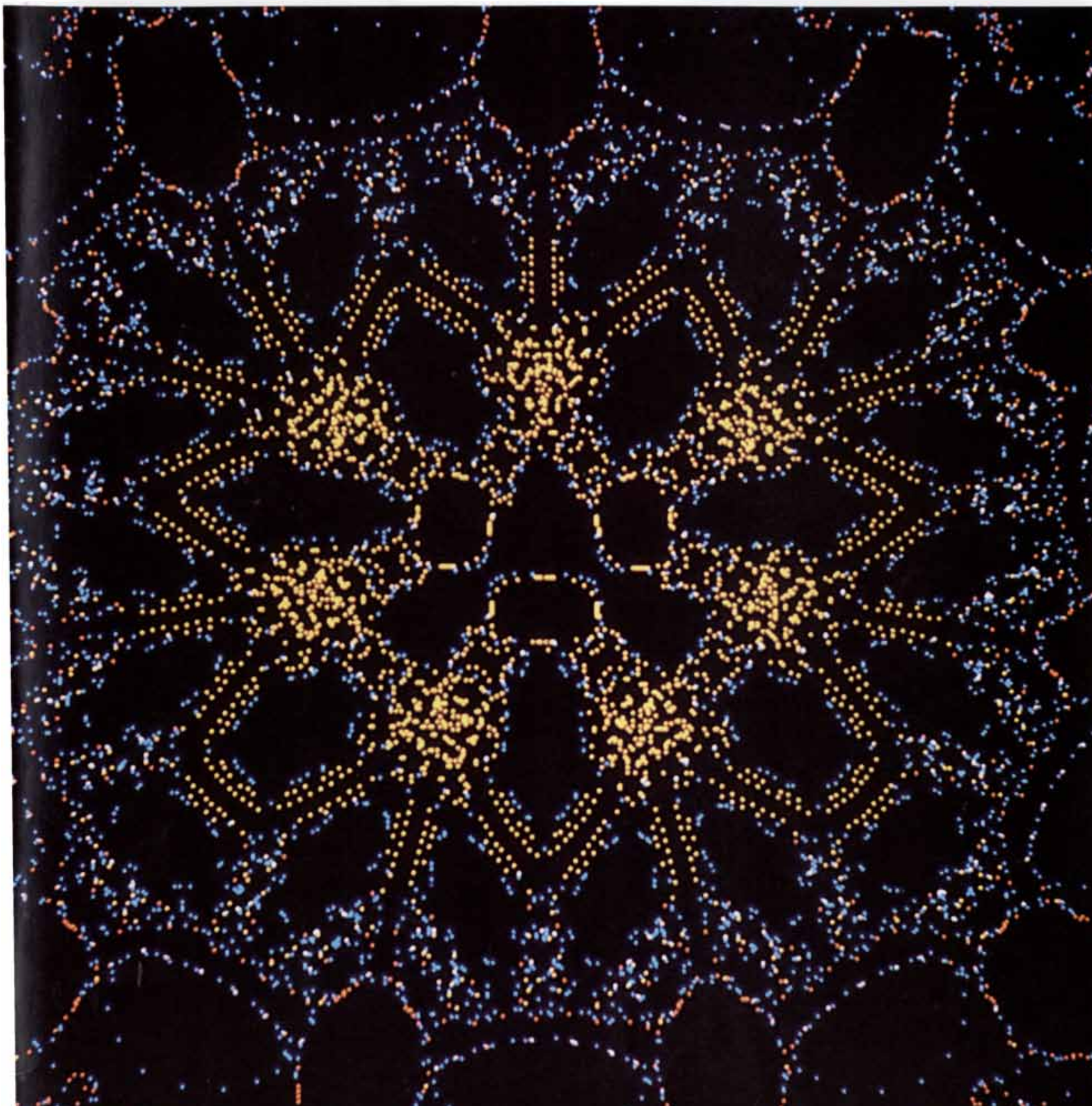
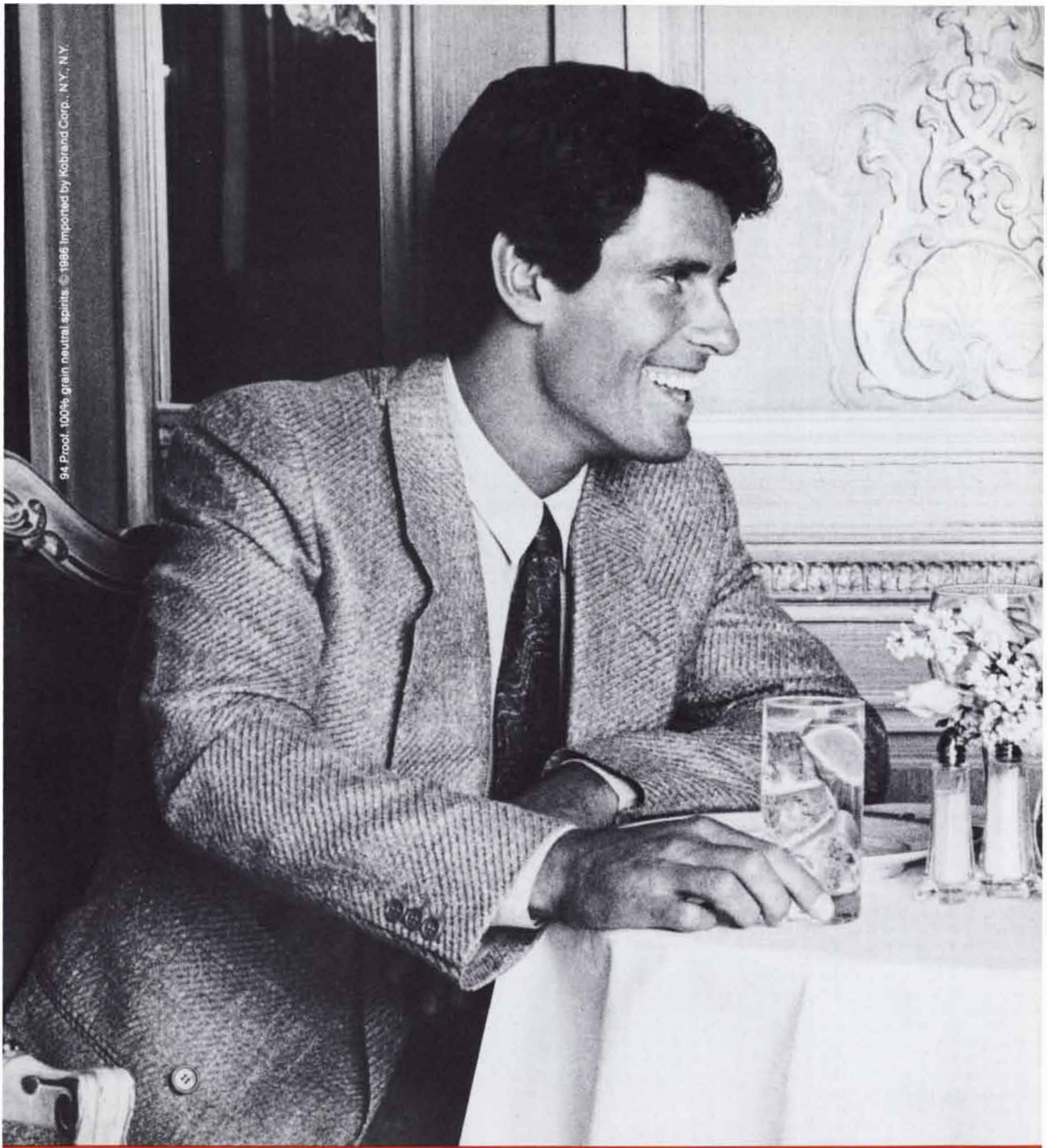
SCIENTIFIC AMERICAN



WALLPAPER FOR THE MIND

$2.50

September 1986

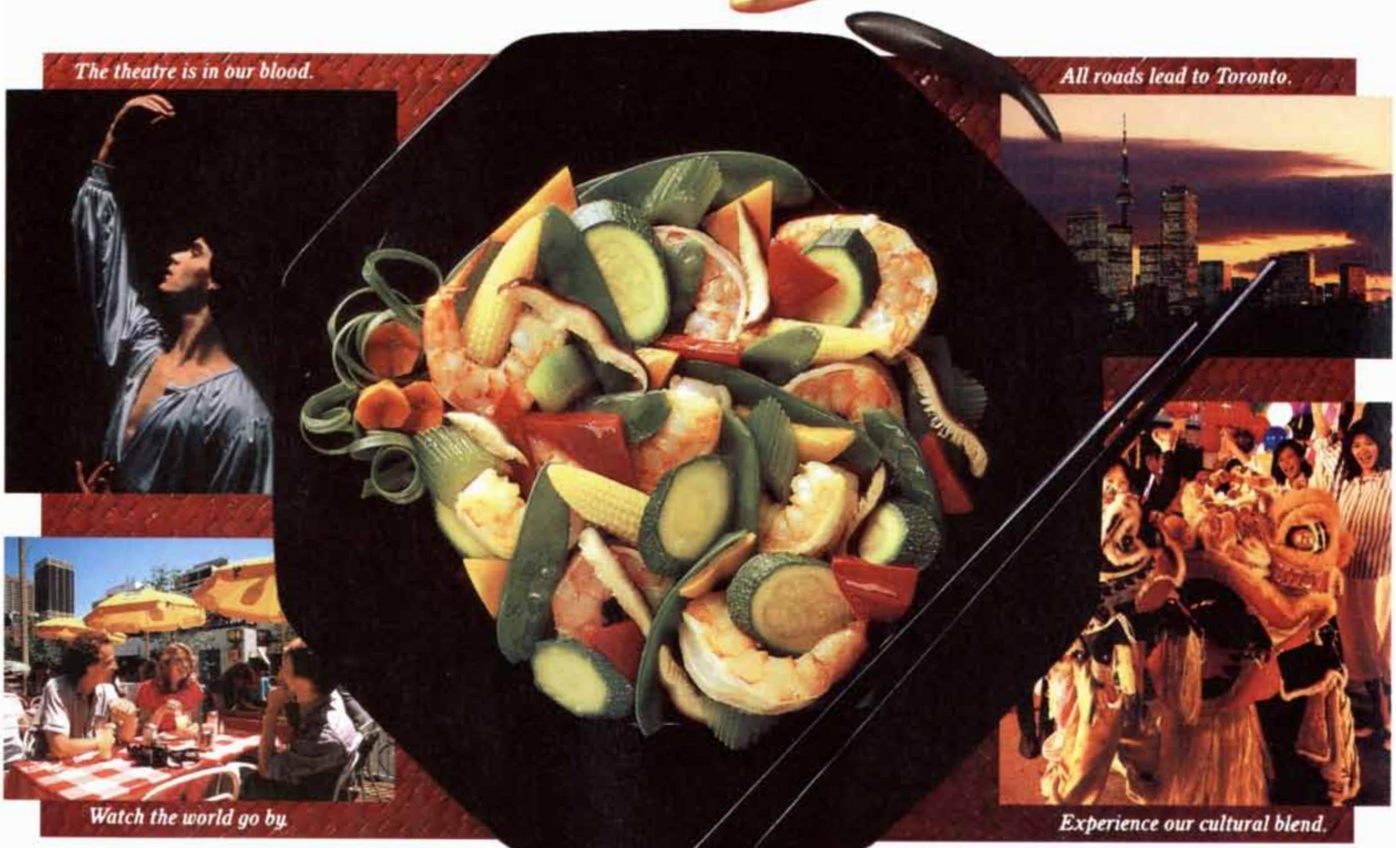

94 Proof. 100% grain neutral spirits © 1986 Imported by Kobrand Corp., N.Y., N.Y.

BEEFEATER
IMPORTED ENGLISH GIN®

The best of times deserve the best of taste.

VISIT ATHENS, ROME & BEIJING THE SAME DAY.

NOW THAT'S INCREDIBLE!



Only in Ontario can you find such an exquisite blending of the world's cultures. Each has brought something unique to enrich our province.

As a matter of fact, coming to Toronto, Ontario is literally like taking a world tour but a lot closer. With the great rate of exchange on the U.S. dollar, a vacation in Ontario is very affordable.

Come on up to Ontario, Canada and experience the very best that the world has to offer.

For more information on vacations in Ontario, call us TOLL FREE at 1-800-268-3735.

ONTARIO *Incredible!*

Established 1845 SCIENTIFIC AMERICAN September 1986 Volume 255 Number 3

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1986 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. 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: $24 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $33 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN
In Other Languages

LE SCIENZE
L. 3,500/copy L. 35,000/year L. 45,000/(abroad)
Editorial, subscription correspondence:
Le Scienze S.p.A., Via del Lauro, 14,
20121 Milano, Italy
Advertising correspondence:
Publietas, S.p.A., Via Cino de Duca, 5,
20122 Milano, Italy

サイエンス
Y880/copy Y9600/year Y13,000/(abroad)
Editorial, subscription, advertising correspondence:
Nikkei Science, Inc.
No. 9-5, 1-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

INVESTIGACION Y CIENCIA
450 Ptas/copy 4950Ptas/year $35/(abroad)
Editorial, subscription, advertising correspondence:
Prensa Cientifica S.A.,
Calabria, 235-239
08029 Barcelona, Spain

POUR LA SCIENCE
25FF/copy 245FF/year 290FF/year (abroad)
Editorial, subscription, advertising correspondence:
Pour la Science S.A.R.L.,
8, rue Férou,
75006 Paris, France

Spektrum DER WISSENSCHAFT
9.80 DM/copy 99 DM/year 112.20 DM/(abroad)
Editorial, subscription correspondence:
Spektrum der Wissenschaft GmbH & Co.
Moenchhofstrasse, 15
D-6900 Heidelberg,
Federal Republic of Germany
Advertising correspondence:
Gesellschaft Für Wirtschaftspublizistik
Kasernenstrasse 67
D-4000 Duesseldorf,
Federal Republic of Germany

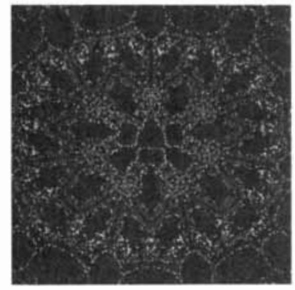
科学
1.40RMB/copy 16RMB/year $24/(abroad)
Editorial, subscription correspondence:
ISTIC-Chongqing Branch, P.O. Box 2104,
Chongqing, People's Republic of China

В МИРЕ НАУКИ
2R/copy 24R/year $70/(abroad)
Editorial correspondence:
MIR Publishers
2, Pervy Rizhsky Pereulok
129820 Moscow U.S.S.R.
Subscription correspondence:
Victor Kamkin, Inc.
12224 Parklawn Drive,
Rockville, MD 20852, USA

TUDOMÁNY
98Ft/copy 1,176Ft/year 2,100Ft/(abroad)
Editorial correspondence:
TUDOMÁNY
H-1536 Budapest, Pf 338
Hungary
Subscription correspondence:
"KULTURA"
H-3891 Budapest, Pf. 149
Hungary

العلوم
1KD/copy 10KD/year $40/(abroad)
Editorial, subscription, advertising correspondence:
MAJALLAT AL-OLOOM
P.O. BOX 20856 Safat,
13069 – Kuwait

Advertising correspondence all editions:
SCIENTIFIC AMERICAN, Inc.
415 Madison Avenue
New York, NY 10017
Telephone: (212) 754-0262 Telex: 236115



THE COVER

The computer-generated image on the cover is a graph of the successive values taken by an elementary mathematical function that is applied repeatedly to its own output (see "Computer Recreations," by A. K. Dewdney, page 14). The image illustrates some of the potential of a simple computer program for the creation of abstract and nonrepetitive patterns. The first point plotted is a seed value of the function, and it determines the value of the second point. Here the seed value is the origin. The second point then determines the value of the third point, and so on, according to the method described on page 15; the values of *a*, *b* and *c* are respectively −.3, .3 and .3. Colors are assigned to the points according to their numerical position in the plotting sequence. Many other schemes for generating computer images have been devised. The image was made by Barry Martin of Aston University in England on a Tektronix Type 4113 terminal, linked to a VAX 11/750 computer. The photograph was made by Barry Brookes, also of Aston.

THE ILLUSTRATIONS

Cover image courtesy of Barry Martin, Aston University

Page	Source
15	David Wiseman, University of Western Ontario
16–19	Barry Martin, Aston University
20	Johnny Johnson (*top*); Tony D. Smith, PICA Pty. Ltd., Essendon, Australia (*bottom*)
22	Tony D. Smith, PICA Pty. Ltd., Essendon, Australia
33	James Kilkelly
34	Edward Bell
35	Alan D. Iselin
36	Haralambos N. Kritikos, University of Pennsylvania
37–39	Alan D. Iselin
41–47	George Retseck
49–60	Hank Iken, Walken Graphics
75	P. A. Cancilla
76–77	Patricia J. Wynne
78	Thomas S. Reese and Morris J. Karnovsky (*left*), Thomas S. Reese and Milton W. Brightman (*right*), Patricia J. Wynne (*bottom*)
79–80	Patricia J. Wynne
81	Gary W. Goldstein and A. Lorris Betz, University of Michigan Medical Center
82	Edward A. Neuwelt
85	Bradford Howland and Denise D. Denton, Massachusetts Institute of Technology
86	Andrew Christie
87–90	Bradford Howland and Denise D. Denton
92–99	Ian Worpole
101	Doug Allan, Animals Animals (*top*); R. J. Hill, New Zealand Oceanographic Institute (*bottom*)
102–106	Tom Prentiss
109	Pat Canganelli
110	Bayerische Staatsbibliothek, Munich
111–113	Pat Canganelli
114–119	Michael Goodman

WHY WE SPONSOR A SHOW THAT DOESN'T TREAT US LIKE A SPONSOR.

AT&T is the national corporate underwriter of The MacNeil/Lehrer NewsHour, but we don't get any special treatment. We don't expect to.

When Robert MacNeil, Jim Lehrer, and their team of correspondents report any kind of news about us, they don't pull any punches. Nor should they.

At AT&T, we're less interested in being handled with kid gloves than we are in the program's objective approach to the news and its uncompromising commitment to quality.

The same commitment to quality we at AT&T bring to our products and services.

All of which led us to support The MacNeil/Lehrer NewsHour in the first place.

AT&T

The right choice.

