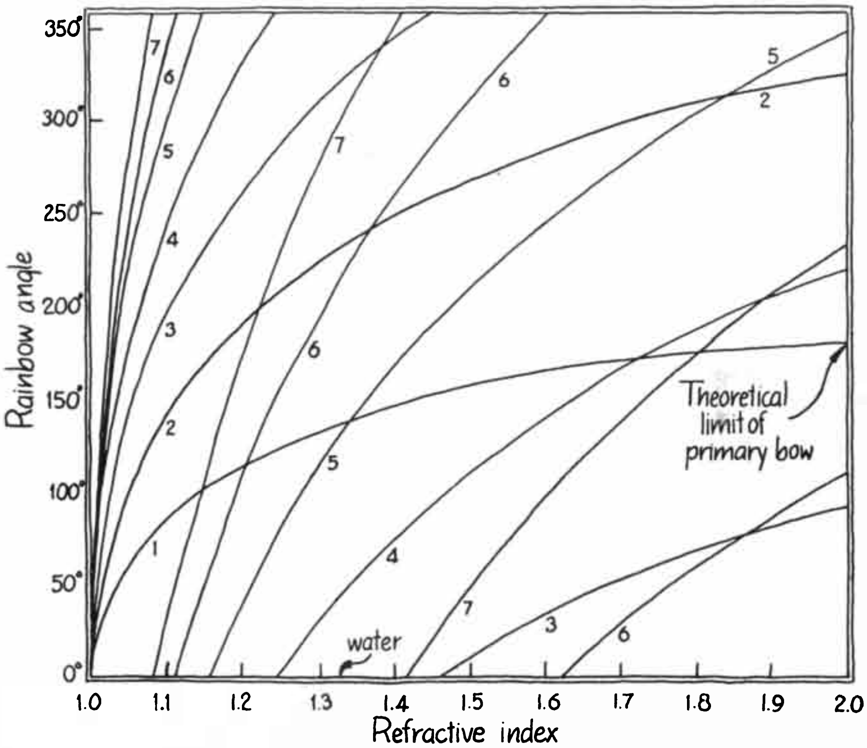


Appearance of the first 13 rainbows

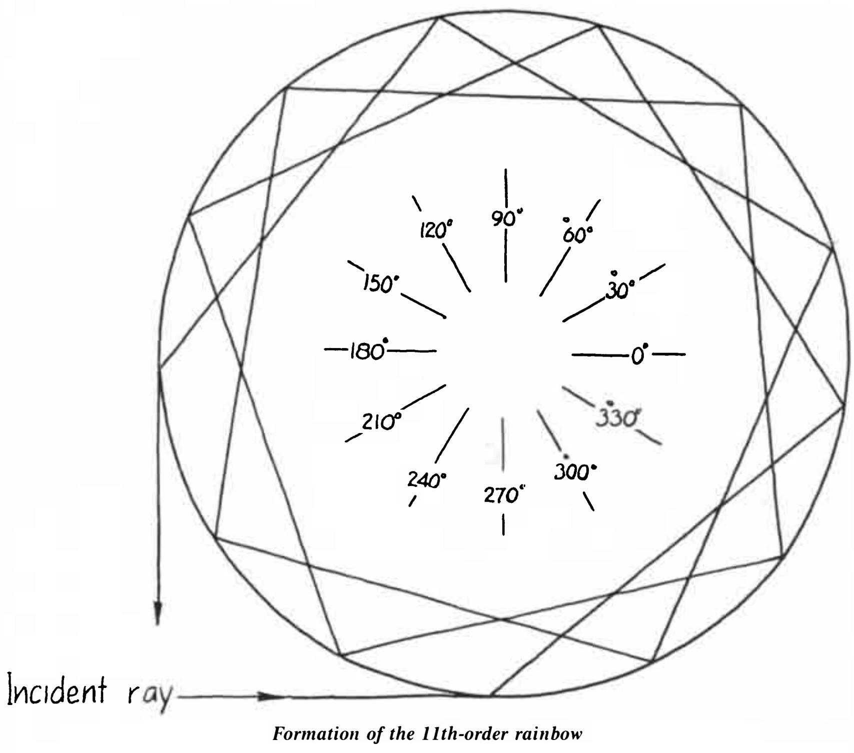
Rainbow order	Angular width	Red	Blue	Side of drop
1	1.72°	137.63°	139.35°	right
2	3.11	129.63	126.52	left
3	4.37	42.48	38.11	left
4	5.58	42.76	48.34	right
5	6.78	127.08	133.86	right
6	7.96	149.10	141.14	left
7	9.14	65.59	56.45	left
8	10.31	17.71	28.02	right
9	11.49	100.86	112.35	right
10	12.65	176.08	163.43	left
11	13.82	93.11	79.29	left
12	14.98	10.19	4.79	split
13	16.15	72.68	88.83	right
14	17.31	155.51	172.82	right
15	18.48	121.69	103.21	left
16	19.64	38.91	19.27	left
17	20.81	43.84	64.65	right
18	21.97	126.58	148.55	right
19	23.13	150.70	127.57	left
20	24.29	67.98	43.69	left

Calculated angles of the first 20 rainbows refraction of 1.749 (in yellow light). Diiodomethane can be obtained from chemical-supply houses such as the Fisher Scientific Company (52 Fadem Road, Springfield, N.J. 07081). I could not find the first- and fourth-order rainbows with diiodomethane because they were almost directly backscattered into the incident beam, and the third and the ninth were buried in glare. The second, the fifth through the eighth, and the 10th were visible.

As the index of refraction increases from that of water to that of corn syrup



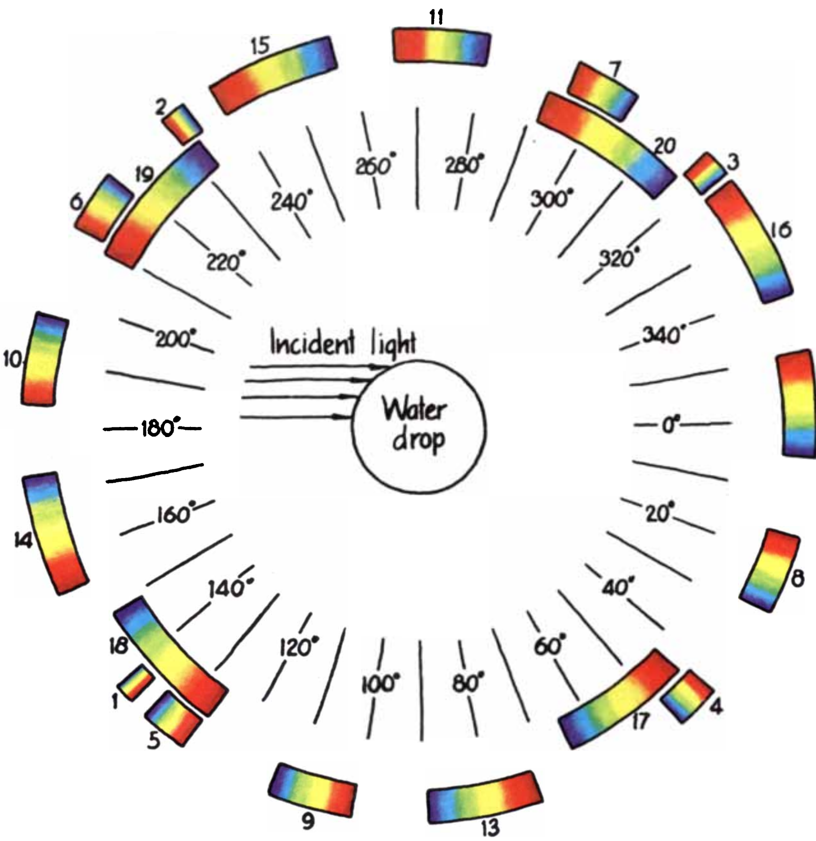
Displacement of rainbows with increasing refractive index