LETTERS

To the Editors:

In "Chernobyl, U.S.A.?" ["Science and the Citizen," SCIENTIFIC AMERICAN, July] Thomas B. Cochran, a scientist at the Natural Resources Defense Council, is given as the source of the information that "hydrogen was released during the 1979 accident at the Three Mile Island-2 reactor near Harrisburg, Pa. There was no chemical explosion because little or no free oxygen was present."

According to an article that was published in the 1985 *Annual Review of Energy* by Richard S. Denning of the Battelle Memorial Institute's laboratory in Columbus, substantial amounts of hydrogen were released by the reaction of steam with the zirconium-alloy tubes containing the reactor fuel; the hydrogen subsequently exploded. Denning states: "Hydrogen produced while the core was uncovered was released to the containment [building]...rapid combustion of the hydrogen (deflagration) occurred in the containment building, rapidly increasing the containment pressure.... The sound of the explosion was heard in the control room, but not until a day later was the significance of the event fully recognized."

Your article gives the impression that because there was no hydrogen explosion in the reactor vessel, there was no such explosion during the Three Mile Island accident. On the contrary: there was a significant hydrogen explosion, whose severity can be judged from Denning's description: "Approximately 50 percent of the zirconium in the core was oxidized in the accident, resulting in the release of hydrogen to the containment atmosphere. It is estimated that approximately 700 pounds of hydrogen were consumed in the combustion event, introducing an overpressure of 28 pounds per square inch gage on the containment. Although the resulting pressure was close to the design pressure of the containment, it was well below the containment's expected failure pressure."

This aspect of the Three Mile Island accident does not seem to be well known. In fact, your observation that "there is little reason to believe U.S. reactors are immune to hydrogen-oxygen explosions" would seem to be rather understated.

LARRY CLIFFORD

Lake Hiawatha, N.J.

To the Editors:

"Modern Windmills," by Peter M. Moretti and Louis V. Divone [SCIENTIFIC AMERICAN, June], is a very interesting article about ancient and modern windmills. I believe, however, that the authors erred in assuming that "the old mills, pumping water and grinding grain, had to start up under load, and to do so they required a strong torque, or twisting force, from the blades in order to turn the grindstones or lift the water."

Actually the only power old grain mills required to start was the power needed to accelerate the moving parts. The millstones in a grain mill were separated by an adjustable gap, and as long as no grain was milled no work was being done.

Once the mill was turning, the miller gauged its speed by counting the tips of the blades as they whizzed by. "Eighty ends" (20 revolutions per minute) was about the allowable top speed; at higher speeds the danger of the blades' being thrown off became quite real.

As the wind picked up, the miller could increase the amount of work being done, and thereby control the speed of the mill, by adjusting the gap between the stones and changing the flow of grain. If the wind became too strong, furling the sails and turning the head of the mill more or less out of the wind were the next steps. It hurt the pride of the old Dutch millers to stop a mill because of a strong wind, and it was always a game to see who could keep his mill going the longest. Millers whose mills pumped water or pressed oil always lost that game, since they could not control the top speed of the mill by making it do more work.

CORNELIS LANGEWIS

Walnut Creek, Calif.

To the Editors:

Mr. Langewis is correct in pointing out our oversimplification of a complex issue. In some kinds of mills, unlike unattended American water pumpers, the load can indeed be reduced for startup. Historical wood windmills, however, had to contend with a great deal of friction in bearings and mechanisms, and so the minimum load was still a substantial fraction of the maximum load. A paddle-wheel rotor or a high-solidity propeller can handle such heavy starting loads. A high-speed, low-solidity turbine, in contrast, has low starting torque; indeed, a Darrieus rotor yields zero torque at rest.

We enjoyed Langewis' description of the work of a Dutch miller. We might add that the task of furling the sails entailed stopping the mill with a blade straight down so that the miller could reef the cloth from the ground or from a gallery. If the miller had not foreseen a rising wind and had waited too long, this could be a difficult task! In a storm the brake could prove insufficient to bring the mill to a stop and would then overheat. There are records of a number of Dutch windmills catching fire as a result and burning down.

Another reader asked us to mention that the ancient paddle-wheel mill design described in our article is still used in some remote areas. He observed mills of that type while he was traveling in 1971 near Herat in Afghanistan.

PETER M. MORETTI

LOUIS V. DIVONE

Oklahoma State University
Stillwater

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

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name

New Address

Street

City

State and ZIP

Old Address

Street

City

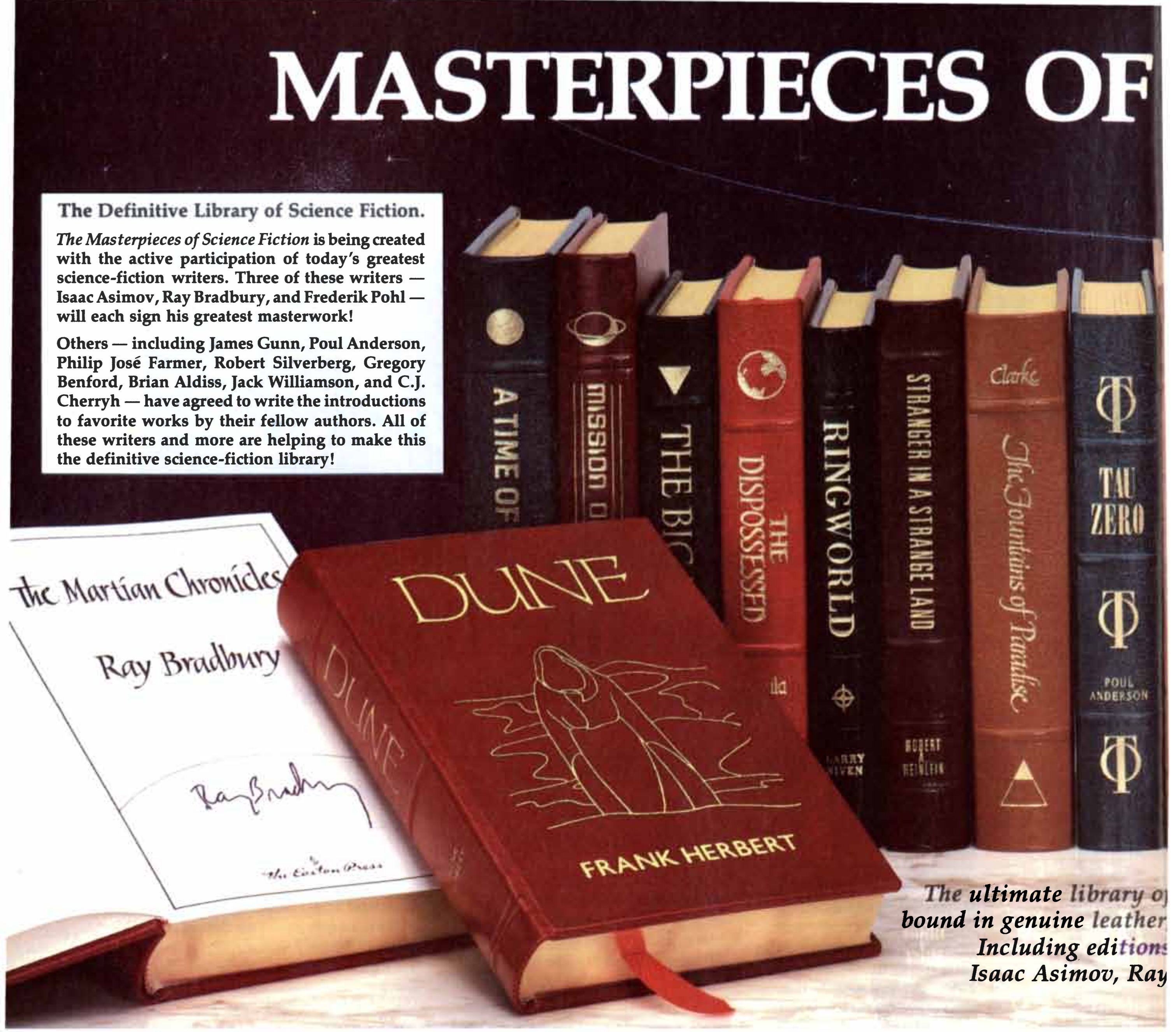
State and ZIP

MASTERPIECES OF

The Definitive Library of Science Fiction.

***The Masterpieces of Science Fiction* is being created with the active participation of today's greatest science-fiction writers. Three of these writers — Isaac Asimov, Ray Bradbury, and Frederik Pohl — will each sign his greatest masterwork!**

Others — including James Gunn, Poul Anderson, Philip José Farmer, Robert Silverberg, Gregory Benford, Brian Aldiss, Jack Williamson, and C.J. Cherryh — have agreed to write the introductions to favorite works by their fellow authors. All of these writers and more are helping to make this the definitive science-fiction library!



***The ultimate** library o*
***bound in genuine** leather*
***Including** editions*
Isaac Asimov, Ray

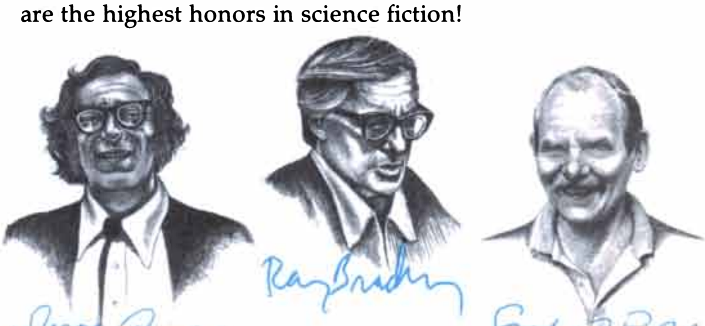
Behold ... the ultimate library of science fiction! For the first time, the greatest masterpieces by the greatest science-fiction writers will be issued in luxurious leather-bound editions.

Timeless, towering works of the imagination, that invite you to explore new worlds, encounter strange phenomena, entertain new possibilities. For the first time all these masterpieces will appear in fine collector's editions. Permanent, leather-bound editions to be treasured ... and to be read again and again — truly a collection that has been long overdue.

A complete collection — from the Originators to the most recent Award winners!

The Masterpieces of Science Fiction will comprise the entire history of the genre. The library will include heirloom editions of works by Jules Verne, H.G. Wells, Edgar Rice Burroughs, and Sir Arthur Conan Doyle. Superb editions from the inventors of science fiction, classics from the 1940's and 1950's, as well as many recent award winners.

There will be splendid editions of Arthur C. Clarke's *2001: A Space Odyssey* ... Fred Hoyle's *The Black Cloud* ... Robert A. Heinlein's *Stranger in a Strange Land* ... Ursula K. Le ‹ *Dispossessed* ... John Brunner's *Stand on Zanzibar* ... Sam Delany's *The Einstein Intersection* ... Frank Herbert's *Dune* ... and Roger Zelazny's *This Immortal*. Many of the works included have received the Hugo Award, the Nebula Award ... or both! These are the highest honors in science fiction!



Personally signed editions of today's most celebrated science-fiction classics!

Recognizing the importance of this collection, three of today's greatest science-fiction writers will hand-sign their greatest for each subscriber. Isaac Asimov will sign

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

SEPTEMBER, 1936: "The past twelve months and the next to come may well be marked in science texts of the future as the crucial period in the development of cosmic ray theory. Within the past year the controversy over the nature of cosmic rays has resolved itself into a more or less general agreement. At least the greater part of the original ray consists of charged particles. There are three chief components: alpha particles, found only high in the stratosphere, electrons and positrons, which are more penetrating, and protons, the most penetrating of all. Where the rays come from and how they acquire their great energy are questions for the future."

"A unique method of 'finger-printing' chemical substances through the use of X-ray-diffraction patterns has been worked out by the Dow Chemical Company. The molecules of any given compound are arranged in a definite geometrical structure, which is a characteristic of this compound only. Equally characteristic of the given compound is the photographic pattern that results from the interaction of X-rays with this definite geometrical structure. The pattern consists of a sequence of lines of definite positions and intensity. It is these lines that give the key to accurate classification."

"Chief among the external aids that the modern archeologist has called to his assistance is air photography. Photographs taken from a vertical point of view will reveal all the markings left on the soil by man's agency. Photographs taken obliquely in a slanting sun will show, even in the growing crops, the outlines of long-vanished foundations and the darker patches left by disturbances of the soil."

"The crowded contents of a large library can now be duplicated on a microscopic scale and yet rendered optically available when desired. By means of micro-photography an original full-size page can be reduced from 150 to 400 times, registered well-nigh imperishably on film and then encased in sealed cartridges ready to be inserted in a suitable magnifying reading machine. A pound of film will carry micro-copies that are the equivalent of a ton of printed matter."

"The deep-seated conviction of almost every red-blooded American motorist that other drivers are less competent can now be exploded. The Bureau of Street Traffic Research at Harvard University has developed a comprehensive group of instruments and tests for measuring driving ability. With the means for grading drivers, the day is not far off when the hopelessly unfit will be taken from behind the wheel, the average driver provided with aids toward increasing his skill and the superior one perhaps given certain privileges as long as they are respected."

"Light from practically all scenes commonly viewed contains a large proportion of polarized light. With polarizing glasses it is possible for the polarized surface light to be prevented from reaching the eye, whereas the light carrying the underlying detail is not polarized and is readily transmitted to the eye. In other words, the glare reflected from surfaces viewed is eliminated."

SCIENTIFIC AMERICAN

SEPTEMBER, 1886: "One of the most remarkable and also interesting events pertaining to Egyptology was the recent unrolling of the mummy of the ancient monarch Rameses II. The unrolling took place at Boulak under the direction of Prof. Gaston Maspero, Director-General of the Excavations of the Antiquities of Egypt, by order of and in the presence of the Khedive of Egypt and a large company of officials and learned men from various countries. The corpse is that of an old man, but of a vigorous and robust old man. We know, indeed, that Rameses II reigned for 67 years, and that he must have been nearly 100 years old when he died."

"At the recent convention of the National Electric Light Association no little time was occupied in a discussion of the expediency of burying the wires. The sense of the convention was decidedly opposed to the project at the present time. When we consider the fact that air is the best insulation and the ground the worst, it is scarcely reasonable to look for an expeditious and easy conquest in the struggle for a similar service underground as the public has become accustomed to receive over the aerial lines. To put all the wires in the metropolis underground is a great and costly undertaking, and to proceed with it without the most conclusive evidence of the practicability of the means employed would be hazardous, to say the least."

"We have several times alluded to the fact that it is possible to obtain fair negatives by arranging a sensitive dry plate in one end of a suitable box, while in the center of the opposite end was a fine needle hole through a thin piece of metal attached to the outer surface of the box. Practically this idea has just been carried out in a small camera recently put upon the market, which, for its compactness, simplicity and novelty, will be likely to lead a great many, young and old, into taking up photography as a pastime."

"At the Baldwin Locomotive Works there are in course of construction four locomotives, designed to be run by soda, which takes the place of fire under the boiler. Inside the boiler will be placed five tons of soda, which, upon being dampened by a jet of steam, produces an intense heat. When the soda is thoroughly saturated (in about six hours), the action ceases, and then it is necessary to restore the fuel to its original state by forcing through the boiler a stream of superheated steam from a stationary boiler. That drives the moisture entirely from the soda, when it is again ready for use. The engines, which are to be run on the streets of Minneapolis, will readily draw four light cars."

"Further experiments have been made to determine the depth to which light penetrates the water of lakes and seas. The place was the Gulf of Nice in water about 550 meters deep. During April the limit of the penetration about midday, in fine weather, was found to be about 400 meters. There is a penetration of about 300 meters all the time the sun is above the horizon, and of 350 meters during eight hours of the day."

"An exhibit at the National Museum shows what a 154-pound man appears like from the chemist's point of view. The suppositious man has been divided into his ultimate elements. There stand all these elements and chemical compounds in glass jars, properly labeled. After the gases, the carbon, the phosphorus and the sulphur have been extracted from the man, there is nothing left of him but metals. One would be surprised to look into this case and see how much a man is weighted down with various metallic substances."

TANDY®... Better Again™

Welcome to the next generation of personal computing.

Introducing four Tandy PC compatibles that redefine the term "industry standard"

Meet our new line of MS-DOS® based personal computers. Each one was designed with a different user in mind. And yet each has one thing in common: the Tandy commitment to excellence. When you buy a Tandy computer, you get more. More power, more speed, more options, more support. All at a lower price than the so-called "industry standard."

The Tandy 1000 EX

Only $799 for a true MS-DOS based computer? It's true. The Tandy 1000 EX runs the same powerful software as an IBM® PC—at up to twice the speed! And every Tandy 1000 EX comes with our new graphics-oriented Personal DeskMate™ software with six user applications. The 1000 EX is also the perfect classroom computer.

The Tandy 1000 SX

Here's a high-performance version of America's #1 PC compatible—the Tandy 1000. The new Tandy 1000 SX is still compatible with industry-standard MS-DOS software. But it can now run many of these programs 50% faster. The 1000 SX comes with 384K and two built-in disk drives—standard. You even get our enhanced six-in-one DeskMate® II applications software. All for just $1199.

The Tandy 3000 HL

If you've already invested in IBM PC/XT® software and hardware, and would like to expand, the Tandy 3000 HL gives you four times the power of the XT . . . for less. The 3000 HL can process twice the data, thanks to its advanced, 16-bit 80286 microprocessor. Plus, the 3000 HL operates at twice the speed of the XT. With a Tandy 3000 HL, you're getting much more than an XT compatible—and for only $1699.

The Tandy 3000 HD

Unmatched in performance, the Tandy 3000 HD is a powerful workstation with both networking and multiuser capabilities. It's compatible with the IBM PC/AT®, offers greater disk storage (40 megabytes), yet is priced at $4299.

American-made quality

These four new computers are built in our own manufacturing plants right here in the U.S.A. They're designed to meet our rigid quality standards, and backed with the best support in the industry. We're proud of our new line, and we want you to see why. Come in for a demonstration.

Available at over 1200 Radio Shack Computer Centers and at participating Radio Shack stores and dealers.

Radio Shack®
COMPUTER CENTERS
A DIVISION OF TANDY CORPORATION

Send me an RSC-17 Computer Catalog

Radio Shack, Dept. 87-A-294
300 One Tandy Center, Fort Worth, TX 76102

Name ____________________
Company ____________________
Address ____________________
City ____________________
State ____________ ZIP ____________
Phone ____________________

Prices apply at RadioShack Computer Centers and participating stores and dealers. Monitors not included. IBM, IBM PC/XT and IBM PC/AT/Registered TM International Business Machines Corp. MS-DOS/Registered TM Microsoft Corp.