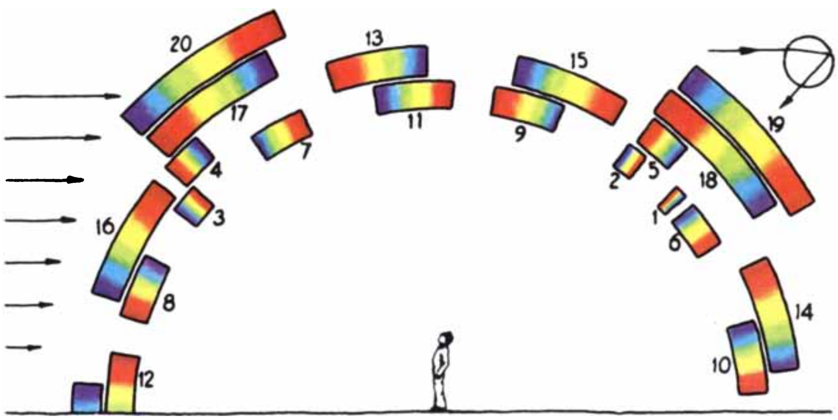
Formation of the 11th-order rainbow

and diiodomethane, the first-order rainbow shifts toward a scattering angle of 180 degrees. If the index of refraction were made exactly 2.0, the angle would reach 180 degrees and the rainbow would be unobservable. Any attempt to see it would block the incident beam. Even if the awkward problem of viewing angle could somehow be overcome, the first-order rainbow would not appear. With an index of 2.0 all the rays enter the drop, reflect from the opposite side and then return to the light source. There is no dispersion and therefore no rainbow. Each of the higher-order rainbows reaches a similar limit at a higher index of refraction.

Not all my grandmother's questions have been answered by this experiment. For example, the faint bows below and above the natural rainbows, which are called supernumerary arcs, have not been explained. Neither has the fading of the rainbow colors as the drops evaporate and shrink. On the other hand, we have seen the delicate colors of more than a dozen rainbows within a single drop of water. What a fabulous sight it would be if all of them were visible in the sky!



Rose of rainbows surrounding a drop



Array of rainbows in the sky

Now – For Only $39.88 You Can Be Using the most Powerful Preprogrammed Calculator Available . . . According to Impartial Critics

Just listing the features and functions of Commodore's SR4190R should prove the absolute truth of the headline. Particularly when thousands and thousands of this model have been sold throughout the world in the last six months at $80.00 and more.

But the "impartial critics" have a lot to say, too, *Machine Design,* a very well-regarded industrial magazine, rated in its February 24th edition 37 personal calculators from 10 major companies, including Hewlett Packard, Texas Instruments, Monroe, Canon, Victor, Casio, National Semiconductor, Omron, Rockwell and Commodore. Out of 21 features, 16 were ideally to be answered "yes." Only *one*—the Commodore SR4190R—could answer "yes" to *all*! Regardless of price . . . because the average price of all 37 calculators was **$109.25.** Compare it with *our* price of **$39.88.**

Here's the way the Commodore SR4190R looked. (If you'd like to see the entire analysis, drop us a note.)

From Machine Design, Feb. 24th:

Handheld Calculators	Commodore SR4190R	
Logic System	Alg	
Display Type	LED	
Trig Functions	yes	✓
Inverse Trig Functions	yes	✓
Common Log	yes	✓
Natural Log	yes	✓
α^x	yes	✓
Reciprocal	yes	✓
Pi	yes	✓
Square	yes	✓
Square Root	yes	✓
Deg/Rad Mode	yes	✓
Rec/Polar Conversion	yes	✓
Hyperbolic Functions	yes	✓
Deg-Min-Sec to Dec Conversion	yes	✓
Metric Conversion	yes	✓
Scientific Notation	yes	✓
Rechargeable	yes	✓
Digits (total)	14	
Storage Registers	2	
Program Steps	0*	
Price ($)	70**	

*You probably won't miss the program steps; see the next quote.

**Our price is $39.88!

Electronics Today International, the British equivalent of *Scientific American* (loosely speaking, of course) rated the SR4190R a few months ago. Here's what they said: "The function list is really quite amazing . . . would seem to have *most functions you could wish to programme readily available.*

"Easy to use. Algebraic logic nice for us humans. Conversions, statistical and graph plotting facilities more comprehensive than any we've seen.

". . . A distinct step upwards . . . and of more use than some that are programmable. *Definitely recommended.*" (All italics ours.)

The compact SR4190R
Only 5-3/4"L x 3"W x 1-1/2"H.
With 49 keys and on/off switch;
91 keyboard functions.

About the only thing left for us to say is that these are among the features you'll find in the SR4190R (forgive us if we're a bit repetitious). Each is equivalent to one or more complete card programs on an HP65 or TISR52, but is activated by a single keystroke:

Poisson, Binomial Probability and Gaussian Distribution ▪ Mean and Standard Deviation ▪ Lineal Regression Analysis ▪ Numerical Integration . . . allowing calculation of area beneath a curve ▪ Permutations and Combinations . . . with no overflow limitations ▪ Factorial Computations and Gamma Function ▪ Complex Numbers Arithmetic ▪ Hours-Minutes-Seconds Mode . . . for time and motion study ▪ Polar/Rectangular, Rectangular/Polar conversions in all quadrants ▪ Percent and Delta Percent ▪ Log and Antilog . . . base e and base 10 ▪ Direct and Inverse Trig Functions ▪ Direct and Indirect Hyperbolic Functions ▪ Xth Root, Xth Power, Square Root, Reciprocal ▪ 11 Back- and Fourth British/Metric Conversions ▪ 2 Storage and 2 Grand Total Memories.

The bright red LED display has 14 characters: 10 mantissa, 2 exponent, 2 signs. The exponents may be entered by the entry key (or automatically, of course) and there are exponent increase and decrease keys.

The SR4190R is fully rechargeable; the nickel-cadmium batteries and the adapter/recharger are included.

The SR4190R is extraordinarily versatile—almost like having five calculators in one. It's a quite complete *business executive* calculator; an extraordinarily powerful *slide rule;* a powerful *statistical* calculator; an advanced *scientific* functions calculator; a comprehensive *conversions* calculator.

All this for only $39.88. How is it possible? *Electronics Today International* may have posed the question best: "About the only question the 4190 won't answer is how Commodore produced it at the price they have." (Mind you, the magazine's comment was based on a £59.90 price . . . or a bit over $100.00 at the current exchange rate.)

The Commodore is a pioneer in offering the consumer Value in personal MOS/LSI electronics. The producer of some of the first business and consumer electronic calculators, the Commodore today has one of the very few vertically integrated companies, designing and producing its own chips and microprocessors, LED and LCD displays and then assembling in its own facilities. For over 20 years, the Commodore's high technology, large volume, vertical integration efficiency and no-frills marketing have given the consumer more features and quality for less money. He will never change his policy.

We're so sure the Commodore SR4190R is exactly right for you that we offer this extraordinary double guarantee: *Use it for 10 days.* If it's not exactly what you want, return it to us for full refund. No questions asked.

Once you decide to keep it, your SR4190R is *guaranteed unconditionally for one full year:* regardless of defect, we'll replace it with a brand new one if you return it to us within the year.

ADVANCED CONSUMER ELECTRONICS, INC.
P. O. Box 730 • Los Altos, CA 94022

Please send me________SR4190R calculator(s) by Commodore at only $39.88 each complete with adaptor/recharger. (Plus $2.50 each for shipping, handling and insurance.)

☐ I'm enclosing check or money order payable to Advanced Consumer Electronics, Inc. for $____________

☐ I'd like to use my charge card:

☐ Master Charge ☐ BankAmericard ☐ American Express

Card No.____________

My credit card expires____________

I understand I can return the SR4190R calculator(s) in 10 days for full refund or credit if I'm not happy with it (or them).

Signature____________

Name____________

Address____________

City____________

State, Zip____________

(Calif. residents: please add 6% ($2.40 ea.): Santa Clara Cty, CA add 6.5% ($3.06 ea.)