THE AUTHORS

KENNETH R. FOSTER and ARTHUR W. GUY ("The Microwave Problem") are interested in the interaction of electromagnetic fields and living tissue. Foster received a Ph.D. in physics from Indiana University in 1971. He began his research on the biological effects of microwaves while serving in the U.S. Navy and has pursued it since 1976 as associate professor of bioengineering at the University of Pennsylvania. Guy has been at the University of Washington since his undergraduate days, with the exception of seven years spent doing research on antennas for the Boeing Aerospace Company. He is now professor of bioengineering and director of the bioelectromagnetics laboratory. From 1970 to 1982 he was chairman of the American National Standards Institute committee that wrote guidelines for human exposure to microwaves. One of his hobbies is operating a ham radio, which he uses for a daily "rag chew" with his father in Montana.

DENNIS H. ROUVRAY ("Predicting Chemistry from Topology") is associate research scientist in the chemistry department of the University of Georgia. Trained as a chemist at the Imperial College of Science and Technology in London, he became interested in applying topology and graph theory to chemistry while on sabbatical in 1970 at the Mathematical Institute of the University of Oxford. Mathematical chemistry was not then an organized discipline; now, he reports, some 500 papers a year are published on the subject, and he has recently founded a new journal dedicated to it. Rouvray held positions in Africa, Asia and Europe before going to Georgia in 1984.

MICHAEL B. GREEN ("Superstrings") is professor of physics at Queen Mary College of the University of London. His interest in string theory began at the Institute for Advanced Study in Princeton, where he spent two years as a fellow after getting his Ph.D. from the University of Cambridge in 1970. He returned to Cambridge in 1972 to work at the Cavendish Laboratory, went to Oxford in 1977 and then to London in 1978. He is a frequent visitor to the California Institute of Technology, where he collaborates with John H. Schwarz, an early pioneer of string theory.

GARY W. GOLDSTEIN and A. LORRIS BETZ ("The Blood-Brain Barrier") began working together in 1977, when Betz was a postdoctoral fellow in Goldstein's laboratory at the University of California at San Francisco Medical Center. In 1979 they both moved to the University of Michigan Medical Center, where Goldstein is now professor and Betz is associate professor in the departments of pediatrics and neurology. Goldstein got his B.A. and M.D. from the University of Chicago; he was a resident in pediatrics and neurology at several university hospitals before taking a faculty position at San Francisco in 1973. Betz has a B.S., an M.D. and a Ph.D. in physiology and biochemistry from the University of Wisconsin at Madison.

PHILIPPE BROU, THOMAS R. SCIASCIA, LYNETTE LINDEN and JEROME Y. LETTVIN ("The Colors of Things") collaborate under Lettvin's leadership at the Massachusetts Institute of Technology. Brou received a Ph.D. in computerized vision from M.I.T. and now works in the Artificial Intelligence Laboratory on problems in graphics, financial analysis, manufacturing and oil drilling. Sciascia is staff neurologist at the Veterans Administration Court Street Clinic in Boston and instructor in neurology at the Boston University School of Medicine. He got his B.S. in biology at M.I.T. and his M.D. from Columbia University's College of Physicians and Surgeons. Linden is assistant professor of manufacturing engineering at Boston University; she has also taught at M.I.T., and at both institutions she has received awards for the excellence of her teaching. Lettvin himself embodies the diversity of his group: he has an M.D. from the University of Illinois and has practiced as a psychiatrist, but for the past 35 years he has been affiliated with the electronics laboratory at M.I.T., and for the past 20 he has been professor in both the biology department and the department of electrical engineering and computer science. He has done extensive research on the transmission of nerve impulses in the brain.

FRANCESCO PARESCE and STUART BOWYER ("The Sun and the Interstellar Medium") share an interest in ultraviolet astronomy. Paresce is senior astronomer at the Space Telescope Science Institute in Baltimore. In 1968 he left his native Italy to attend the University of California at Berkeley. After obtaining his Ph.D. he stayed on as a research astronomer until he moved to Baltimore in 1983. At Berkeley he worked with Bowyer, who is professor of astronomy. Bowyer got his Ph.D. at Catholic University with a thesis in X-ray astronomy. In the late 1960's he started developing instruments for extreme-ultraviolet observations. At the time, he writes, "no astronomy had been carried out at these wavelengths. Hence I started a program to identify diffuse sources. It turns out that one of the strongest components of the diffuse background is provided by inflowing interstellar helium—the topic of this article." Bowyer is the principal investigator for a NASA spacecraft that will survey the extreme-ultraviolet sky.

JOYCE R. RICHARDSON ("Brachiopods") became interested in brachiopods while at the New Zealand Oceanographic Institute. "The brachiopods were of particular interest," she writes, "because they provide the best record (in length, continuity, abundance, detail) of life on the earth." She is now resident scholar at the Museum of Victoria in Melbourne, but she has retained an attachment to the fjords of New Zealand: "My free time has been spent in planning a commercial enterprise—a project that my partner (an engineer) and I hope will protect a special environment by making it self-supporting. We have produced plans to construct an underwater observatory (to a depth of 80 feet) within Milford Sound as a resource for tourists (to view the astonishing range of life within a sheltered fjord basin) and scientists." Richardson has an M.A. in zoology from the University of New Zealand and a Ph.D. in geology from the University of Melbourne.

VERNARD FOLEY and WERNER SOEDEL ("Leonardo's Contributions to Theoretical Mechanics") have often pooled their scholarly resources—Foley is associate professor of history and Soedel is professor of mechanical engineering at Purdue University—to study topics in the history of technology. They have written three earlier articles for SCIENTIFIC AMERICAN: "Ancient Catapults" (March, 1979), "Ancient Oared Warships" (April, 1981) and "The Crossbow" (January, 1985). Foley's Ph.D. (from the University of California at Berkeley) was in European history with an emphasis on science and technology. Soedel was born in Prague and belonged to the Sudeten German minority that was evicted to Germany in 1945. There he took a degree in engineering and worked as an automotive engineer before coming to the U.S. in 1963. He received his Ph.D. from Purdue in 1967.

Q. Would you buy a pair of shoes from this man?

A. Thousands of Lands' End customers do, for reasons this interview makes clear.

Above is a rare photo of Jim Jennings, the "old shoe dog" who governs the buying and selling of Lands' End shoes. Rare, because Jim is seldom caught in repose. But Carol Sadtler, one of our resourceful creative people, cornered him one day, and what follows are excerpts from an interview that may lead you to rethink whatever prejudice you may have against ordering shoes from a catalog. Even ours.

Carol: *How long have you been in the business, Jim? And how does that affect your buying shoes for your customers?*

Jim: Let's see... about 20 years, I guess. When you've been around that long, you see a lot of companies come and go. You get to know the good, reliable ones. The ones that make the Lands' End kind of shoe.

Carol: *And what kind of shoe is that?*

Jim: The kind of shoe that goes with the clothing we sell. Not "high fashion", or a "hot seller". But classic. We offer casual shoes, as you know, and a very controlled line of dress shoes. Always in style. And as well made as possible.

Carol: *What steps do you take to get that kind of quality?*

Jim: Extra steps, frankly. When we look at a shoe, we'll see what we can add to it to give it more quality. Like a leather lining instead of vinyl, so the shoe wears longer and breathes better. Or a wool lining instead of acrylic in a pair of slippers. Or we'll add a better sole.

(JIM PAUSES, RUMMAGING AROUND IN SOME SHOE BOXES AND COMES UP WITH A PAIR OF LANDS' END "DUSTY BUCKS".)

Here's a perfect example. We took these traditional Bucks. Nice suede leather, but the rubber soles wore out in a big hurry. And they were heavy. So we added a lightweight Vibram® sole, with a real tough skin to make it long-lasting. Now, instead of a Buck, we give you a Buck-and-a-quarter.

Carol: *They warned me about your puns, Jim. Let's get serious again. I'm told you visit manufacturers quite often. Right? What do you look for?*

Jim: Actually, as a company we probably do make more factory visits than most. That's really the only way to maintain quality. You just can't run our business from an "ivory tower". And when we do visit, we look to see how things are organized. And at how many points the shoes are inspected. After a while, you can tell how neat and careful the work is at every step.

Carol: *Do you literally look over a handsewer's shoulder?*

Jim: You bet. You look to see how neat and even the stitches are.

Carol: *That kind of quality control must take a lot of travel.*

Jim: It does. But we're fortunate in having built up a very strong domestic structure of suppliers. Which means for the customer that we can keep an eagle eye on quality, and keep the shoes they want in stock too.

LANDS' END
DIRECT MERCHANTS

Carol: *But you buy shoes overseas, too?*

Jim: Yes, indeed. We go where we can get the best shoe. Italy, for example, for the fine leathers and excellent craftsmanship for some of our dressier shoes.

Carol: *How do you do all these extra things and still keep prices as low as they are?*

Jim: We have several things going for us. We don't operate on the normal 50% retail mark-up. We're direct merchants, with no middlemen to raise costs. Also, we order in large quantities. And we have long-standing relationships with our suppliers, which helps.

Carol: *What would you say to the person who hesitates to order shoes from a catalog because of fit?*

Jim: We make an extra effort to standardize sizes within our offerings. We've used the same lasts—which determine fit—for ten years or more in many of our shoes. And we fit-test a shoe on 20 to 25 people before we offer it.

Carol: *Gee, you seem to have things all worked out. Don't you have any problems?*

(LONG PAUSE. JIM LAUGHS A LITTLE NERVOUSLY. FINALLY, HE ANSWERS.)

Jim: Frankly, the hardest thing to do is managing the variety of sizes and colors we offer in each style. We try our darndest to fit everybody, and that means we have to keep track of something like 1863 offerings.

Carol: *That ought to keep you busy. And your customers happy.*

Jim: Well, Carol... that's the idea!

* * * * *

It is, indeed. Whether it's Lands' End shoes you're buying, or any one of our other catalog items. All quality offerings, all priced to reflect remarkable value, all GUARANTEED. PERIOD.

Call our toll-free number for a free catalog. Or send in the coupon below. We'll tell Jim Jennings he can expect to hear from you. Okay?

Please send free catalog.
Lands' End Dept. Q-71
Dodgeville, WI 53595

Name ____________________

Address ____________________

City ____________________

State __________ Zip __________

Or call Toll-free:

800-356-4444

COMPUTER RECREATIONS

Wallpaper for the mind: computer images that are almost, but not quite, repetitive

by A. K. Dewdney

Ordinary wallpaper is printed by a rotating cylinder engraved with a design. As the cylinder turns it prints the same design over and over again. Only a computer, however, can reproduce certain richly embroidered patterns I call wallpaper for the mind. These patterns do not repeat themselves, at least not exactly; instead each pattern continually manifests itself in new contexts and configurations, left and right, up and down. From one apparition to the next, what is changed and what is preserved?

The color swatches in my current sample book demonstrate the results of three widely differing techniques. The computer programs responsible for the images range in difficulty from the extremely simple to the merely easy. They come from three readers: John E. Connett of the University of Minnesota, Barry Martin of Birmingham, England, and Tony D. Smith of Essendon, Australia.

Connett's program is based on the circle, but it celebrates the varieties of design based on the square. The apparent conundrum compels me to name it CIRCLE^2. In a nutshell, it applies the formula for the graph of a circle, $x^2 + y^2$, to assign a color to the point that has the coordinates x and y. I shall give the details below. In the meantime you may be boggled, as I was, to discover that there is much more to this wallpaper than a set of concentric circles; specifically, if you back away from the wall, intricate patterns of delicate squares may also emerge [*see illustrations on opposite page*]. There are mysteries here.

It is perhaps not surprising that CIRCLE^2 was inspired by the Mandelbrot set, which I described here in August, 1985. The set is the discovery of Benoit B. Mandelbrot of the IBM Thomas J. Watson Research Center; the riot of form and color that surrounds the Mandelbrot set is based on a single mathematical function, applied repeatedly to its own output for each complex number in a region of the plane. Whenever the iterated value of the function reaches the magnitude 2, the number of iterations needed to reach that magnitude determines the color of the corresponding point.

Connett, who had no color monitor at his disposal, assigned black to points that reach 2 after an even number of iterations and white to points that reach 2 after an odd number of iterations. Creditable images of the Mandelbrot set appeared, but Connett was led to explore other formulas. He selected the formula $x^2 + y^2$ and discarded iteration altogether. His program systematically scans a gridded section of the plane; at each point (x,y) it computes the formula and then truncates the resulting value to an integer. If the integer is even, the point (x,y) is colored black; if it is odd, the point is colored white (left blank).

I fear I have just lost half of my audience. They already understand the program, and they have rushed off to type it into a nearby computer. It is that simple. In algorithmic shorthand CIRCLE^2 is an input section followed by a double loop:

```
input corna, cornb
input side

for i ← 1 to 100
    for j ← 1 to 100
        x ← corna + (side × i/100)
        y ← cornb + (side × j/100)
        z ← x² + y²
        c ← int(z)
        if c is even, then plot (i,j)
```

First the program causes the computer to call for the coordinates (*corna,cornb*) for the lower left corner of the square to be examined. The variable *side* is the length of the side of the square of interest. For example, if the user types −15 and −20 for the corner coordinates and 87 for the side, the program plots a 100-by-100 array of points within a square region of the plane 87 units on its side, whose lower left corner is the point (−15,−20). In my outline of the program I have assumed the iteration limits run from 1 to 100, but they must be adjusted so that the square lies within the boundaries of the output device to be used. On my monitor these limits outline a smallish square on the screen.

The double loop marches through the square grid and for each index pair (i,j) computes the coordinates of the point (x,y) to which the pair corresponds. The loop then squares x and y, assigns the sum of the two squares to the variable z and truncates the sum. The largest integer less than or equal to the sum is stored as the variable c. If c is divisible by 2, the point (x,y) is plotted, presumably as a colored pixel on a monitor or as a black dot in printed output. If c is odd, no point is plotted.

Readers who want to re-create Connett's wallpaper need not get too anxious about whether they have the colors right. Most patterns are equally striking with the colors reversed. Indeed, even more than two colors are possible: instead of determining whether c is even or odd, divide it by the number of colors you want. The remainder after the division can then be assigned to a distinct color. For example, two, three and four colors were chosen to generate respectively the top, middle and bottom images in the illustration on the opposite page.

The smaller the square under examination is, the closer the plane appears to the viewer and the greater the magnification of CIRCLE^2's image is. Unlike the procedure for coloring neighborhoods of the Mandelbrot set, however, Connett's program does not yield an infinite regress of progressively smaller patterns. At high magnifications about the origin (0,0) a set of concentric circles appears. At still higher magnifications there is a large black disk in the middle of the screen: the truncated-integer value of every point in the disk is zero. Then all the screen is blackness.

You can better appreciate the beauty of Connett's wallpaper by reducing the magnification: back away from the wall. The concentric circles dissolve into an intriguing arrangement of primary and secondary circles resembling a moiré pattern. New wallpaper designs appear like magic, seemingly different with every lower magnification. Is there an infinite regress lurking here? It seems a vexing question, but I am confident readers will be able to shed some light on it before I publish the most interesting responses three months from now.

At Aston University in Birmingham, Barry Martin was also inspired by the Mandelbrot set. Martin adopted Mandelbrot's idea of iterating a formula from a numerical seed, but it is there that the similarity ends. Whereas Mandelbrot's patterns emerge from complex numbers, Martin's wallpaper is based on iterations of ordinary real numbers. Moreover, the numerical seeds for the Mandelbrot set are the points, infinite in number, found throughout a region of the plane; Martin's program grows its patterns from only one seed.

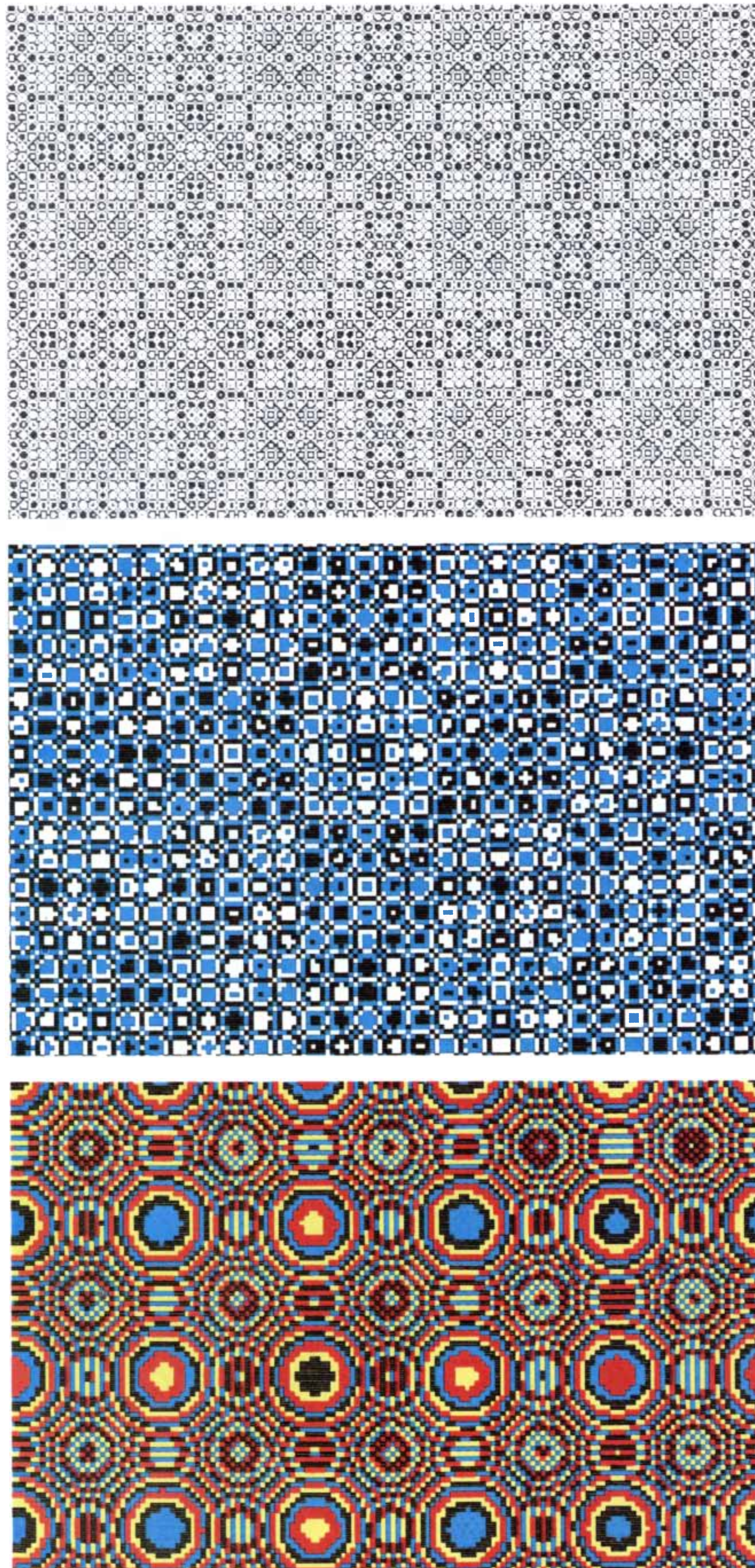
For example, Martin suggests trying the following pair of formulas, which can generate stunning, highly detailed images in many colors [*see the cover of this issue and the illustrations on the next page and pages 17 and 19*].

$$x \leftarrow y - \text{sign}(x) \times [\text{abs}(b \times x - c)]^{1/2}$$
$$y \leftarrow a - x$$

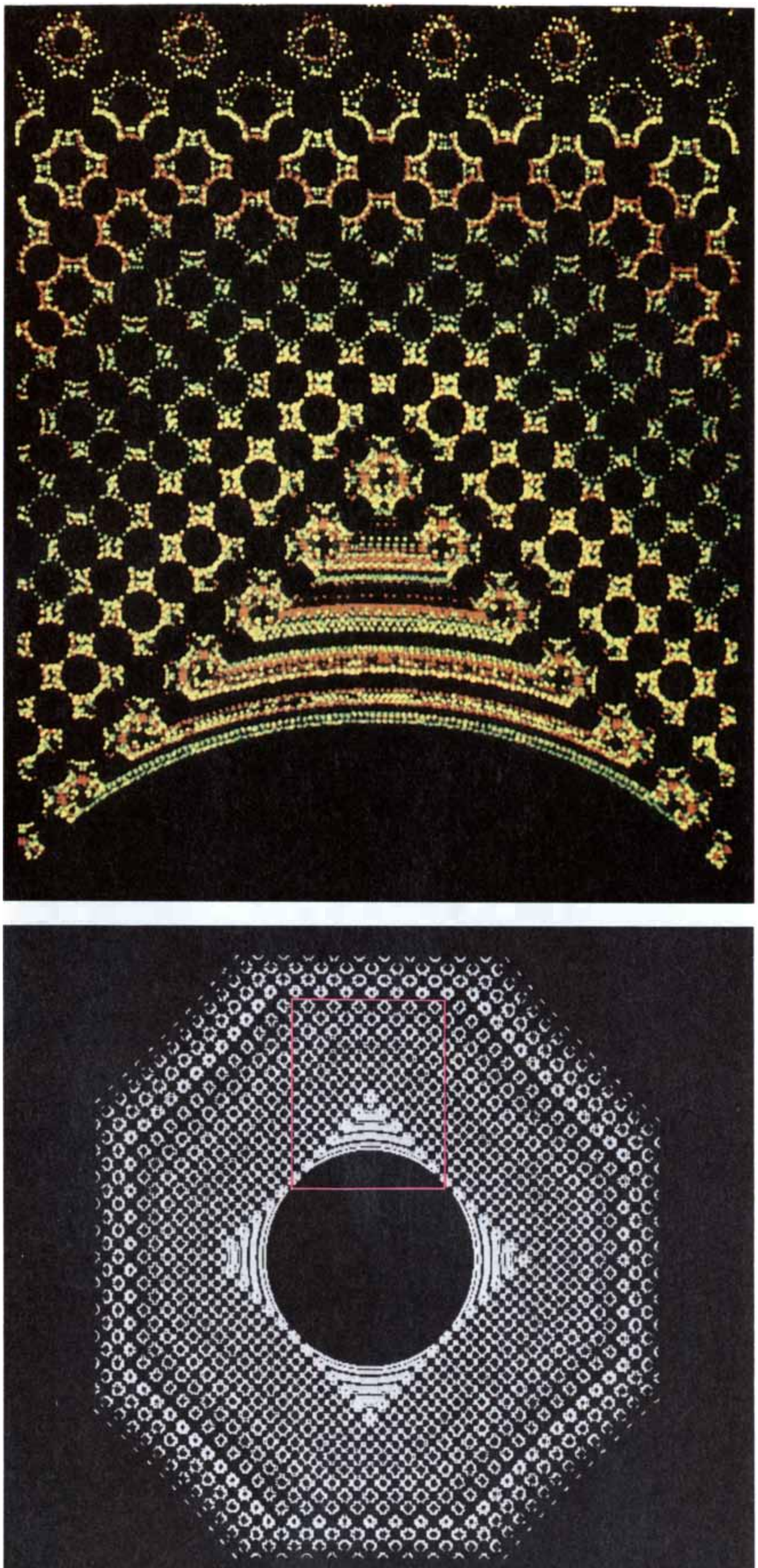
Here the function sign(x) takes the value 1 or -1, depending on whether x is positive or negative; the function abs($b \times x - c$) is the absolute value of the expression $b \times x - c$. The patterns can vary widely according to the values assigned to the letters a, b and c, which are numerical constants in the formula.

The formulas themselves are written in a form of mathematical shorthand: it is understood that one set of values is used for x and y in the formulas to the right of the arrows and a new set of values is computed from them on the left. The new values for x and y then replace the old values to the right of the arrows, and the calculation is repeated. In this way the program I call HOPALONG hops from one point to another. It begins at the point for which x and y are both equal to 0, namely at the origin. The next point might be at the upper right, the one after that at the lower left. A computer draws the points so quickly that one has the impression miniature electronic rain is falling on the screen: hundreds and then thousands of points drop onto the monitor. Soon a pattern begins to emerge. For example, if a is set equal to -200, b to .1 and c to -80, a broadly octagonal pattern is formed [*see bottom of illustration on next page*]. If the pattern is magnified and each point is colored according to the number of hops needed to reach it, the pattern becomes a wonderful cartouche [*see top of illustration on next page*]. For other values of a, b and c new designs appear: try setting a to .4, b to 1 and c to 0 [*see top illustration on page 17*], or set a to -3.14, b to .3 and c to .3 [*see bottom illustration on page 17*].

The algorithm for HOPALONG is al-



Circles and squares modulo 2, 3 and 4, from John E. Connett's program CIRCLE²



"Fractal" wallpaper from Barry Martin's program HOPALONG

most as easy to appreciate as the one for CIRCLE[2]:

```
input num
input a, b, c

x←0
y←0
for i←1 to num
   plot (x,y)
   xx←y − sign(x) × [abs(b × x − c)]^1/2
   yy←a − x
   x←xx
   y←yy
```

Another stampede of readers rushes off to type in the program. Your rewards for lingering here are a more detailed explanation of Martin's program and a description of the third kind of wallpaper program.

In order to run HOPALONG one enters the total number of iterations as the variable *num;* one also enters values for *a, b* and *c.* The larger the value of *num,* the finer the detail of the pattern. For example, if *num* is 10,000, the program will plot 10,000 points on the screen, but for some values of *a, b* and *c* that is only the beginning. If *a* is −1,000, *b* is .1 and *c* is −10, the pattern at low magnification resembles the rind of a four-lobed lemon [*see bottom of illustration on page 19*]. When the run of the program is extended from 10,000 points to 100,000 and then to 600,000, the filigree becomes increasingly ornate [*see top of illustration on page 19*].

The algorithm can work as it stands, but it can also be enhanced; for example, one might add a facility for moving off-screen points or compressing off-screen regions into the visible region. If such features are added, three more parameters for determining position and scale must be specified at the start of the program. The body of the main loop must then be modified: just after *x* and *y* are computed the enhanced version of HOPALONG adds the position changes to *x* and *y* and multiplies the result by the scale factor.

The wallpaper analogy was not lost on Martin: "It seems to me we are on the verge of a pattern-generating explosion that has great commercial implications, e.g., we can expect to see 'designer' wallpaper and textiles within the next few years. Patterns will be produced by the customer merely by the selection of a few numbers." Martin is equally sanguine about the implications for mathematical biology. Look once again at the four-lobed lemon. The enlargements show details strongly reminiscent of vascular bundles: could it be the outer rind of a monocotyledon in cross section? About these and other patterns Martin writes: "Clearly these curious configurations show us that the rules responsible for the construction of elaborate living tissue structures could be absurdly simple."

Readers might enjoy exploring the patterns generated by a different pair of iteration formulas, also suggested by Martin:

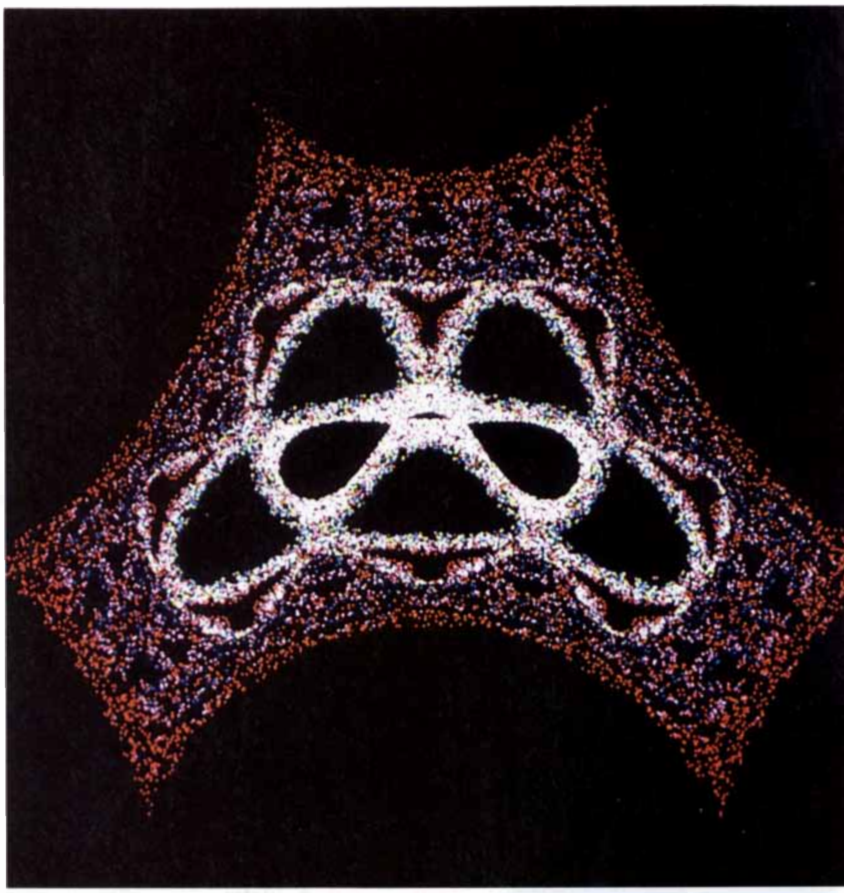
$$x \leftarrow y - \sin(x)$$
$$y \leftarrow a - x$$

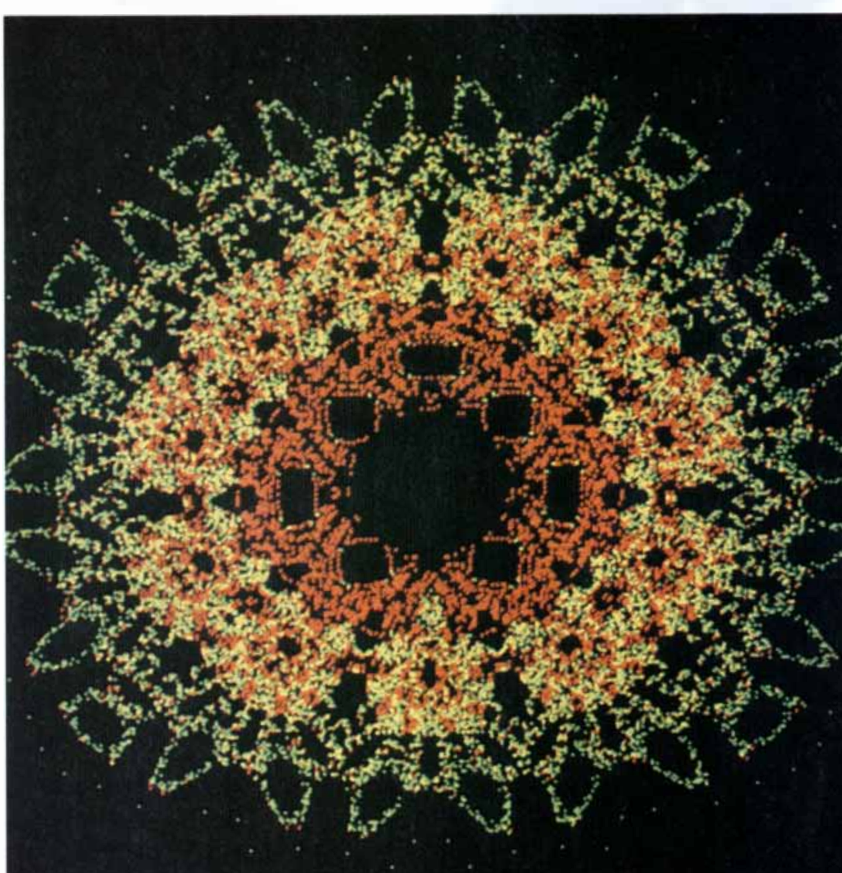
In these formulas only the variable a must be specified. Martin has discovered a most interesting series of patterns when a lies within .07 of the number pi.

The third kind of mental wallpaper must be reserved for rooms devoted to heavy thinking. The patterns range from Persian complexities to Incan riots [*see illustrations at bottom of page 20 and on page 22*]. The methods that lead to them could hardly be less like the techniques described above. Tony D. Smith of PICA Pty. Ltd. in Australia has devised intricate variations of the self-replicating cellular automaton invented in 1960 by Edward Fredkin of the Massachusetts Institute of Technology [see "Computer Recreations," by Brian Hayes; SCIENTIFIC AMERICAN, October, 1983]. There is astonishing potential for pattern in this idea, and Smith has begun to explore an amusing corner in the vast space of possibilities.

What is the Fredkin cellular automaton? Imagine an infinite, two-dimensional grid of square cells. At any given moment each cell is in one of two possible states: living or dead, so to speak. Somewhere an imaginary clock ticks away. The fate of each cell is determined by its four edge-adjacent neighbors: if the number of living neighbors is even for one tick of the clock, the cell will be dead at the next tick regardless of its previous state. On the other hand, if the number of living neighbors is odd, the cell will be alive at the next tick. The same rule is simultaneously applied to every cell on the planar grid.

Fredkin's automaton is closely related to the game of Life, invented by John Horton Conway of the University of Cambridge and often discussed in this department. Fredkin's automaton, however, was discovered earlier than Conway's, and it is much simpler. Moreover, it has an amazing property not shared by Life: any initial configuration of live cells grows through a series of generations (ticks of the clock) into four copies of itself. After several more generations there are 16 copies, 64 copies and so on. The finest wallpaper appears during intermediate gen-