BOOKS

Linear perspective, M. C. Escher, the great Dutch famine of 1944–45 and other matters

by Philip Morrison

THE RENAISSANCE REDISCOVERY OF LINEAR PERSPECTIVE, by Samuel Y. Edgerton, Jr. Basic Books, Inc., Publishers ($15). THE MAGIC MIRROR OF M. C. ESCHER, by Bruno Ernst. Random House, Inc. ($15). Copernicus was born in 1473; that 500th anniversary rang throughout the world of science. Here we celebrate, along with an imaginative and experimental art historian (of Boston University), a much less arbitrary moment in the growth of the Renaissance point of view. One day in the year 1425 "a short, middle-aged man arrived at the piazza" between the still unfinished cathedral of Florence, pride of the Republic and of its rich and generous bankers and wool merchants, and the famous multicolored marble baptistry that still faces the portal of the great church. That man was the sculptor, architect and artisan-engineer "Pippo di Ser Brunellescho the Florentine." He had with him a small painted panel about a foot square, with a curious hole in it, and a mirror of the same size. He stood within the portal, we may guess, looking out toward the baptistry, with the painting held in one hand "oddly enough, obversely right up against his face." In the other hand he balanced the mirror, which he dropped in or out of the line of vision from time to time. "Is that a mirror reflection in his eyes, or a glint of satisfaction?" He has cleverly matched his painted image (as seen looking through the hole in the painting at its reflection in the mirror) with what the eye sees. It was the first modern demonstration of linear perspective with lines converging to the vanishing point. Indeed, the author of a treatise of the 1460's ascribed that "subtle and beautiful" discovery to Brunelleschi's "considering what a mirror shows to you." It had to be a plane mirror, of flat glass, which at the time was rare and expensive, and so it was the mirror that determined the scale of the small painting.

Professor Edgerton takes the reader to that same portal and shows us with his own mirror and Nikon and a few persuasive plans, drawings and photographs what was the probable field of view, where the painter stood to take his cue from the mirror and even the little tilt he may have given the mirror to balance the sky and the pavement satisfyingly in his image. Other scholars have suggested that Brunelleschi painted directly on the mirror (impractical and unlikely) or that he made the discovery first intricately on paper, merging the elevations and plans of the church he was no doubt preparing. It was of course Brunelleschi himself who closed the big unfinished edifice with one of the greatest cathedral domes in Europe during those very years.

Edgerton's happy experiment makes a strong case (stronger than he seems to think) for Brunelleschi's being the first to paint in perspective, for his learning it from the images in plane mirrors and for his tellingly demonstrating his discovery to his scholarly and painterly friends with his dramatic image-superposition scheme. At the same time, give or take a year or so, his friends the painter Masaccio and the sculptor Donatello created the first works of art clearly to employ linear perspective. His patrician friend the learned Leon Battista Alberti, who was unique in the breadth of his training in science and in letters and who had taken up painting as a relaxation, wrote the first explicit treatment of linear perspective as a device for painters.

Alberti prepared a version of his work in Italian (for the artists to read), which he dedicated to Brunelleschi among others, 10 years after the play with mirrors in the Piazza del Duomo. The wonderful painted panel has not survived, but Brunelleschi's biographer, writing in the 1480's, stated, "I have had the painting in my hand and have seen it many times," and he describes the panel size, the peephole "small as a lentil on the painting side of the panel" and the care, delicacy and accuracy of the work. For the sky the painter had put burnished silver leaf, to reflect the natural clouds so that they might be "moved by the wind when it blows."

Only a fifth of this fresh and lively small book is given to an account of the analysis and reenactment of Filippo Brunelleschi's invention; most of it is a study in the history of ideas from paintings and relevant texts. Alberti gains the most attention; one of the clearest perceptions is into his language, which serves as a sign for the centrality of human experience in his period. He speaks of concave surfaces as "the inner surfaces of egg shells" and of a plane as consisting of "many lines...joined closely together like threads in a cloth," making colorful metaphors of the world of everyday experience to replace "the black and white of Euclid."

Edgerton is anxious not to overpraise his protagonists; by the very title of the book he allots first place to the ancients. He asserts that Euclid himself, and Ptolemy too (particularly in the projections described for mapmaking in the *Geographia,* which was first seen in the Christian West in about 1400 in Florence), had possessed the main ideas of Alberti and his contemporaries.

What is plain is that linear perspective fits the grand metaphysics of the Renaissance. It was their "symbolic form," to use the term Edgerton draws by way of the art historian Erwin Panofsky from the philosopher Ernst Cassirer. It constructs a space in which for one moment and one position the observer becomes objective, a space based on his experiences having been abstracted impartially for all. In the end it was more science than art that came to prosper from this discovery on the part of the artists. The entire quality of linear perspective, its empty space free from boundary or change, suits well the spaces of Newton, the lens of Hooke and the traveler's notebook of Humboldt.

Nowadays science has matured. Our projective spaces do not exclude Newton, but they span much more. Our world is buzzing, various, half-chaotic, half-patterned. One impartial standpoint is no longer enough; time and chance happeneth to it all. The artists too have made peace with this world view, but they have not led it. It has been a long time since the perceptions of science have been pioneered by the painter. Remarkably enough, one modest Dutch printmaker came to express visually in his meticulous and ingenious woodcuts and lithographs a great deal of what our bold new postulates and richer instrumentation have added to the lucidity first grasped in Alberti's city.

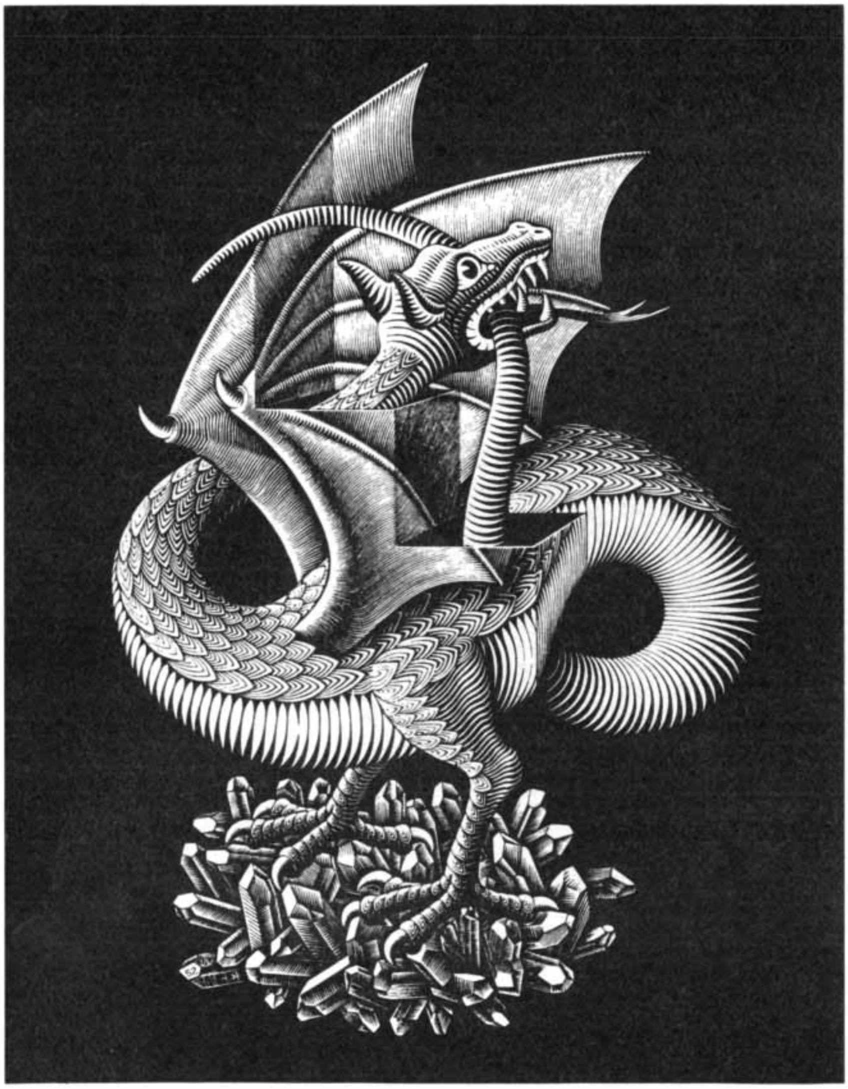
Bruno Ernst was a friend and neighbor of M. C. Escher's, and a teacher of physics and mathematics and editor of a mathematics magazine for Dutch students. In this fine book he presents many of Escher's best-known prints in the context of models, sketches, personal confidences, influences and the hopes Escher expressed to him and to others. Vanishing points? Of course Escher used linear perspective with uncommon crispness. But here is the famous "High and Low," where there are two vanishing points, one dragging our gaze up to the tiled ceilings, the other pulling it down to the tiled floor. Relativity has entered; the

same small patch is at once the ceiling and the floor, depending on where the observer stands. The finished lithograph is reproduced, and with it four or five sketched versions and Escher's own account of how one vanishing point can and should be made to function for two observers, high and low.