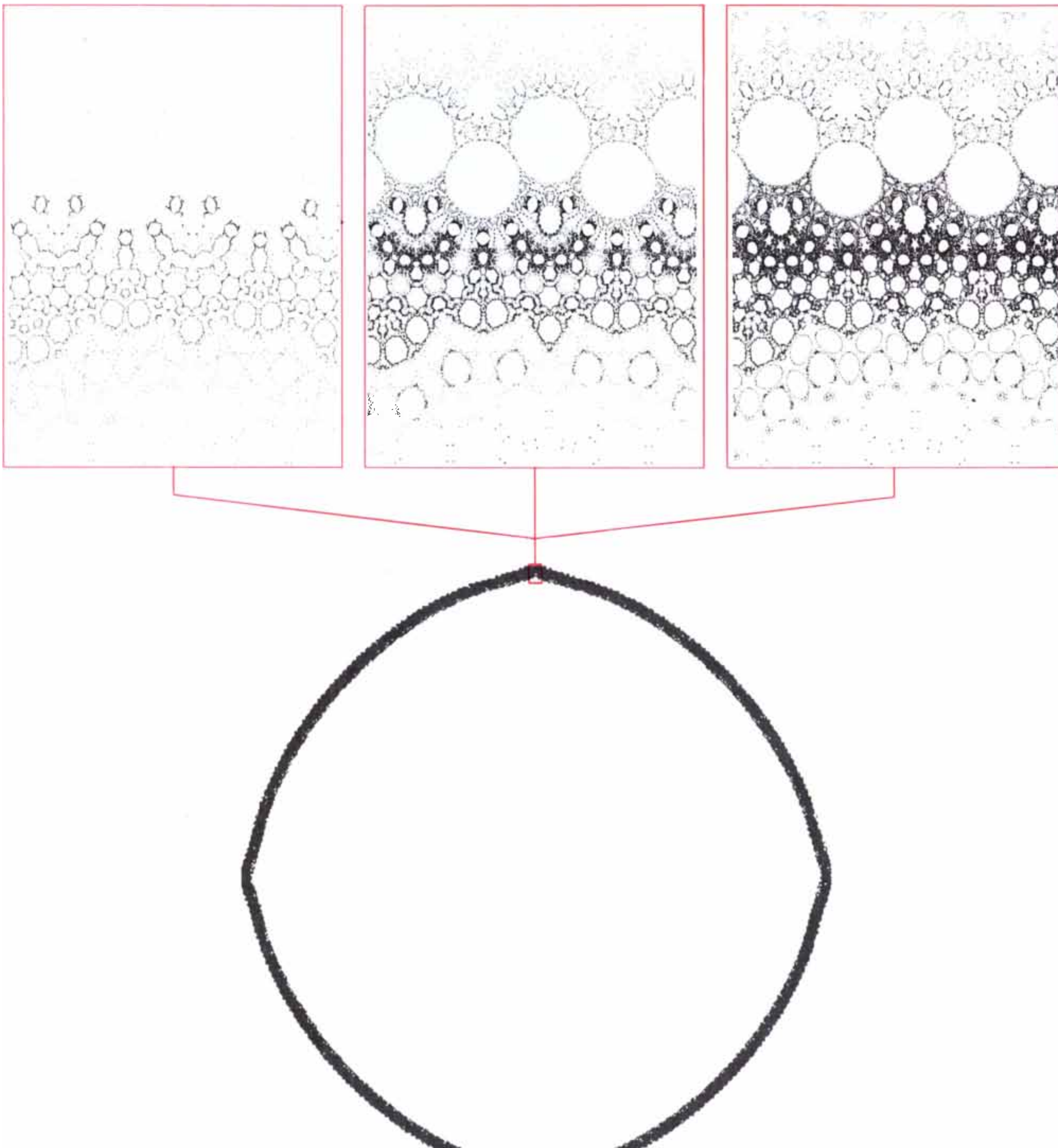


Mandalas on the number 7, generated by Martin's iteration formula

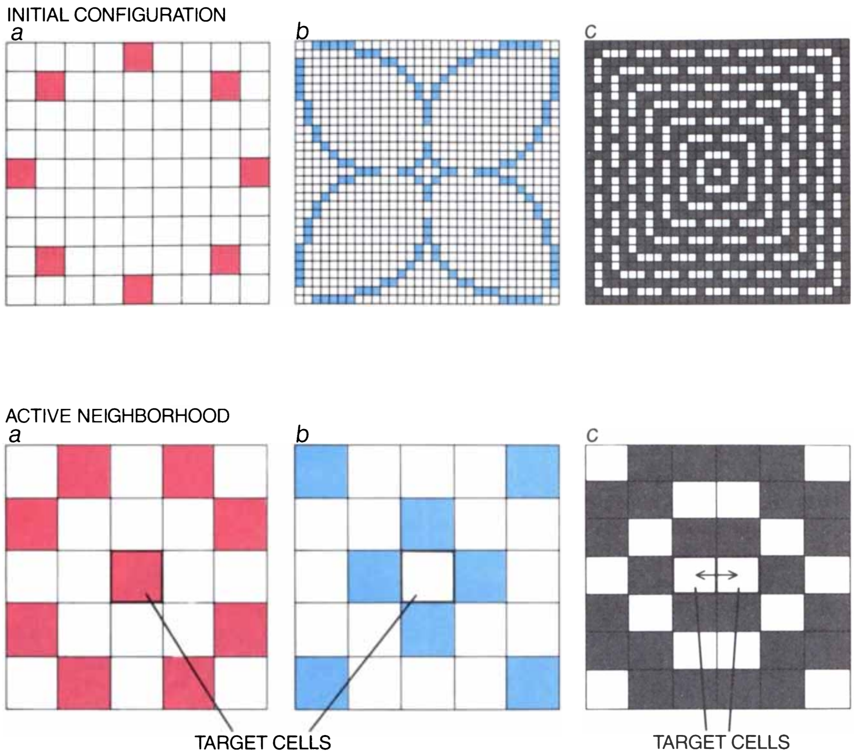
erations: in the generations between the ones in which the original population of live cells is replicated.

Smith's program for printing generalized Fredkin wallpaper is a highly versatile one called PATTERN BREEDER. The rules that determine the fate of a cell in PATTERN BREEDER need not depend solely on the state of the four edge-adjacent cells. In fact, before running the program one can specify the configuration of surrounding cells that will constitute the active neighborhood of each cell. The program then applies the same even-odd rule that was chosen in Fredkin's original automaton. At each tick of the clock if the number of living cells in the active neighborhood is even, the target cell will be dead in the following generation. Otherwise, the cell will be alive.

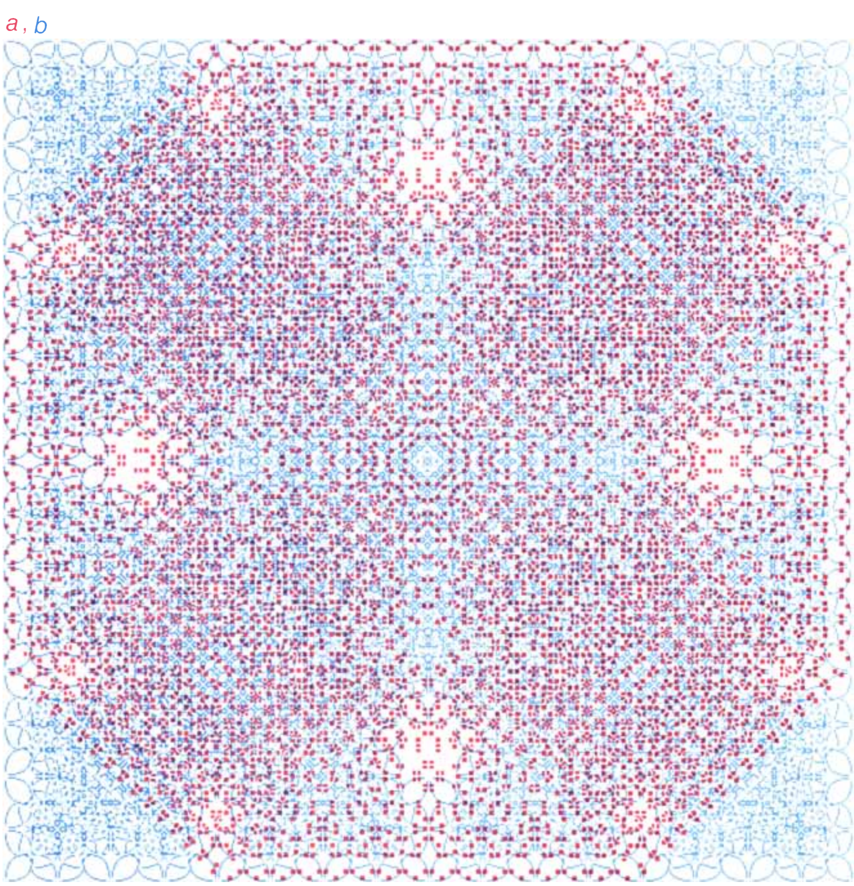
PATTERN BREEDER goes to work on any initial configuration of cells supplied by the user. For example, the initial configuration designated *a* in the top illustration on the next page gives rise to the red part of the pattern shown in the illustration immediately below it: for each target cell and for every stage in the evolution of the pattern the active neighborhood is the same. It is a complex pattern itself, which includes all the colored cells



A model plant stem and its vascular bundles, generated by HOPALONG



Rules for generating cellular automatons in Tony D. Smith's program PATTERN BREEDER



Two images generated by PATTERN BREEDER *and superposed*

within a 5-by-5 matrix, also designated *a* in the top illustration at the left. Note that the active neighborhood in this case includes the target cell itself. To apply the rule, count the number of live cells that coincide with the cells in the active neighborhood; if the target cell is currently alive, it too is included in the count. Initial configurations and the active neighborhood associated with each of them are also shown for two other images. The ones labeled *b* correspond to the blue part of the pattern in the bottom illustration at the left and the ones labeled *c* correspond to the illustration on page 22. There is an added complexity for active neighborhood *c:* the target cell itself oscillates in succeeding generations from left to right and back again in the center of the neighborhood.

Readers impatient to generate their own magnificent patterns may forgo the algorithmic struggle by writing to Smith at P.O. Box 256, 38 Ardoch Street, Essendon, Victoria 3040, Australia. Smith's program incorporates a library of initial configurations and active neighborhoods from which the user can choose; the program is currently available only for Apple Macintosh computers.

I am not about to outline PATTERN BREEDER in its full sophistication, but I shall describe a simpler program called FREDKIN. Readers with a bit of additional programming acumen can then convert FREDKIN into a more general program with some of the features of PATTERN BREEDER.

```
    input initial pattern
S   for each cell in the screen array
        count ← 0
        for each neighbor of that cell
            if neighbor alive
              then count ← count + 1
        if count even
          then cell ← 0
          else cell ← 1
               plot cell
    input go
    go to S
```

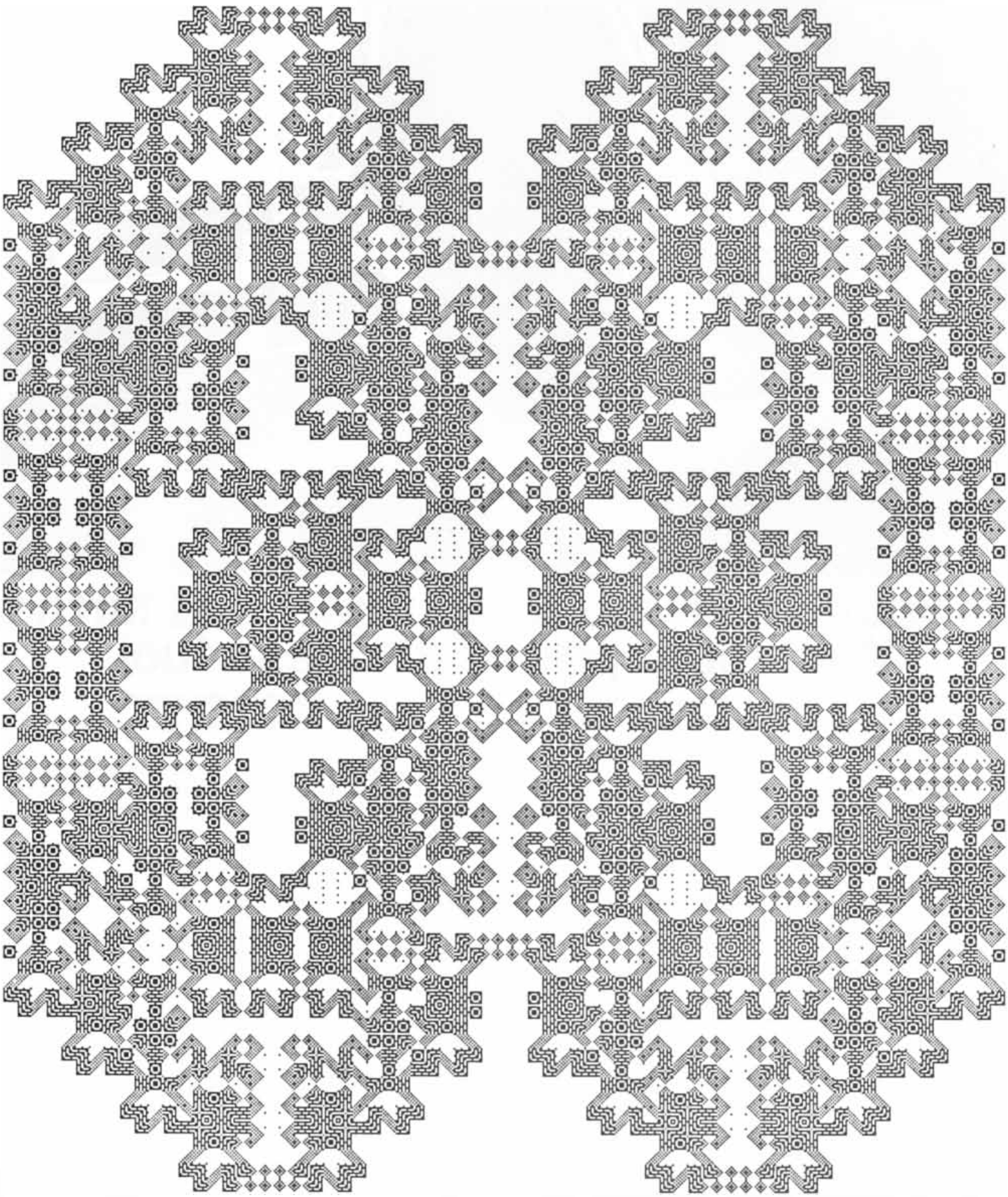
One of the great joys of writing algorithms is that so many levels of description are available. The line between descriptive generality and irresponsible vagueness is a fine one. Readers will note that FREDKIN, as specified above, occupies a slightly more rarefied stratum than the algorithms I outlined previously. For example, the instruction "Input initial pattern" will take several instructions to implement in any practical programming language. Any such instructions would involve a double loop, with two indexes *i* and *j*. Another double loop is concealed in the instruction

"For each cell in the screen array." Here the two indexes supply the coordinates of points on the display screen or the printer.

Inside the main program loop FREDKIN simply carries out the rule for the evolution of a given pattern. It counts the number of living neighbors for each cell (i,j); then, if the number is odd, it plots or prints the cell as a single point. The last instruction of the main loop calls for input of the variable *go*. The user may type any number at this stage in order to get FREDKIN to generate the next pattern. In this way the execution of the program can be halted at will if a particularly pleasing pattern appears. Sometimes a bit of trickery and a go-to statement are useful. This strategy is not structured programming, but it works just fine.

I shall be pleased to print, in three months' time, the finest mental wallpaper that readers are able to supply.

c



A Mayan abstraction from Smith's wallpaper program

The rest will make up a new sample book to be presented if and when I return to this topic.

Last month's column on computer magic left the inner logic of two tricks unrevealed. The first of them was based on the well-known "Stop!" trick of Harry Lorayne, a New York magician. A volunteer chooses a card from the middle of a well-shuffled deck and then replaces it. The volunteer deals the deck into six smaller hands and finds the hand that includes the chosen card. The computer then instructs the volunteer to deal out the hand; after each card is dealt the computer tells the volunteer whether to continue dealing or to stop, thereby indicating the position of the chosen card in the deck.

How does the computer stop at the chosen card? The answer is that when the volunteer deals out the six hands, one card from the central six cards of the original deck is dealt into each of the six piles. Each of the six cards occupies the same relative position in its pile, namely the fourth. Because the volunteer deals only from the pile that includes the chosen card, the computer can always pick the card by stopping the deal after the fourth card.

The second unexplained trick was based on the key pad of an ordinary calculator. A volunteer enters two three-digit numbers into the calculator, multiplies them and calls out all but one of the digits. How does the magician guess the missing number?

According to the magician's instructions, each three-digit number selected by the volunteer must come from the digits along a row, a column or a main diagonal of the key pad. Such a number has the mystical property of "3-ness" after casting out 9's. When the digits of the numbers are added together, their sum can be a one- or a two-digit number. If the sum is a two-digit number, the sum of the two digits is a one-digit number. In either case the final one-digit result is a multiple of 3. For example, along the first column the number 471 is reduced to 12 and then to 3, which is a multiple of 3. On the other hand, when the two three-digit numbers selected by the volunteer are multiplied and the product is subjected to the same peculiar treatment, the result is 9. The procedure is equivalent to multiplication modulo 9—hence the term "casting out 9's." If one of the digits in the product, say x, is missing, the result of casting out 9's for the remaining digits is $9 - x$. This number is computed by the magician as the volunteer reads off all but one of the digits. In most cases knowing $9 - x$ makes x instantly computable.

BOOKS

A positive-sum strategy for productivity that, Thurow says, does not quite add up

by Lester C. Thurow

THE POSITIVE SUM STRATEGY: HARNESSING TECHNOLOGY FOR ECONOMIC GROWTH, edited by Ralph Landau and Nathan Rosenberg. National Academy Press ($35).

Clearly this book is intended to be a major statement about both economic policy and economic reality. Its distinguished and powerful roster of 43 authors includes chief executive officers of several major U.S. corporations and financial institutions, the president of a major university, leading economists and natural scientists from Harvard, Stanford and other centers of learning, and Government officials. The editors are Ralph Landau, a vice-president of the National Academy of Engineering, consulting professor of economics at Stanford and former chairman of the board of the Halcon SD Group, and Nathan Rosenberg, chairman of the economics department at Stanford. The essays taken together offer a diagnosis of the peculiar plight that has afflicted the U.S. economy like a persistent cold since the end of the Vietnam war. Although many of the authors offer genuine insight, collectively they have, I think, got the diagnosis more wrong than right.

When it comes to one central theme of the book, the authors are certainly accurate: technology can be harnessed to raise living standards, while minimizing the use of labor and raw materials, so that we can escape from a zero-sum society in which every person's income gain must be matched by someone else's income loss.

With the replacement of some copper by fiber optics in telecommunications, the miniaturization of electrical equipment by microelectronics and the use of aluminum in long-distance power transmission, copper has gone from being a high-priced element in short supply 15 years ago to selling today for less in real terms than it did at the bottom of the Great Depression.

By aggressively harnessing technology the Japanese are fast closing the eight-to-one gap in living standards that existed between Japan and the U.S. in the 1950's; they may well succeed in their announced goal of having the world's highest real standard of living by the year 2000.

There is no argument about the desirability of a more rapid advance in what engineers know as technical change and economists know as productivity growth. The authors and editors run into difficulty when they explain why technical change seems to have slowed down in the U.S. The difficulty grows as they attempt to fashion a remedy that will accelerate growth in American productivity.

Productivity growth (the advance in output per hour of work) is the economist's measure for how rapidly new and more efficient technologies are being embedded into economic activity. When they are examined, the productivity growth statistics form a dismal picture of America's technological progress. From 1948 to 1965 productivity increased at a rate of 3.3 percent per year. After 1965 a gradual but very persistent decline began. From 1977 to 1985 productivity growth averaged only .7 percent per year. In 1985 nonfarm business productivity actually fell .3 percent. Productivity often falls during recessions (reductions in overhead labor are slower than reductions in output), but 1985 was not a recession year. It marked another kind of ominous event: it was the first year since the data have been kept in which a fall in productivity was not accompanied by at least one quarter of economic decline.

Furthermore, 1985 was marked by intense international competition in which virtually every American industry lost market share to foreign producers. Low-productivity firms were driven out of business. Given such pressure to become more efficient, 1985 should have seen rapidly rising productivity. It did not. Why?

Whatever is happening, it is happening only in America. Productivity growth rates in every market economy fell after the first OPEC oil shock in 1973. (The conventional explanation is that investments were focused on saving energy rather than on raising labor productivity.) Yet the rest of the industrial world rebounded after the second OPEC oil shock in 1979 and since then has enjoyed productivity growth rates that are four to six times those the U.S. has posted.

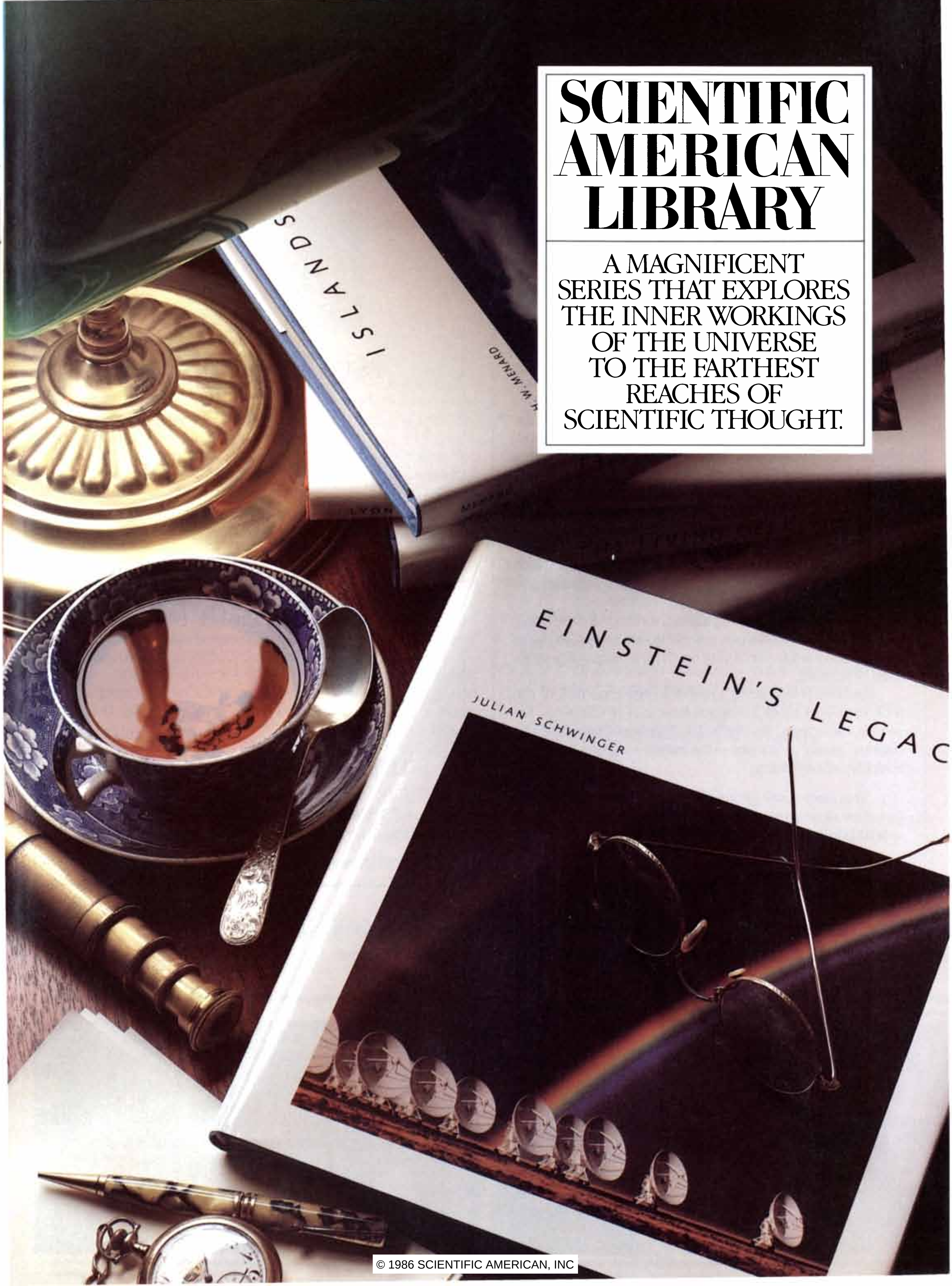
Although *The Positive Sum Strategy* is written as if the slowdown in U.S. productivity growth is a complicated mystery, a simple, straightforward diagnosis leaps out at the reader. As the editors state in their overview, "the truly fundamental factors which can increase the rate of growth permanently are the *rate* of technical change (change resulting primarily from expenditures upon R&D) and the increase in the quality of the labor force." Harvey Brooks, professor of applied physics and of technology and public policy at Harvard, elaborates the theme. He points out that America currently invests less in civilian research and development than its major industrial competitors do—1.7 percent of the G.N.P. for the U.S. versus 2.5 percent for West Germany and 2.3 percent for Japan.

Nor does the quality of the American work force measure up. The U.S. has more functional illiterates (13 percent of the adult population), graduates a smaller percentage of its population from high school, has a greater percentage of high school graduates who are undereducated in mathematics and produces a smaller percentage of engineers among its college graduates than any of its main competitors do. The result is a work force low in quality when it is judged against any that existed in this nation's past, or any existing today in nations that are major competitors of the U.S. Given low-quality inputs, why should anyone expect high-quality output?

Yet, as the authors point out, this simple story is only part of the answer. Investment in plant and equipment is also a factor. Most new technologies require new capital equipment; a labor force whose skills have been improved usually requires new tools to exploit its enhanced abilities. Nations or industries that invest more usually adopt newer technologies faster. The U.S. is simply investing less in acquiring new tools than its competitors do. It seems obvious that the U.S. will have to invest more.

Throw in both entrepreneurs willing to take risks and ample market opportunities, other points mentioned by several of the authors, and America is home free. Or is it?

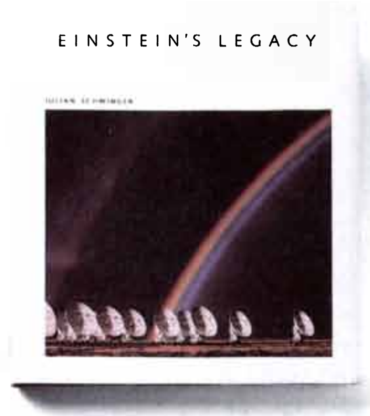
If one looks at the authors' views of


SCIENTIFIC AMERICAN LIBRARY
A MAGNIFICENT SERIES THAT EXPLORES THE INNER WORKINGS OF THE UNIVERSE TO THE FARTHEST REACHES OF SCIENTIFIC THOUGHT.
ISLANDS
H. W. MENARD
EINSTEIN'S LEGAC
JULIAN SCHWINGER

Your Premier Volume

EINSTEIN'S LEGACY

Julian Schwinger



Nobel laureate Julian Schwinger tells the story of one of the twentieth century's greatest achievements—the theory of relativity—which is largely the work of one man: Albert Einstein.

The groundwork was laid in the seventeenth century by Isaac Newton, who unified motion on Earth and in the heavens in a framework of absolute space and time; then, in the nineteenth century by James Clerk Maxwell, who unified electricity, magnetism, and light. But it was left to the sixteen-year-old Einstein to glimpse, for the first time, that the theories of these two giants were incompatible.

Schwinger makes a lively narrative of Einstein's quest for the reconciliation of this conflict, a quest that led to the unification of matter and energy, and of space and time in his special and general theories of relativity.

The special theory has had its awesome confirmation in mankind's command of the nuclear force. Schwinger shows how the general theory, in turn, has stood up for 70 years to experiments drawn from the theory by Einstein, himself, and by his successors, employing ever more ingenious instrumentation and carrying the proofs ever farther beyond the decimal point.

In the last chapter, Schwinger describes space-age experiments made possible by the technologies that incorporate the theory itself. Their outcomes may not only secure further confirmation of Einstein's legacy, but face it with difficulties that lead on towards a still more comprehensive theory.

Julian Schwinger was awarded the Einstein Prize in 1951, the National Medal of Science in 1964, and the Nobel Prize for physics in 1965.

He is currently University Professor of the University of California, Los Angeles. He received his Ph.D. from Columbia University and has been on the faculty at Purdue University and Harvard University. Through the years, he has done theoretical work in various areas of both classical and quantum physics.

"I am happy to report, that his book makes for delightful and instructive reading. Technical demands on the reader do not exceed the most elementary algebra.

"It is particularly gratifying that the subject's intricacies, including glimpses of non-Euclidean and Riemannian geometries, and the theory's early successes, . . . are so well conveyed here in simple language. Altogether, this well printed and pleasingly illustrated book is an ideal gift for the curious non-expert." —NATURE

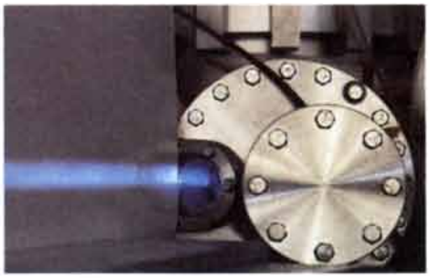
250 PAGES, 200 ILLUSTRATIONS

COVER PHOTO: The rainbow arches over the Very Large Array radio telescope of Soccoro, New Mexico. For nearly three generations, Einstein's legacy has withstood experimental test by instrumentation not yet invented when Einstein did his thinking. © Doug Johnson

TO GUARANTEE SWIFT DELIVERY OF YOUR FIRST VOLUME OF THE SCIENTIFIC AMERICAN LIBRARY, WE WILL SHIP *EINSTEIN'S LEGACY* VIA UPS.

Radio telescopes
Stephen Northup/Black Star © 1980

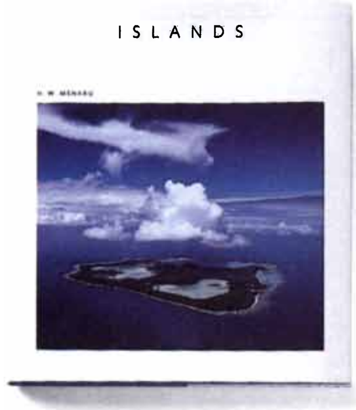
Einstein and Chaplin in Hollywood, 1931
The Granger Collection



Synchrotron light produced at Brookhaven National Laboratory

ISLANDS

H.W. Menard



In a lifetime of scientific adventure on the broad reaches of the Pacific Ocean, H.W. Menard discovered in the Earth's most romantic islands new insights into the forces deep in the Earth's interior.

His central contributions to plate tectonics, the unifying theory of geology, explain the layout of the island arcs, the atolls and the seamounts on the map of our globe. He also tells the story of the evolution of life, of human settlement and of the exploration, in the latest centuries, of these islands.

Sharing in Menard's adventure, the reader will also share his understanding of the Earth as a dynamic system. Photographs and cross section diagrams illuminate the text and make vivid connection between the beauty of nature at the surface and the processes driving the system in its depths.

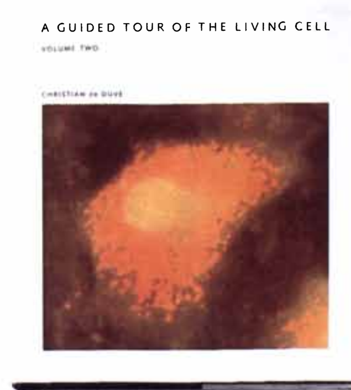
H.W. Menard was a founding member of the Scripps Institution of Oceanography, where he was professor of Geology. In the course of his distinguished career, he served as Technical Advisor in the President's Office of Science and Technology and as Director of the U.S. Geological Survey.

"In a very real sense, this book is his scientific legacy."—ROGER REVELLE, PROFESSOR OF SCIENCE AND PUBLIC POLICY, UNIVERSITY OF CALIFORNIA, SAN DIEGO

230 PAGES, 152 MAPS, PHOTOGRAPHS AND ILLUSTRATIONS

THE LIVING CELL

Christian de Duve



In A GUIDED TOUR OF THE LIVING CELL, Christian de Duve invites us to accompany him on an engrossing expedition through the trillions of cells that make up the human body.

The journey is divided into three itineraries. In the first, we examine a cell's outer and inner membranes and their two-way commerce with the exterior. In the second, we call at all the major organelles—the cell's own organs—and observe them manufacturing their own special products. Our third itinerary takes us into the nucleus itself, where we see DNA steering the reproductive biochemical machinery of the cell.

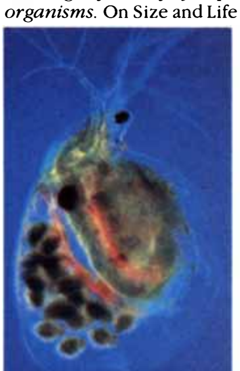
Christian de Duve is Andrew W. Mellon Professor at The Rockefeller University in New York. He was a joint recipient of the Nobel Prize for medicine in 1974.

"It is hard to imagine a better introduction to the world of the cell and all it means to everyone."
—JOHN BARKHAM REVIEWS

2 VOLUMES, 422 PAGES, 425 ILLUSTRATIONS

"The difference [between this and other book series] comes in the quality of
With the handsome volumes in front of us, what response can

The largest forms of life reproduce slowly, more ponderously, and live much longer than smaller organisms. On Size and Life *examines why.*



A molecular view of the plasma membrane. Electr
through an axopod of the heliozoon Echinosphae

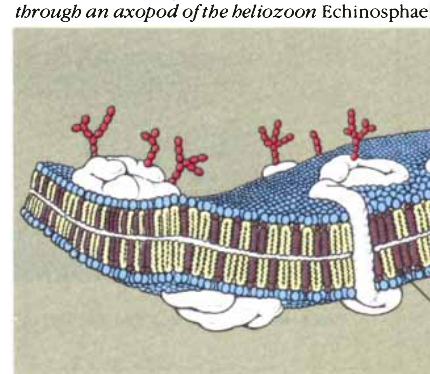
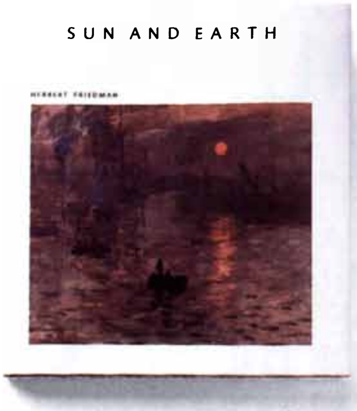


PHOTO CREDITS—Daphnia: James Bell, Photo Researchers, Inc./Ant: al History Photographic Agency/Illustration drawn by Neil O. Hardy/Electron microg
Mutual Research Corporation

SUN AND EARTH

Herbert Friedman



As we can tell with our unaided senses, the distant sun bathes the Earth in radiant energy. From instruments on the ground and in orbit, we have now learned that the Earth revolves around in the Sun's extended atmosphere—the two bodies are in contact. An out-rushing solar wind engages our outer atmosphere and sets into action the processes that reach all the way to the planet's surface.

In SUN AND EARTH, the space-scientist Herbert Friedman combines an historical narrative with interpretive accounts that retain the awe and wonder with which scientists and poets have regarded the sun's life-bestowing characteristics.

SUN AND EARTH also provides a comprehensive picture of how the sun shapes our environment and explains the impact of solar variability on the contemporary problems of survival.

Herbert Friedman is chief scientist emeritus of the E.O. Hulburt Center for Space Research of the Naval Research Laboratory. In 1969, Dr. Friedman received the National Medal of Science—the nation's highest honor for scientific achievement.

"It covers more ground in an authoritative fashion than any contemporary book I've seen."
—WILLIAM C. LIVINGSTON, ASTRONOMER

251 PAGES, 195 ILLUSTRATIONS

OTHER VOLUMES IN THE LIBRARY:

SUBATOMIC PARTICLES
Steven Weinberg

MATHEMATICS AND OPTIMAL FORM
Stefan Hildebrandt and Anthony J. Tromba

POWERS OF TEN
Philip and Phylis Morrison and the Office of Charles and Ray Eames

HUMAN DIVERSITY
Richard Lewontin

PERCEPTION
Irvin Rock

DRUGS AND THE BRAIN
Solomon H. Snyder, M.D.

THE SECOND LAW
P.W. Atkins

THE SOLAR SYSTEM
Roman Smoluchowski

THE SCIENCE OF MUSICAL SOUND
John R. Pierce

FOSSILS
George Gaylord Simpson

CONSTRUCTING THE UNIVERSE
David Layzer

FIRE
John W. Lyons

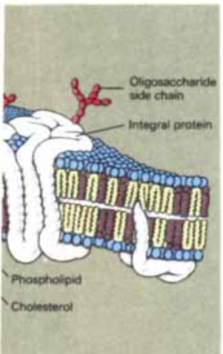
ON SIZE AND LIFE
Thomas A. McMahon and John Tyler Bonner

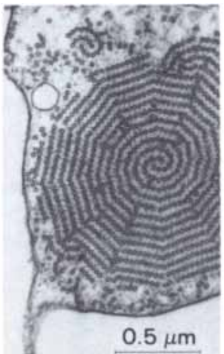
THE TIMING OF BIOLOGICAL CLOCKS
Arthur T. Winfree (in preparation)

EXTINCTION
Steven M. Stanley (in preparation)

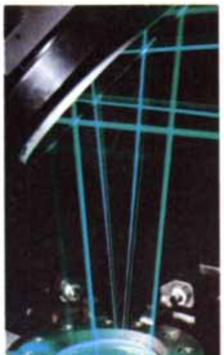
e authors, who combine eminence in their fields with proven gifts of exposition.
ere be but immediate surrender to this firepower?" —Nigel Calder, NATURE

nicrograph of a section n nucleofilum. *From* The Living Cell.





The "crowning" of a forest fire. Laser-beam spectroscopy. The copper converter. An experimental warehouse fire. From Fire.



h courtesy of Dr. L.E. Roth and Dr. Y. Shigenaka/Fire in trees: L. Burr, Gama-Liaison/(, Inc./Laser: Courtesy of Sandia Labs/Plastics Fire Test: Courtesy of Factory

who will have to change standard operating procedures to bring about the necessary improvements, it would seem that the obligation falls only on government.

Government should "lower real interest rates, which are affected by monetary policy, deficits, and the uncertainty of government policies"; government should "adopt a tax structure that really *would* promote saving and investment" while recognizing that incentives such as "the capital gains tax differential have been beneficial for innovation and risk taking"; government should "adopt sensible regulatory policies and work to ease ...fears about poisons, illnesses, and other hazards"; government "should reduce the rate of increase in...expenditures, including perhaps even defense, but increase R&D expenditures for basic and generic research"; government should "avoid excessive...intervention in markets," and government should remember that "national planning and real innovation are incompatible." As James Brian Quinn, professor of management at Dartmouth College, states: "Because of the driving power of such motivations, maintaining a climate friendly to entrepreneurs should be a central focus of any economic growth policies in the United States." And in its other activities government has "overemphasized equity," both legal and economic, and has pressed "for greater equity in the distribution of income even at the expense of possible loss of efficiency."