Here too is Escher's remarkable grid in which the scale grows steadily as we encircle a point in the center of a drawing. The print space bulges enormously; the window mullions of the print gallery reveal to us not a ship print but a great ship in the harbor. Scale has become explosive; what Escher had done, the "learned gentlemen ... once tried in vain to convince me," was to draw on a Riemann surface. He was not trained in mathematics, but he came to its conceptual schemes by his artist's vision and his determination to present that vision consistently. "So far as I was concerned it was merely a question of a cyclic expansion or bulge, without beginning or end." Even specialized modern theory can speak to a reflective percipient.

Escher's themes are marshaled and exemplified: different structures of space interpenetrating; the flat surface made to fill, to contain an approach to infinity within a limited bound, to reflect becoming no less than being; the conflict between the optical devices of perspective and the intricate algorithms of the unconscious by which we judge depth and position in pictures. It is all a rich legacy, to which every admirer of Escher will repair for a sense both of the man and of his methods. The little horsemen ride by, the one melting into the other; the last pages show us Escher's last work, a twining of three elegant, colorful snakes into a double interlacing of rings, suggesting infinity at both the edge and the center.

The book begins with a magic mirror, Escher's mirror, which sets its little beasts free to walk around on a tiled table. Overleaf we see the end point of linear perspective, Andrea Pozzo's great baroque trompe l'oeil ceiling in Rome. The plenitude of the world lies still uncaught by art or by science. The book shows a color snapshot of Ernst with Escher a few weeks before the artist's death, with the touching quotation: "I consider my work the most beautiful and also the ugliest." It is here again remarkable how one can catch in apparently conventional work of student days the prefiguration of an artist's maturity.



"Dragon" (1952), from **The Magic Mirror of M. C. Escher**

THE LAST GREAT SUBSISTENCE CRISIS IN THE WESTERN WORLD, by John D. Post. The Johns Hopkins University Press ($12.95). FAMINE AND HUMAN DEVELOPMENT: THE DUTCH HUNGER WINTER OF 1944/45, by Zena Stein, Mervyn Susser, Gerhard Saenger and Francis Marolla. Oxford University Press, Inc. ($12.95). Bourbon and Hapsburg had won; Napoleon was on St. Helena. In this early summer of the year 1816 the Belgians were preparing to celebrate the first anniversary of Waterloo. But there was something wrong in the very look and feel of the summer sun. A New York physician wrote: "The human eye could thus, during the long days, gaze on the great luminary ... deprived of its dazzling splendour and radiance; ... numerous dark spots were discovered on its face, without the help of telescopic ... glasses." Many thought those sunspots were significant, but others noted the long twilights, a blue moon, blood red sunsets. Far from the Atlantic world the volcano of Tomboro 200 miles east of Bali had erupted in 1815, killing tens of thousands by explosion and ashfall. Hundreds of cubic kilometers of dust entered the air to spread around the world in the dustiest eruption of three centuries (three times the dust of Krakatoa in 1883).

The year 1816 would be the year without a summer. The data are clear enough: Milan, Geneva, Paris, Salem—at each station the mean temperatures were down from the years before, without any exceptions for the months of June through September, by some four to eight degrees Fahrenheit. Professor Dewey of Williams College summed up the year: "Frosts are extremely rare here in either of the summer months; but this year there was frost in each of them.... June 6th ... snowed several times.... July 9th, frost, which killed parts of cucumbers. August 22nd, cucumbers killed.... Very little Indian corn became ripe in the region."

The result was famine, one not matched in the Western world since that time and certainly the worst since the year 1710. The author, an economic historian at Northeastern University, paints a careful picture of economic distress and its political and social consequences across all Europe. That unsettled postwar world was ill-prepared for

so disastrous a harvest. Grain prices nearly doubled in 1816 and 1817 in a dozen European markets (and in the U.S.), and in not one was there an appreciable decrease. Many starved in Hungary and even in Switzerland, where the price of bread rose four- or fivefold compared with 1814 levels. Interior New England suffered heavily; in Vermont farm families lived on hedgehogs and boiled nettles.

Not all the hungry were resigned. Strikes had not yet become the mode, but everywhere there were demonstrations outside bakers' shops, looting and rioting, and vagrant bands often attacked farm granaries. In Suffolk they carried a flag inscribed "Bread or Blood." There was extensive rioting in Brussels on the second anniversary of Waterloo. A mass German emigration was set off: "Whole multitudes of emigrants with their families" went down the Rhine to the Dutch ports and on to America; others went to Russia. "A sort of stampede took place from cold, desolate, worn-out New England to this land of promise." The writer was referring to Ohio.

Actually "the old biological regime" had already been shattered everywhere in Europe. Population was steadily on the rise; we are not quite sure why. High peaks in mortality had been the usual consequence of subsistence crises up to the second half of the 18th century; the deaths in hungry years offset the years of growth. The famine of 1816–17 was the last of those continental peaks; death rates rose all over Europe, most steeply in Switzerland, Italy and the Carpathians. In 1710, the previous great crisis, mortality had been enormous from pestilence, particularly plague. This time typhus took some toll, notably in East London and among the Swiss and the Irish, but plague was unimportant. The author contends that plague had left European cities largely because of structural change, from fire-prone wood and shingle walls and roofs to brick and stone. In the 18th century most of the ports of western Europe had become fireproof, and thus also inhospitable to the black house rat. It was not lack of food directly but social turbulence that had formerly linked famine and pestilence.

Good weather and low grain prices returned by 1820, but the volcano had pulled the trigger of history. Governments moved to the right all over Europe. Political repression grew (but so too did the prosaic but positive administrative programs that have ever since tempered the challenge of scarce grain to the life of the poor in Europe). By 1819 the dust from the distant crater had touched off a bizarre foreshadowing of Europe's deepest tragedy. The culpable students of Bavaria began a campaign of disorder that ended in violent anti-Semitic riots all over Germany and as far away as Copenhagen, "to the taunting cry of Hep-Hep," an obscure phrase followed by "Jude verreck" ("Jew, drop dead"). The military put an end to these "outbreaks of the vulgar masses," as Prince Metternich saw them. Once such outbreaks appear, he argued, "no security exists, for the same thing could arise again at any moment and over any other matter."

It is four decades since that "same thing" arose again in its most terrible array, now as state policy. One of its smaller cruelties was a famine induced in the populous and well-run cities of western Holland—Amsterdam, Rotterdam, The Hague and others—by the deliberate act of the Occupying Authority and the Wehrmacht. It was 1944, late in their day; the Allied forces had driven for the Rhine bridges without success. The Dutch rail workers brought all rail traffic to a stop as soon as the battle began in late September. In reprisal all transport to the big Dutch cities, including food supplies, was embargoed by the Germans. When water traffic was allowed as a concession in November, an early freeze prevented the barges from moving. "By that time the acute food shortage had become a famine." That half of the Dutch people who lived in the western towns suffered true famine until liberation in May, although rural people had eaten well enough all over the country. But the embargoed cities starved. "The ordinary person... consumed in two or three days all that was given for the whole week."

By January hospitals were filled with the starving. Nurses and physicians worked day and night without extra rations. A slice of bread and a tea substitute was their breakfast; two potatoes and some watery sauce was lunch; dinner was one or two slices of bread and a plate of soup. Special starvation hospitals were opened in February. People dropped in the streets; the old ones, who could not roam in search of food, stayed in bed and died, out of sight of the doctors. The mortality figures show 10,000 dead from hunger. One graph tells the story: a deep notch appears in the plot of the monthly food ration. That notch is the bite of the long hunger winter.

Why dwell on such a mournful story? People starved in greater numbers in gallant Leningrad and in the camps all over the greater Reich. But this calm, meticulous, well-written book turns the famine into a heroic large-scale experiment on human development. This starving population was in no way in disorder. The births and the deaths were meticulously recorded as usual. In 1971 and 1972 these authors checked up. They took 2,000 birth records at random among all the birth registers of the famine city of Amsterdam (and of a control city in the south). The names they chose were those of baby boys. In his 19th year each of them would come to be examined medically and psychologically by the doctors of the draft board, under a uniform and well-recorded protocol. In the validation sample 3 percent had died and 2 percent had emigrated, and there were the expected 3 percent of legal exemptions from the draft. In all, 97 percent of the births of the preceding generation were either represented by induction medical records or lost to known channels. The missing names were distributed through the lists randomly.

About 40,000 people were exposed to prenatal famine. The coded tapes of the Dutch military hold the data, the best large-scale survey we have of the health of mature people who were exposed to famine before birth. The controls were less than perfect. They are twofold: comparisons in time (among individuals born before, during and after that famine winter) and in place (with those born in southern cities not held by the Germans). History provides no well-managed laboratories. The control cities no less than the famine ones were literally decimated—one in every 10 babies dead—by a mysterious epidemic in the summer of 1945. But the controls retain enough leverage to show, for instance, that babies born to mothers who were hungry during the last trimester of pregnancy were small ones. This result is fortified by a quite elaborate set of analyses, testing multiple correlations between the size of rations and hospital measurements. The significance of the best results is better than 1,000 to one.

From this nutritional experiment carried out on pregnant women that bitter winter the authors (a team of epidemiologists at the Columbia University College of Physicians and Surgeons) have woven their compact, explicit, table-filled book, a model of perception and of honesty. They begin with reflective accounts of the biological and social questions they might face, they devise tentative models for the correlations they turn up, they put forward extra little tests, subdivided samples and controls to tease out the meaning of the data in their study. The message they bring is on the whole a welcome one. In this brief description of the work the message is summed without the evidence that gives it strength. It is the evidence that is the burden of the report, a statistical study whose findings lead to strong conclusions. Hunger cuts the birthrate and the size of infants. The poor suffered most. But overall there is one unequivocal finding. Among the young men reaching 19 who had been born to women exposed to famine no detectable effect was found on mental performance, physical stature or health, with one exception: there was a larger percentage of rare congenital anomalies of the central nervous system, such as cerebral palsy. Above some nutritional threshold, at least, the mother protects her fetus.

The senior authors have pursued their studies of the origins of public health from the ghettos of Johannesburg to those of Manhattan. They have "exploited the collective woe of a whole people" in search of hope for the future. From the Dutch famine they can now make plain that one winter's cruelty is largely reversible. But what of the myriads of infants born to poorly nourished mothers who remain poorly nourished during a lifetime of poverty thereafter? Hunger did not leave the West with the dust of Tomboro.

ISLAND OF ISIS: PHILAE, TEMPLE OF THE NILE, by William MacQuitty. Charles Scribner's Sons ($14.95). EARLY HYDRAULIC CIVILIZATION IN EGYPT: A STUDY IN CULTURAL ECOLOGY, by Karl W. Butzer. The University of Chicago Press ($8). The lines form at dawn for those who would view the tomb spoils of young Tutankhamun, the New Kingdom ruler whose golden treasure has been displayed in London, Japan, the U.S.S.R. and this year in American cities. The unprecedented popularity of the show has a real consequence on the Nile: the receipts go to UNESCO to pay for saving the temples of Philae from the river waters. Philae is a small granite island in the Nile just above the First Cataract, where once the Nile flowed through deep rocky channels and wide pools over rapids to the sea 700 miles away. The polychrome walls and the aisles of lotus capitals of half a dozen temples there have been famed for a long time; Philae is the "pearl of Egypt." They are not ancient, as Egypt reckons monuments. They were built over five centuries, mostly by the Greek Ptolemies and by the Roman rulers who followed the last of the Ptolemies, the murdered son of Julius Caesar and Cleopatra VII Philopator. All these classical structures paid tribute to the old deities of Egypt. We see a Ptolemy kneeling on the sacred boat of Isis, offering to the goddess. Carved on an unfinished building with superb floral capitals, the emperor Trajan burns incense and offers wine to the sacred family: Isis, Osiris and Horus.

Philae has been ill-used in the 20th century. Even before that the Coptic iconoclasts carved their crosses and defaced the reliefs of an infidel past. The first dam at Aswan was flung across the Nile between 1898 and 1902, less than a mile below Philae. The water level each year stood high on the walls of Philae, 10 feet above ground level. Year after year the waters rose and fell in the cycle of the Nile year, half-concealing the monuments nine months out of 12. It developed that "the submersions were not altogether bad." They saved some wear and offered some protection from souvenir hunters; the waters receded only during the hot season, when there were few tourists.

The paintwork has vanished, but the British engineers of the dam had carefully underpinned the monuments to prevent structural damage. At first the lake level was set low in response to people who wanted to preserve the monuments. Winston Churchill wrote at the time: "This offering of 1,500 millions of cubic feet of water to Hathor . . . is the most cruel, the most wicked and the most senseless sacrifice ever offered on the altar of a false religion. The State must struggle and the people starve in order that the professors may exult and the tourists find some place to scratch their names."

The High Dam came to the Nile, four miles upstream from the old barrage, between 1964 and 1971. Philae is the only temple caught between the two dams; it was almost submerged and was subjected to a daily rise and fall of several yards. Like the ancient cliff temple of Abu Simbel 175 miles upstream, the carved sandstone wonders of Philae would be doomed by the new regime of the Nile unless something was done. UNESCO and the Egyptian government, with international engineering consultants, have moved the buildings of Philae. In March of 1972, as the waters first rose, the Italian contractors began to drive a double row of 320,000 sheet-steel piles through the mud and silt to the bedrock river bottom, a double wall enclosing the entire island, some 20 or 30 acres. A great sandbank three miles away was loosened by high-pressure jets of river water, and the slurry was carried in pipes to fill the corridor between the two walls of piling. The surplus water leaked out and the contained sand held back the Nile. Then the enclosed space was drained by the same pumps, which toward the end were installed in deep wells to dry the surface.

Once dry, the temples were taken apart and then reassembled stone by stone on another island, less than half a mile away, that stands eight or 10 meters higher than Philae. Philae is no longer to be seen above the newly blue water, but the temples, "made to last for eternity," are safe on their new landscaped site, which is scheduled for completion this summer. (Nile water provided the pumping power, with a diesel standby against failures.) Fine photographs in color and black and white, descriptions of Philae and of many other temples of the Nile, a quick account of the Ptolemies (who seem half terror, half romance) and a valuable context complete a handsome volume, a model of popular but careful work, at once a chronicle and a traveler's introduction.

Early Hydraulic Civilization in Egypt is brief, with graphs and tables, maps, geologic sections and line drawings, and its text is composed by typewriter. It is neither easy reading nor appealing to the eye, but it is a remarkable example of the best new archaeology, going a long way toward placing Egypt in a wider context of human development. Even the general reader senses that the riches of Egyptology have too long been isolated; their findings remain aloof from the common pool of understanding, showing "an introspective emphasis on content rather than context."

This candid, quantitative, critical and original study is by a well-known anthropologist-geographer. He has for years worked not only on the Nile but also on another river that flows from the same plateau but is untamed by man, the Omo. The flavor of his work is in that relation: comparative and skeptical, he finds in the lower Omo a model of the prehistoric Nile.