Put bluntly, the authors recommend that, with the exception of spending more on R&D, government should get out of the way, reduce its role in the American economy and let the market work its magic. Even if this analysis were correct (and I shall demonstrate that it is not) it creates a credibility problem for the industrialists, scientists and economists writing these essays because it sounds so horribly self-serving. Increase the government R&D expenditure programs that benefit me but cut the government social programs that benefit everyone else. Cut my capital gains taxes but raise consumption taxes on other Americans. What has not been conveyed with credibility is the very important idea that achieving advances in technology and productivity is not a selfish special-interest-group issue but a general issue that is vital to the living standards of every individual in every interest group in America.

Gaining credibility for one's recommendations is relatively easy. One has merely to recommend a few measures that adversely impinge on one's own immediate standard of living rather than only measures that affect someone else's economic well-being. This the authors have not done, and as a result the book will have little political credibility whatever its scientific and economic merit.

Where the analysis breaks down technically is in its failure to follow the advice of Harvey Brooks to "scan and adopt foreign" practices. There is no scanning of Europe, although Europe has a trade surplus with the U.S. that approaches the surplus achieved by Japan. And where the volume does scan foreign practices (in the essays on Japan by Masahiko Aoki, professor of Japanese studies and economics at Stanford, and Daniel I. Okimoto, associate professor of political science at Stanford) the scanning has no noticeable impact on the recommendations.

If too much American equity is the problem, for example, how do the authors explain the fact that among industrial countries the U.S. (with the exception of France) has the most unequal distribution of income? American income inequalities are 50 percent greater than those of Japan and 35 percent greater than West Germany's.

If excessive government spending is the problem, how do the authors explain the fact that the OECD has just announced that Japanese government now spends a larger fraction of Japan's G.N.P. than American government (local, state and Federal) does of the U.S. G.N.P.? How do they square their analysis with the fact that since the Japanese have only a small defense budget, social spending is now a greater proportion of government outlays there than it is in the U.S.? If overall government spending is the problem, why is productivity worst in the country—the U.S.—that now has the smallest government sector among all major industrial countries?

Are high taxes the problem? What do the editors and authors who call for lower taxes have to say about the contribution by Dale Jorgenson, professor of economics at Harvard? Jorgenson shows that the effective American corporate tax rates were far higher in the 1950's and 1960's, when productivity was growing at a rate in excess of 3 percent, than they are now, when productivity is growing at less than 1 percent per year.

If more capital investment is needed, why is it that investment in plant and equipment is a larger fraction of the U.S. G.N.P. now (11.6 percent from 1977 through 1985) than it was then (9.5 percent from 1948 through 1965), whereas productivity and technical change are occurring more slowly now than they were then? If entrepreneurship is a key, why is it that productivity is growing much more rapidly in Japan and Germany even though no one would describe those countries as more entrepreneurial than America?

Why is it that Japan, a society preeminent for its "administrative guidance," and Germany, where governments own major fractions of what in America would be private industries, perform well? Why is it that the editors virtually ignore in their recommendations what has worked so well in Japan? Albert Bowers, chief executive officer of the Syntex Corporation, in his article on how Japan caught up with the U.S. provides the familiar list of key policies (targeting specific markets, infusing selected industries with low-cost capital, inhibiting foreign access to key domestic Japanese markets, forgoing immediate profits, and so on). Yet the editors refer only in passing to this example as they press their policy agenda on the reader.

With the exception of an excellent article on the economic handicaps of the American legal system by Milton Katz, professor of law at Harvard, what the authors completely miss is the role of social organization, or what might be called "soft" rather than "hard" sources of technology growth. If American productivity is to grow more rapidly, organizations other than government will have to change their standard operating procedures.

If, for example, one looks carefully at the .5 percent per year rate of growth of nonfarm business productivity between 1978 and 1985, one discovers some interesting facts. During those years American business firms reduced their blue-collar payrolls by 1.9 million workers, or 6 percent, while increasing the business G.N.P. by 18 percent (after correction for inflation). If one produces 18 percent more while reducing inputs by 6 percent, one has achieved a 24 percent increase in productivity. Divide that number by seven years; the calculation shows that the blue-collar workers of America on the factory floor were generating a rate of growth of productivity in excess of 3 percent per year—world class.

What were those same firms doing when it came to their white-collar work forces? They were adding 10 million white-collar workers—a 21 percent increase in employment. Since output was only up by 18 percent, white-collar productivity actually fell in American industry. There are now 58 million white-collar workers and only 30 million blue-collar workers on American payrolls. As a consequence the fall in white-collar productivity wiped out much of the gain in blue-

We're making waves in medical technology.

The Lithotripter. Designed, developed and engineered by Dornier, it employs shock waves to replace surgery in the removal of kidney stones.

Described as a "medical miracle" by the Secretary of Health and Human Services, the Lithotripter eliminates the pain, discomfort and much of the cost normally associated with surgical procedures. It was developed by Dornier as a technological refinement of research which had originally been targeted toward studying the effect of shock waves on supersonic aircraft. It is the only device of its kind in the world today that has successfully treated over 100,000 kidney stone sufferers.

At Dornier, we're working on other major medical devices employing shock wave technology. We believe it is the wave of the future, not only in providing improved treatment, but in making high technology work to help lower the cost of medical care.

DORNIER

DORNIER MEDICAL SYSTEMS, INC., 824 LIVINGSTON COURT, MARIETTA, GA 30067

collar productivity. If one adds in service workers, many of whom indirectly work for business, the decline in what is essentially white-collar productivity would be even greater.

By 1985 American businessmen were talking about excessive white-collar overheads, but in spite of the talk and the occasional newspaper article about firms reducing middle-level management, the data show that quite the reverse was happening. In 1985 the real business G.N.P. rose 2.7 percent. Yet executives and managers were added to American payrolls at the rate of 5.6 percent—more than twice as fast as output was growing. At the same time administrative support staff (there are now 17.3 million such people on American payrolls) were being added at the rate of 3.5 percent, or 30 percent faster than output was growing.

Industry-by-industry comparisons of productivity between American and foreign firms, such as the famous studies done of the relative costs of producing the Ford Escort, always show that much of the American cost (productivity) handicap is due to relatively high white-collar overheads.

Since the authors focus on the situation that prevails on the factory floor—a part of the firm where little is wrong—something important is being left out of the analysis they offer. Even if all the authors' recommendations were implemented, little would change in the average American office; yet that is where the major problems lie. When it comes to the office, there has been an immense absorption of hardware and software but no improvement in productivity. Why?

Several of the authors talk about what Burton Klein, professor of economics at the California Institute of Technology, calls "the propensity to engage in risk taking." Such phrases have little content when one is talking about office productivity. If the willingness to take risks means, for example, being willing to invest in new technology, then there has been a lot of risk taking in the office. Office automation is sweeping through American business at prodigious rates. In recent years office computers have accounted for a very large percentage of total American industrial investment and the American office is on the average much more computerized than the offices of any of its foreign competitors are, yet office productivity tumbles in the face of that technology. Enormous amounts of capital (risk taking?), technology and highly skilled workers were going in (exactly the factors called for by the authors), but negative productivity was coming out.

The fault does not lie with American government; it lies precisely with American firms. Deregulation (of oil, transportation and banking) arrives and there is much less enforcement of the antitrust regulations, and yet private American white-collar bureaucracies grow bigger. Whatever else they may be, American managers are not world class when it comes to managing themselves.

This is not the place for a long discussion of why white-collar overheads have grown so rapidly in American firms, but I would argue that the trend cannot change without some fundamental restructuring of the way American firms manage and organize themselves.

Accountants provide one good example. Computerized accounting, it could be presumed, would have enabled American firms to operate with fewer accountants. Yet the number of accountants on American payrolls has risen from 1.0 million to 1.3 million between 1978 and 1985. The invention of a new technology has simply led to ordering up old accounts more frequently and inventing new forms of accounting (management information systems, cost accounting, inventory control, and so forth) that were not previously feasible, but not to thinking how any of these systems could really improve the ultimate results.

Perhaps the answer is to be found in foreign systems of participatory management, such as those that let blue-collar workers do inventory control, that do not send enormous quantities of information up the corporate pyramid and thousands of orders down the pyramid. That is a dangerous thought, since it would require the authors to tell corporations that they too are inefficient and will have to make changes in standard operating procedures. The fault lies not in Caesar, dear Brutus, but in thee.

Before following the prescriptions of these authors and thereby allowing a shift in taxes that will cut consumption or reductions in government programs that provide benefits to the citizenry, the average American voter might well ask why he should believe American industry is likely to spend whatever extra income is surrendered to it any more wisely than it spent the billions of dollars recently put into office automation.

He might also ask why he should pay more taxes to support more research and development when one author (Edwin Mansfield, professor of economics at the University of Pennsylvania) argues that "technology is being transferred across national boundaries more rapidly than in the past." Gaining a technological edge is essentially the name of the game being recommended, yet no one explains how American voters or workers, as opposed to American firms, get a technical edge in a world of multinational or transnational firms for whom instant offshore production (Kodak, for example, made no attempt to manufacture its new eight-millimeter video camera in America) is a way of life.

That average voter might also ask for a comment on the argument made by the chief executive officer of the Phillips Corporation, widely reported in the European press at the end of March, that America would eventually be completely driven out of the electronics business because U.S. firms had given up on consumer electronics—the market that can provide the volume necessary to support investment in new process innovations, which can then be carried over to industrial and military electronics. Why should managers be supported who are making such strategic mistakes?

The average American might be forgiven for thinking that something other than government must be wrong with the way American management is organizing itself and making decisions if nonunionized American firms can, in a very brief period of time, go from a 90 percent market share in 8K RAM chips to a 10 percent market share in 256K RAM chips. In an excellent Ph.D. thesis Charles Ferguson, a Massachusetts Institute of Technology graduate student, documents the fall of the American semiconductor industry. It was precipitated by a shrewdly thought-out and well-executed strategy developed by government and industry in Japan; the strategy worked because of the fragmented structure of American industry and the shortsighted decisions and high turnover that flowed directly from the way American industry is organized.

The article by Gordon Moore, chief executive officer of Intel, briefly mentions some of the structural problems. "An excess of venture capital is...often disruptive to the company the entrepreneurs leave when they start their venture.... A major project, important for our international competitive position,...was delayed for a year or more because of the people leaving for a start-up." How does one build a successful company that lasts for more than one generation if each group of bright young engineers wants to start up its own company and become rich? Like too much ice cream, too much entrepreneurship can give the economy indigestion. Yet the authors call for still more entrepreneurship.

There is, to be sure, some dissent by

authors from the dominant themes. Paul David, professor of economics and history at Stanford, makes the point that markets do not always lead to optimal results. (He cites the awkwardly arranged QWERTY keyboard of the typewriter and computer as an example.) But nowhere does this realization seem to inform the recommendations that are offered by the editors and authors.

Several authors see a weakness in manufacturing and process R&D. Harvey Brooks recognizes that "U.S. emphasis on product as compared with process innovation relative to other countries may. . .be a factor" in America's poor performance. Robert Malpas, managing director of the British Petroleum Company, points out that "a production engineer is near the bottom of the engineers' merit scale." Gordon Moore sees that "if we lose the manufacturing battle to our overseas competitors, the research and development that creates the opportunities is in real danger of losing the revenue stream that is required to support it."

Industry is exhorted to do more process R&D and to give those who do it more credit. No explanation is offered, however, for why American firms have underinvested in process R&D or why they are now suddenly going to change those habits. In those efficient free markets, if Americans do not invest in process R&D, it must be because it is not profitable for them to do so, and no amount of exhortation or extra funds for investment will change the amount or prestige of such investments. If the lack of process R&D is a real problem, then one must face the fact that somehow the market mechanism has failed. No government bureaucrat gave orders to avoid investment in process R&D or tinkered with the tax system to thwart such investment.

In an article in *Technology Review,* "Designing the Designers: Computer R&D in the United States and Japan," D. Eleanor Westney and Kiyonori Sakakibara describe the process of moving technologies into production in the computer industry in Japan. Instead of the rigid separation between R&D, manufacturing and management common in the U.S., new engineers first begin in research and then essentially follow their ideas into production engineering and finally into management. Who could have a better incentive to manufacture a new product efficiently than those who first thought it up? It is just such organizational changes that will have to occur if America is to regain its edge in process innovation. And if American firms are not willing to make such changes in their standard operating procedures, then new products, such as semiconductor chips, will remain American products for progressively shorter periods of time.

Other forms of social organization also play a role in suppressing productivity. Although there is an article on the role of large banks in financing industry by John Reed, the chairman of Citicorp, and his colleague Glen Moreno, nowhere is there a discussion of how the bold forays of financial vikings affect new technologies and the growth of productivity. In the American financial system the highly leveraged firms that are produced by mergers or that become highly leveraged in order to fight off the raiders have no funds (after mandated interest payments are made) to invest in new technologies. In a climate of merger mania, industry becomes a plaything of finance. Assets are reshuffled, fortunes are made, yet the ultimate source of more wealth, higher productivity, grows ever more slowly.

In Japan the Minister of International Trade and Industry is regarded as more powerful than the Minister of Finance. In the U.S. to even compare the power of the Secretary of Commerce with that of the Secretary of the Treasury is absurd.

Such perceptions underscore what a society thinks is important. So does a country's behavior when what is good for finance conflicts with what is good for industry. In January, 1985, the British pound hit $1.04. Britain's manufacturers should have been dancing in the streets. For the first time in 50 years they would have been able to expand their world market share. Yet it was a disaster for the City of London. Who would want to hold British financial assets with such a value for the pound? Overnight the British government raised interest rates from 9 to 14 percent to restore the position of its financial firms, and in doing so it foreclosed the opportunities that might have been open to its industrial firms. When it comes to such relationships, the U.S. is fundamentally Anglo-Saxon: if it is forced to choose, it will be pro-finance and anti-industry. In order to compete with countries that emphasize industry, that bias may have to be turned upside down.

This collection of essays might also have benefited from a historian's view of American economic development. Harvey Brooks comments that "the demand for durable steel rails for the railroads was a major factor driving technological innovation in the burgeoning steel industry"; he does not mention that the more efficient British steel industry was kept out of the

American rail market by the protectionist policies of the U.S. Government. Nor does Brooks remind readers of the book that those railroads were also built with the aid of government land grants.

The first unique contribution of the Federal Government to progress in industry was the concept of interchangeable parts. Eli Whitney developed that manufacturing system in 1798 to fulfill a contract from the War Department to make muskets. No one in the private sector would take a risk on mass production, and the initial contract in fact paid Whitney more than it would have cost the Continental Congress to buy handmade muskets in Europe. He also delivered late and incurred large cost overruns, which were picked up by the Government. Yet the early and widespread use of interchangeable parts gave America an edge that enabled it to catch up with Britain economically during the 19th century.

The nation's civilian aviation industry was built on aircraft that had been developed for military transportation. Government construction projects such as the interstate highway system created the internal markets that allowed construction-machinery manufacturers to gain the economies of scale enabling them to dominate world markets for many years.

America's own history is far more complicated than simply letting unfettered markets rip.

The editors and authors frequently refer to technology as a black box whose intricate inner workings no one understands, whereas they imply that the market mechanism is crystal clear. Pump in high-quality inputs, keep government from fouling up the mechanism, and the market will always yield high-quality outputs.

As economists and managers know too well, the market mechanism is unfortunately at least as opaque as technology. Consider the prescription of increasing the economic rewards (reducing the taxes) for those much prized individual risk takers. At no time did technical progress occur faster than it did during World War II, yet at no time in American history did individuals reap less of their rewards in the form of personal income or wealth. Maximum tax rates were in excess of 90 percent. Today's taxes on high-income individuals are far below those paid at any time since the Great Depression, yet the adoption of new technologies and productivity growth is slower than it has been at any time since the onset of the Industrial Revolution. That clear market mechanism grows ever more opaque.

Established 1845 SCIENTIFIC AMERICAN September 1986 Volume 255 Number 3

The Microwave Problem

Is exposure to low levels of microwaves a hazard? How strict should exposure limits be? These issues remain in dispute in part because some findings on the biological effects of microwaves are ambiguous

by Kenneth R. Foster and Arthur W. Guy

Air-traffic-control systems, police and military radar, earth-to-satellite television broadcast systems, long-distance telephone equipment, medical diathermy devices and microwave ovens all generate microwaves. Except when microwaves are deliberately applied to the body for therapy, little of this invisible energy ever reaches the public. Is exposure to such low levels of microwave energy hazardous to human health?

This question is easier to ask than to answer. The interaction of microwaves and living organisms has been studied much more extensively than most other potential environmental hazards have, and yet the result has been continuing public and scientific controversy.

Communications equipment is a source of widespread—albeit extremely weak—public exposure to microwaves. A typical dispute arises when a company petitions a county zoning board for permission to install a microwave communications facility. The company's spokesmen might explain that the intensities, or power levels, of transmission will result in peak exposure levels that are some thousandths or millionths as high as allowable limits. These limits in turn are considerably below the levels known to produce biological damage.

Opponents argue that low levels of microwave energy may pose some as yet unproved danger and that their safety is still to be demonstrated. They point to the many biological effects that have been suggested at one time or another (by animal studies) to be related to exposure to low-level microwave energy, such as changes in immune-system functions, altered behavior, changes in the permeability of the brain to molecules carried in the blood, damage to chromosomes and the development of cancer.

Controversy continues in part because some of the data provided by approximately 6,000 studies in the 40 years since microwave technology was introduced are inconsistent and inconclusive. Whereas it is known that exposure to high levels of microwaves can burn human tissues or cause heat stress, no clear-cut damage to human beings from low-level radiation has been demonstrated. On the other hand, exposure to low levels of microwaves cannot be proved free of hazards. This lack of consensus does not result from flawed research alone. The cause is more fundamental: the normal process of risk assessment yields data that can be subject to different interpretations and can generate controversy, whether or not a hazard is ever demonstrated.

Given the inconclusive state of the published evidence, we do not argue here that exposure to low-level microwave energy is either hazardous or safe. Rather, we shall describe the process by which known hazards have been quantified. We shall also review the development of the first major American standard for limiting exposure to microwaves and the rationale behind the standard. We shall then review several case histories that indicate why assessing the possible hazards of microwaves—and low-level environmental agents of all kinds—is so challenging.