Agriculture came late to the Nile delta because local hunting and gathering worked so well. The irrigation then was natural, as geologic sections and test drilling confirm. The delta was no swamp but a complex array of turtlebacks of sand and silted natural levees. Game was common, and the settlers used the riverbank for a single crop season, naturally irrigated by the rise and fall of the Nile. Then came artificial irrigation, visibly celebrated on the carved mace head of the Scorpion King, about 3100 B.C. A limited input of labor could much improve the natural irrigation, with short ditches, earth dams, hand buckets for raising water. The shadoof—the pole-and-bucket lever—came in with the New Kingdom; its works too concentrated on regulating the high-water Nile. There were still no summer crops. The Nile gradient is only one in 12,000, and the complex radial canalization that was developed for the Mesopotamian plain was of little use here. Only after the advent of animal-powered chains of buckets, as late as the Ptolemies, could water be lifted several meters and low-water crops become practical.

There is a kind of mandala, a key circular diagram carefully assembled from primary data of the 18th century of our era, that fully displays the round of the Nile year. Crop follows crop on the floodplain, in a state of perennial irrigation. Wheat, barley, chickpeas, onions, lentils, clover, dates, sugar, cotton and more round out the cycle of the year. Full irrigation sustains a growing population, five million by Roman times. That irrigation was not a radial, branching network but linear, flood basin by flood basin along the lengths of a few main channels. Although the state was present early, Egyptians did not become city dwellers, or even producers of crops for market. They were villagers, who managed their own irrigation without a strong central irrigation bureaucracy, for which there is no evidence in the "plethora of Old Kingdom titles." Whatever made the Old Kingdom, it was not the pat stages from population stress to irrigation, to central management of the waters and finally to the despot king. We need to look at the facts of land and water and resist the seduction of a prematurely unified theory. On the other hand, there are causes to be winkled out; history in Egypt, no less than elsewhere, has a strong coupling to the changing land, as strong as to church, state, technique or symbol.

The New World Primates: Adaptive Radiation and the Evolution of Social Behavior, Languages, and Intelligence, by Martin Moynihan. Princeton University Press ($12.50). The New World has no great apes, and large tailless primates entered it overland only in the past 100,000 years or so, by the most extreme of estimates, although they have lived in the Old World for many times that span. We have kinfolk in the Americas, all right, the monkeys of the tropical forest. Their home is the great and ancient South American lowland forest, where most savanna seems recent and the flowers and the birds are present in rich assembly (roughly double the species diversity of Africa). The very parrots there are almost exact analogues to our little monkey cousins: "tropical, arboreal, diurnal, characteristic of forested regions... often highly gregarious, with complex systems of... signals, usually brightly colored in intricate patterns, omnivorous or at least very catholic in their choices of foods, and sometimes good manipulators."

That intriguing remark is typical of this unusual small book. Dr. Moynihan, a primatologist of the Smithsonian Tropical Research Institute, has spent 17 years watching monkeys in the wild and in captivity amidst the dense forests of Panama and the uppermost Amazon. He reports his findings in a wider context, from lemurs to chimpanzees. This is a personal, reflective, almost quizzical book, full of perceptive generalizations and conjectures, written close to the animals themselves, strikingly illustrated by the author's own lively ink sketches, simple and yet evocative of the quick, clever little creatures he knows and describes. He has watched them all, pretty clearly, the marmosets and the howlers, the spider monkeys, *Aotus* (the only nocturnal species) and more. The most monkeylike are the organ-grinder's monkeys, the wary capuchins. Out of this experience he puts and tentatively answers the questions implicit in the subtitle of the book.

The capuchins seem close to tool-using when they pound hard-shelled nuts on surfaces to crack open the nuts. They explore and play with everything they can lay their hands on; they "are very prone to take things apart," and they can put things together "when the spirit moves them." They bite most fruits they encounter, rejecting many but accepting an impressive diversity. They count many insects in their diet, and the common name for the species in Colombia is maize-eater. Alone among American monkeys they come down from the trees to the forest floor, even for excursions into open country, although they remain mainly arboreal. Their thumbs oppose, and their tails are almost as good as hands. Moynihan reports an hour-long game of catch with a stolen handkerchief, caught about equally often by hand or tail tip and sometimes thrown in an "undertail" lob.

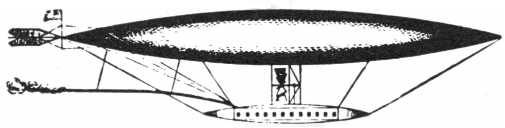
It is hard not to expect more from these small relatives. Most species have complex, visually stereotyped languages of display and gesture, with dozens of elements. They emit a vocal repertory of squeaks, whistles, trills, barks and grunts, social in use, even sequential. Many sound spectrograms are reproduced but not much is yet known about the use of the sounds. Most of the signals—visual, tactile and vocal alike—are innate, unvarying from occasion to occasion. There is no production of new sentences in evidence. Curiously, under strong social stress in conditions of captivity a few individuals of some species develop unusual special movements and postures, often elaborate in form. The author calls these "quirks," and he notes that the other monkeys pay little attention to them and do not pass them on by imitation. Human beings can nonetheless judge the state of the animal by the quirks. (We badly miss a specific description of a few quirks.)

Here Moynihan sees the essence of our language. Human language is learned, whatever preadaptations it implies. Our ancestors were selected, he conjectures, for the ability to learn new signal quirks, so to speak. Why? He argues that the use of tools meant the need for new signals to describe them; the number of signals had to match the increasing number of useful objects; innate signals were too slow to change and too few. It was flexibility, true learning, that was rewarded by selection. The capuchins are still caught in the forest; they would have trouble among those dense intertwining branches with quirks, which still involve hands and arms. But more ground foraging, much more tool use, more need for vocal signals, more selection for inventiveness.... "It would be interesting to see what happens to them in the next million years or so."

To Fly like a Bird: The Story of Man-powered Aircraft, by Keith Sherwin. Distributed by Sterling Publishing Co., Inc. ($6.95). The dual role of engine and engineer is certainly a taxing one. The frontispiece shows the author, a little winded, sweating at the controls and crankshaft of *Liverpuffin,* a man-powered aircraft built by the students in the department of mechanical engineering at the University of Liverpool. The craft flew, although another picture shows the author "surrounded by the crashed remains" when a gust caught the machine and overturned it on the runway. These great-winged propeller-driven flying bicycles—to give them a trivial if accurate name—are exquisitely delicate devices. The best performer, *Jupiter,* built by Royal Air Force cadets, has flown three-quarters of a mile. Its expert pilot is sure that his experience helps; he must operate simultaneously in two different modes, "the upper half geared to control whilst the lower half must be geared to the motive power." *Jupiter* is a balsa-and-spruce wonder, very clean, covered with plastic film. Its weight empty is 145 pounds; its wingspan is 79 feet. Among the 15 planes listed worldwide that do work (here in the U.S. we have so far only a few hopeful aspirants) there are featherweight designs with about half that wing loading. Man-powered flight is short on power and must strive for lift at low speed.

Of course the aspiration to fly by the sweat of one's brow is age-old. The first such flight (we set the ancient Minoan pioneers aside) was in 1935, in a plane built by two engineers at Junkers. They flew 80 times, their best distance being about 790 yards. An Italian rival seems also to have succeeded about then. The new epoch began in the late 1950's, when a committee of the Royal Aeronautical Society found a donor, the industrialist Henry Kremer, to offer a big prize (now 50,000 pounds) for flight under rather stiff conditions: in a figure-eight pattern half a mile long, higher than 10 feet throughout, without power storage or the aid of a buoyant gas, in still air (wind under 10 knots). It now appears that this task is all but impossible, and Professor Sherwin favors a retreat to a sporting stance, not aiming at the Kremer Prize for the present but rather trying to build up international experience, with many designs and the development of skill by experienced pilots.

This small, knowing narrative treats the entire delicious topic in context: its history, the related but distinct activity of hang-gliding, recent efforts, a simple review of design and operational problems, comments on the activity as an ideal student project and its development into a less exacting sport. The same author earlier published a text on the engineering design of man-powered aircraft, but this book is the first for the general reader. The photographs are illuminating and the stories are intimate.



BIBLIOGRAPHY

Readers interested in further exploration of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE RECOMBINANT-DNA DEBATE

THE MANIPULATION OF GENES. Stanley N. Cohen in *Scientific American,* Vol. 233, No. 1, pages 24–33; July, 1975.

MOLECULAR BIOLOGY OF THE GENE. J. D. Watson. Addison-Wesley Publishing Co., Inc., 1976.

NATIONAL INSTITUTES OF HEALTH GUIDELINES FOR RESEARCH INVOLVING RECOMBINANT DNA MOLECULES. Public Health Service, U.S. Department of Health, Education, and Welfare, June 23, 1976.

RECOMBINANT DNA RESEARCH: BEYOND THE NIH GUIDELINES. Clifford Grobstein in *Science,* Vol. 194, No. 4270, pages 1133–1135; December 10, 1976.

NOTES OF A BIOLOGY-WATCHER: THE HAZARDS OF SCIENCE. Lewis Thomas in *The New England Journal of Medicine,* Vol. 296, No. 6, pages 324–328; February 10, 1977.

RECOMBINANT DNA: FACT AND FICTION. Stanley N. Cohen in *Science,* Vol. 195, No. 4279, pages 654–657; February 18, 1977.

THE ATMOSPHERE OF MARS

MARINER 9 OBSERVATIONS OF THE MARS NORTH POLAR HOOD. Geoffrey A. Briggs and Conway B. Leovy in *Bulletin of the American Meteorological Society,* Vol. 55, No. 4, pages 278–296; April, 1974.

COMPOSITION AND STRUCTURE OF THE MARTIAN ATMOSPHERE: PRELIMINARY RESULTS FROM VIKING 1. A. O. Nier, W. B. Hanson, A. Seiff, M. B. McElroy, N. W. Spencer, R. J. Duckett, T. C. D. Knight and W. S. Cook in *Science,* Vol. 193, No. 4255, pages 786–788; August 27, 1976.

ISOTOPIC COMPOSITION OF THE MARTIAN ATMOSPHERE. Alfred O. Nier, Michael B. McElroy and Yuk Ling Yung in *Science,* Vol. 194, No. 4260, pages 68–70; October 1, 1976.

STRUCTURE OF MARS' ATMOSPHERE UP TO 100 KILOMETERS FROM THE ENTRY MEASUREMENTS OF VIKING 2. Alvin Seiff and Donn B. Kirk in *Science,* Vol. 194, No. 4271, pages 1300–1302; December 17, 1976.

VIRAL HEPATITIS

HUMAN VIRAL HEPATITIS: HEPATITIS-ASSOCIATED ANTIGEN AND VIRUSES. Arie J. Zuckerman. North-Holland Publishing Company, 1975.

PROCEEDINGS OF A SYMPOSIUM ON VIRAL HEPATITIS in *The American Journal of the Medical Sciences,* Vol. 270, No. 1, July–August, 1975; Vol. 270, No. 2, September–October, 1975.

APPROACHING THE CONTROL OF VIRAL HEPATITIS TYPE B. Joseph L. Melnick, Gordon R. Dreesman and F. Blaine Hollinger in *The Journal of Infectious Diseases,* Vol. 133, No. 2, pages 210–229; February, 1976.

ADVANCES IN VIRAL HEPATITIS: REPORT OF THE WORLD HEALTH ORGANIZATION EXPERT COMMITTEE ON VIRAL HEPATITIS. WHO Technical Report Series, No. 602, 1977.

BIOLOGICAL REGENERATION AND PATTERN FORMATION

PRINCIPLES OF REGENERATION. Richard J. Goss. Academic Press, Inc., 1969.

SUPERNUMERARY LIMBS IN AMPHIBIANS: EXPERIMENTAL PRODUCTION IN *NOTOPHTHALMUS VIRIDESCENS* AND A NEW INTERPRETATION OF THEIR FORMATION. Susan V. Bryant and Laurie E. Iten in *Developmental Biology,* Vol. 50, No. 1, pages 212–234; May, 1976.

PATTERN REGULATION IN EPIMORPHIC FIELDS. Vernon French, Peter J. Bryant and Susan V. Bryant in *Science,* Vol. 193, No. 4257, pages 969–981; September 10, 1976.

THE DEVELOPMENT OF PATTERN AND FORM IN ANIMALS. Lewis Wolpert, edited by John J. Head. Oxford University Press, Inc., 1977.

THE SOLIDIFICATION OF CEMENT

KINETICS OF PHASE CHANGE, 1: GENERAL THEORY. Melvin Avrami in *Journal of Chemical Physics,* Vol. 7, No. 12, pages 1103–1112; December, 1939.

SOME STUDIES OF THE HYDRATION OF PORTLAND CEMENT USING HIGH VOLTAGE (1MV) ELECTRON MICROSCOPE. D. D. Double in *Materials Science and Engineering,* Vol. 12, No. 1, pages 29–34; July, 1973.

THE CONVERSION OF HIGH ALUMINA CEMENT/CONCRETE. C. Bradbury, P. M. Callaway and D. D. Double in *Materials Science and Engineering,* Vol. 23, No. 1, pages 43–53; April, 1976.

THE HYDRATION OF PORTLAND CEMENT. D. D. Double and A. Hellawell in *Nature,* Vol. 261, No. 5560, pages 486–488; June 10, 1976.

ANIONS OF THE ALKALI METALS

COLLOQUE WEYL IV: ELECTRONS IN FLUIDS—THE NATURE OF METAL-AMMONIA SOLUTIONS in *The Journal of Physical Chemistry,* Vol. 79, No. 26, pages 2789–3081; December, 1975.

ELECTRONS IN LIQUID AMMONIA. J. C. Thompson. Oxford University Press, Inc., 1976.

ELECTRON-SOLVENT AND ANION-SOLVENT INTERACTIONS. Edited by L. Kevan and B. C. Webster. Elsevier Scientific Publishing Co., Inc., 1976.

ALKALI METAL ANIONS: AN UNUSUAL OXIDATION STATE. James L. Dye in *Journal of Chemical Education,* Vol. 54, No. 6; June, 1977.

THE COMPOUND EYE OF INSECTS

THE SIZE OF OMMATIDIA IN APPOSITION EYES. Horace B. Barlow in *The Journal of Experimental Biology,* Vol. 29, No. 4, pages 667–674; December, 1952.

PHYSICAL ASPECTS OF THE SENSE ORGANS. Hl. de Vries in *Progress in Biophysics and Biophysical Chemistry,* Vol. 6, pages 208–264; 1956.

THE VISUAL SYSTEM OF INSECTS. Timothy H. Goldsmith and Gary D. Bernard in *The Physiology of Insects: Vol. II,* edited by Morris Rockstein. Academic Press, Inc., 1974.

THE COMPOUND EYE AND VISION OF INSECTS. Edited by G. Adrian Horridge. Oxford University Press, Inc., 1975.

PHOTORECEPTOR OPTICS. Edited by A. W. Snyder and R. Menzel. Springer-Verlag New York, Inc., 1975.

GAUSS

MEN OF MATHEMATICS. Eric T. Bell. Simon & Schuster, Inc., 1937.

DISQUISITIONES ARITHMETICAE. Carl Friedrich Gauss, translated by Arthur A. Clarke. Yale University Press, 1966.

CARL FRIEDRICH GAUSS: A BIOGRAPHY. Tord Hall. The MIT Press, 1970.

WERKE. Carl Friedrich Gauss. G. Olms Verlag, Hildesheim–New York, 1973.

MATHEMATICAL GAMES

WHAT IS MATHEMATICS? Richard Courant and Herbert Robbins. Oxford University Press, Inc., 1941.

GENERALIZED "SANDWICH" THEOREMS. A. H. Stone and J. W. Tukey in *Duke Mathematical Journal,* Vol. 9, No. 2, pages 356–359; June, 1942.

THE AMATEUR SCIENTIST

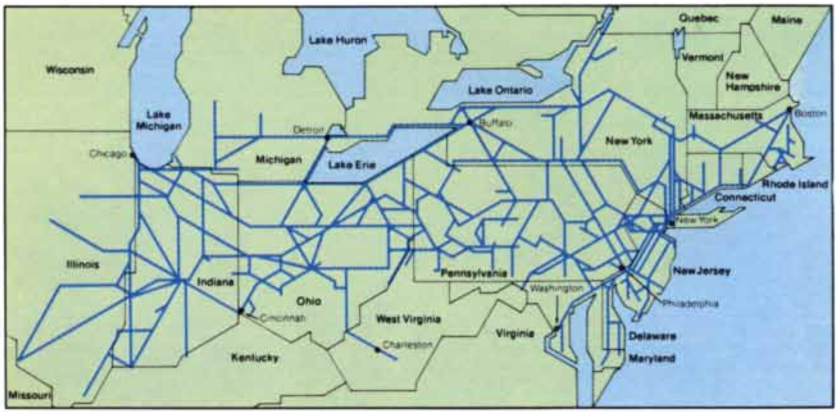
THE RAINBOW: FROM MYTH TO MATHEMATICS. Carl B. Boyer. Thomas Yoseloff, 1959.

MULTIPLE RAINBOWS FROM SINGLE DROPS OF WATER AND OTHER LIQUIDS. Jearl D. Walker in *American Journal of Physics,* Vol. 44, No. 5, pages 421–433; May, 1976.

How Conrail is building a better railroad for energy-hungry industries in the Northeast

Over half of America's manufacturing plants and factory workers are located in the bustling Northeast. Forty percent of these industries depend on energy that is produced by coal. And much of the coal is delivered by Conrail.®

Conrail has a formidable mission. On April 1, 1976, we began the task of consolidating six bankrupt, deteriorating railroads in the Northeast. Our goal: to physically rehabilitate the railroads, to improve service—in effect, to build an entirely new railroad. And to operate at a profit—projected by 1980.



Conrail network covers 16 states in the Northeast, the District of Columbia, two provinces in Canada. Area includes 55% of America's manufacturing plants, 60% of her factory workers.

A new approach brings results

In order to turn a profit, we're challenging the status quo at every turn. We're tackling hundreds of problems along the way. We're asking, "Is there a *better* way to do it?"

There is.

1. We've consolidated some of our accounting functions. Savings: about $9.5 million the first nine months alone.

2. We've consolidated our computer and communication services. Savings: $5.3 million a year.

3. We've got better ways to purchase and handle materials. Anticipated savings: $20-$30 million a year.

4. We want to increase the amount of time our freight cars are in use.

By making each car work harder, we can avoid buying thousands of new freight cars over the next 10 years. Savings: tens of millions of dollars.

5. We've got stricter quality control. On work performed and on purchased goods. In 1976, we rejected about $4 million worth of material that didn't measure up to our standards.

Rotary dumper dumps 100-ton coal car in just 2 minutes. This year, Conrail expects to deliver almost 1 million carloads of coal.

Vital to the Industrial Heartland of America

We're a key lifeline to the economy of 16 states and the District of Columbia.

We deliver raw materials and finished goods to the bustling Northeast which has nearly 55 percent of America's manufacturing plants and almost 60 percent of her factory workers.

The alternative

If Conrail should falter or fail, nationalization could become a reality. And that would cost Americans millions of dollars annually in extra taxes. No one wants that.

We've got to succeed.

We're *going* to succeed.

We've got a better way to run a railroad.

CONRAIL®

"A better way to run a railroad"

Consolidated Rail Corporation, Philadelphia, Pa.

Mr. Watt, until you took steam engine

Look where the science of energy is today.

Despite brilliant scientific advances, the problem of energy is where it's always been: Man is still trying to make the most of his sources — the sun, the wind, the atom, coal, gas, water, oil, and the heat of the earth.

Take the sun. Every 40 minutes, the earth receives enough solar energy to meet man's needs for a year. The trick still is to collect and store it at economically practical cost.

Consider coal. America's 400-year supply can be burned, or turned into gas and oil. Cleaning up coal's act and converting it at acceptable cost are problems yet to be solved.

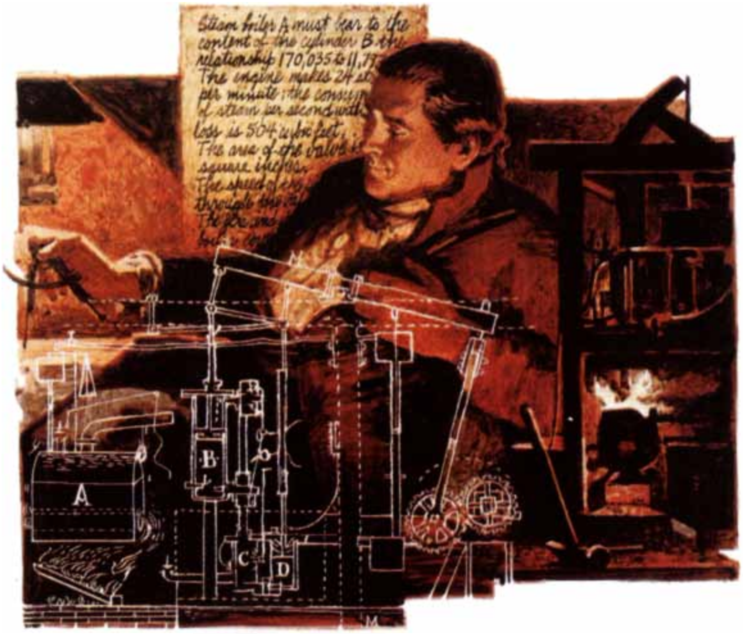
There are tough problems involved with the use of every form of energy. Rockwell is addressing a variety of these problems in a variety of ways through a variety of technologies.

Our annual report talks about those efforts. If you'd like a copy, just write Rockwell International, Department 815S, 600 Grant Street, Pittsburgh, PA 15219.

America is using 46% of its hydroelectric power potential. But even in this area, expansion and efficiency must grow side by side. Rockwell is building a computerized system that will help the Grand Coulee Dam make the most of its power potential by controlling the demand-output functions.

Rockwell valves standing up to 15 feet high and weighing up to 25 tons are in nearly every U.S. nuclear power plant. To slam the door against high-pressure steam in three or four seconds, our big valves must generate 1,000,000 pounds of thrust — as much as rockets did when they put America's first satellites into orbit.

Thomas Savery patented the first steam engine some 38 years before James Watt was born. But the early engines burned great quantities of coal wastefully, producing little power. It remained for Watt, while on a Sunday stroll through a Glasgow park, to figure out how to triple — yes, triple — the amount of work a coal-fired steam engine could do. Though not the father of the steam engine, Watt did become the father of a radically redesigned engine that made a critical stride in efficiency. He even figured out a way to measure the amount of work a steam engine could do. He called it *horsepower*.



that walk, the was just a toy.

Today, 2% of America's energy comes from nuclear power. By 2000 AD, that number must reach 25-30% through accelerated plant construction — or this nation could be in trouble. Rockwell equipment and technology are involved in today's plants — and in a new breeder that could produce as much electricity from 1.5 tons of uranium as today's plants produce from 150 tons.

According to American Petroleum Institute estimates, the U.S. has less than a 20-year supply of known oil and gas reserves. In present and future fields, you will find our McEvoy oil-field equipment helping to bring out those precious and diminishing energy reserves with a minimum of waste.

One Rockwell technology in development extracts oil from coal. Another extracts gas from coal. A third lets power plants burn high-sulphur coal without fouling the air.

Though solar-energy use is in its infancy, developing technologies will assure it an increasing role in man's future. Along with a variety of projects we've already launched, Rockwell is working with other companies on a solar electric power plant designed to serve a community of 10,000 people — for starters.

Rockwell International

...where science gets down to business


THE SEAGRAM'S GIN
AND TONIC SECRET.
Seagram's
Extra Dry
Gin
Use ice cubes made of tonic water.
Then add Seagram's Gin
and your usual amount of tonic.
Seagram's. The Perfect Martini Gin. Perfect all ways.