Concern over the biological effects of microwaves must be viewed in the context of experience with lower-frequency electric fields, which were exploited for technology before microwaves were. Microwaves are a part of the electromagnetic spectrum, and their band extends from 300 megahertz to 300 gigahertz; that is, it includes waves with oscillation frequencies ranging from 300 million hertz, or cycles per second, to 300 billion hertz. Those frequencies are higher (and the wavelengths corresponding to the frequencies are shorter) than those of standard radio and television signals. Above the microwave band there are, in order of increasing frequency, infrared radiation, visible light and "ionizing" radiation: ultraviolet radiation, X rays and gamma rays.

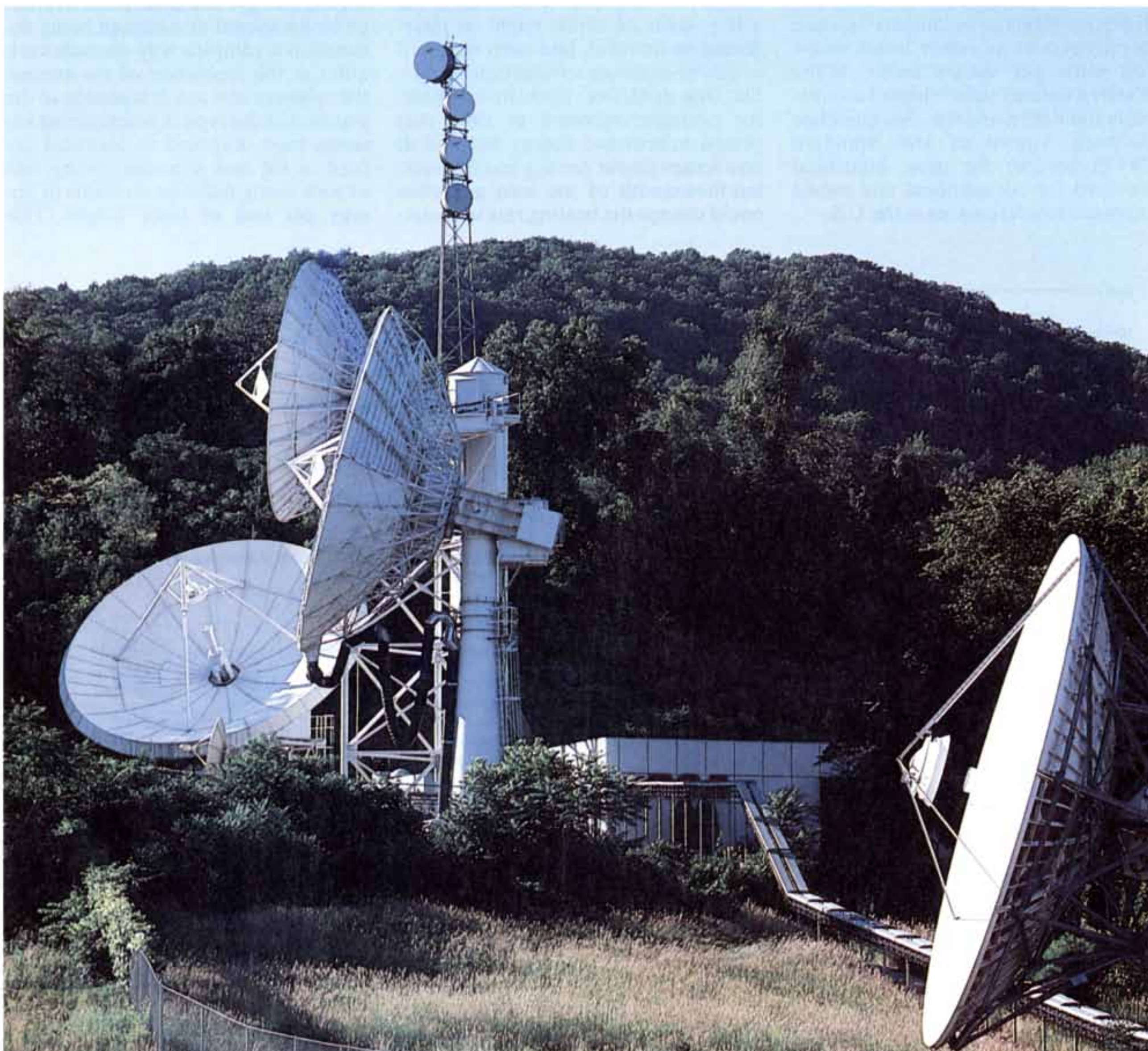
We should point out that ionizing radiation is qualitatively very different from microwave energy in its effects on biological systems. As one photon, or energy packet, of an ionizing ray passes through a substance, the photon breaks chemical bonds (even in the absence of any appreciable heating) and causes neutral molecules to become charged. Such ionization can damage tissues. In contrast, the energy of a photon of a one-gigahertz microwave is only one six-thousandth of the kinetic energy possessed by a molecule in the body owing to normal thermal agitation, even less than the energy needed to break the weakest chemical bond. This does not rule out the possibility that weak microwave energy can directly alter molecules of tissue, but it does not suggest a mechanism by which significant changes could occur.

By the time microwave technology was introduced, during World War II (when radar changed the course of the war), electromagnetic fields at lower frequencies had been employed for therapeutic heating for most of the century. Although the value of therapeutic heating was not disputed, the one obvious potential hazard was that of burns or other damage resulting from excessive tissue heating. In addition controversy raged among investigators over claims that high-frequency electromagnetic energy produces biological effects in various organisms, effects that some believed did not arise from simple heating alone.

In the decade after World War II microwave technology continued to develop rapidly, and with this progress came widespread human exposure to the energy. Although few injuries of any kind were reported, the possibility of hazard became increasingly disturbing. Because of the military's extensive application of relatively high-powered radar equipment, the services took a lead in investigating microwaves, establishing research programs in the early 1950's that continue in expanded form today. Also in the early 1950's, many investigators and Government officials began to consider setting exposure guidelines.

One of those people was Herman P. Schwan of the University of Pennsylvania. Based on theoretical estimates of tissue heating, he recommended in 1953 that human exposure to microwave energy be limited to a maximum average "power density" of 100 watts per square meter.

Schwan's calculations showed that exposure to this incident power level should raise the temperature of any region of the body by one degree Celsius or less, and add heat at a rate comparable to that generated by the body's normal physiological processes. The limit was roughly one-tenth as intense as bright sunlight and perhaps one-fiftieth as intense as the power levels of diathermy equipment. Schwan



ANTENNAS (*large "dishes"*) **at a communications station in Vernon, N.J., transmit microwave signals to satellites; other antennas** (*smaller disks*) **relay energy to receivers on the earth. The signals are put to a variety of purposes, including satellite television broadcasts. Transmissions from these antennas, and those from the many others that make Vernon a hub of microwave satellite communications, result in exposure levels to the public that are well below the safety limits imposed by the state. (Backyard dishes that merely receive transmissions are not a source of exposure.) Vernon is one of several places in the U.S. where citizens' groups have protested the installation of microwave generators, fearing that exposure to even low levels of microwave energy may pose health hazards.**

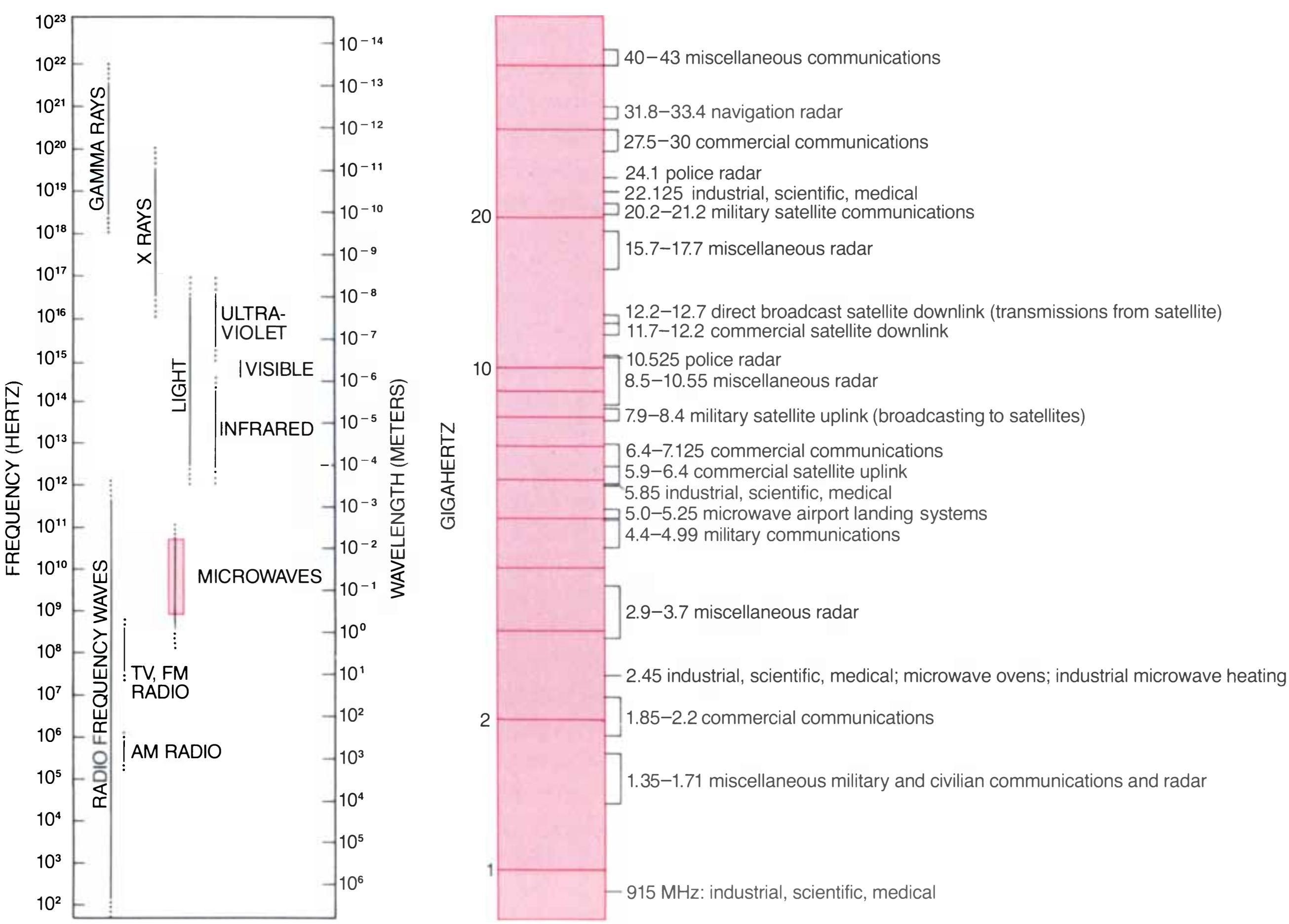
also calculated that power densities of about 1,000 watts per square meter might produce heat damage to the body in some circumstances. The proposed limit, then, allowed a safety margin of roughly 10.

Schwan directed his advice to the U.S. Navy, which had expressed concern about microwave safety, and his proposal was eventually embodied in a formal standard by the United States of America Standards Institute, later renamed the American National Standards Institute (ANSI). ANSI, a private organization that recommends safety guidelines of many kinds for different industries, adopted the standard in 1966 after reviewing the scientific literature and finding no convincing evidence for damage in animals exposed to microwaves at power levels below 100 watts per square meter. Many Western nations soon adopted comparable standards, and the ANSI guideline (formally known as ANSI Standard C95.1) became the most influential standard for occupational and public exposure to microwaves in the U.S.

The original ANSI standard, which covered frequencies between 10 megahertz and 100 gigahertz, remained essentially the same for many years, but by 1982 ANSI's periodic reexamination of the guideline suggested a change was in order. By that time too the safety of exposure to low-level microwave fields—ones transmitted at power levels too low to cause significant heating—had become a subject of popular, political and scientific debate.

Public concern had been heightened by several factors. Between the 1940's and the 1970's many studies had addressed the biological effects of microwave energy. Most of the reported effects occurred at power densities above 100 watts per square meter, but a few, some of which might be interpreted as harmful, had been reported at power densities substantially below the ANSI guideline. (One investigator, for example, reported in 1968 that pulsed microwave energy beamed at an average power density less than one ten-thousandth of the ANSI guideline could change the beating rate of isolated frog hearts, sometimes stopping a heart entirely. That finding could not be confirmed by other investigators.) People had also heard news reports that the Soviet Union had for many years beamed low levels of microwave energy at the American Embassy in Moscow, reports that often included speculations about health effects. In addition the Soviet Union and the Warsaw Pact countries had set exposure limits for the general population at levels 100 to 1,000 times lower than any American standard.

The major impetus for ANSI's revision of the guideline was data from studies employing improved methods of dosimetry, or measurement of absorbed energy. The absorption of energy by an animal or a human being depends in a complex way on such variables as the frequency of the energy, the subject's size and orientation to the waves, and the type of transmitting antenna used. Exposed to identical energy, a rat and a human being will absorb vastly different amounts of energy per unit of body weight. This

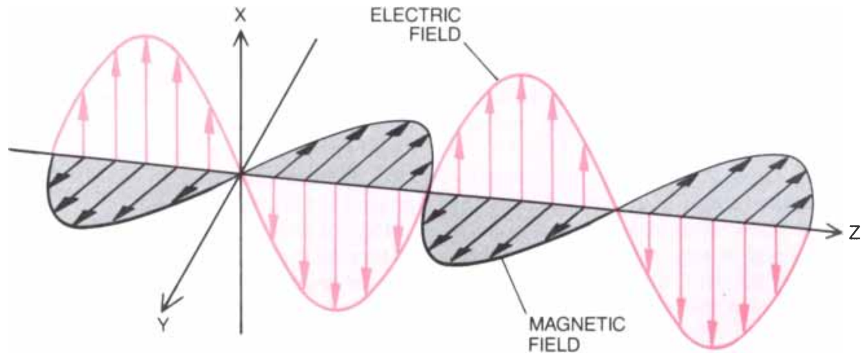


MICROWAVE BAND of the electromagnetic spectrum (*left*) extends from approximately 300 million to approximately 300 billion hertz. Microwaves are exploited for an array of applications, only a few of which are listed at the right. Military and civilian radar and communications functions are concentrated in the range of .3 to 40 gigahertz (billions of hertz). Medical, industrial and scientific technologies employ several frequencies, but the ones most commonly applied are 915 megahertz (millions of hertz) and 2.45 gigahertz.

complexity had long been recognized, but before the late 1960's few investigators carefully measured the amount of energy absorbed by their subjects. Beginning in the 1960's engineers undertook studies that led to a fuller understanding of energy absorption and to better experimental techniques for measuring it in animals.

Several methods developed in the past two decades calculate absorbed energy by measuring the temperature rise in models of living subjects. One of us (Guy) introduced the "split phantom" technique in 1968. Workers shape plastic foam into a hollow model of an animal, fill it with a gel that has electrical properties resembling those of living tissue and briefly irradiate it. Then they split the model open and photograph it with an infrared camera to record temperature increases. Workers can also determine energy-absorption patterns by computer simulation. The animal or exposed person is modeled as a simple ellipsoid or cylinder. More elaborate mathematical models, consisting of block figures, have also been studied.

By 1982 all the new methods agreed that the amount of energy absorbed varies widely with the frequency of the radiation and the size of the body. Under unfavorable circumstances a person might absorb up to 10 times as much energy at frequencies between 70 and 100 megahertz as at higher frequencies, depending on the individual's orientation to the waves. Electromagnetic waves consist of electric and magnetic fields that are perpendicular to each other and to the direction in which the waves are traveling [*see top illustration on this page*]. Maximum absorption occurs when waves impinge on the body from the side, where the electric field is parallel to the body's long axis and the magnetic field is perpendicular to the front of the subject. The human body is an efficient antenna for waves in the 70-to-100-megahertz range; it is said to "resonate" with the fields at those frequencies.

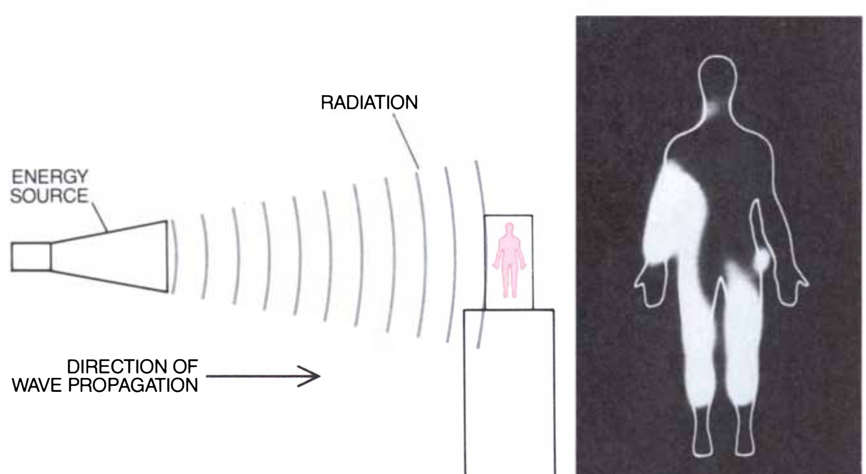


ELECTRIC AND MAGNETIC FIELDS of microwaves and other forms of electromagnetic radiation are perpendicular to each other. They are also perpendicular to the direction of wave propagation. When the radiation travels in the direction that is shown by the *z* axis, the electric field (*color*) is parallel to the *x* axis; the magnetic field is parallel to the *y* axis.

To adjust for these resonance effects, in 1982 an ANSI committee reviewing the standard decided to make power-density limits dependent on frequency, with the overall goal being to limit the energy absorbed by the body. Whereas the old guideline had specified only the intensity of the energy incident on the body, the new one attempted to limit average entire-body absorption levels to .4 watt per kilogram of body weight. (Somewhat higher peak absorption levels might be allowed for partial-body exposure.) When the body is at rest, it normally generates heat at twice that rate, and it generates much more heat during moderate exercise. Compared with the previous guideline, the 1982 standard called for significantly reduced power densities at resonance frequencies, and it applied to a wider range of frequencies, from 300 kilohertz to 100 gigahertz [*see top illustration on page 38*]. The incident power level was calculated by averaging power over six-minute periods; thus the standard allowed brief exposure to high power levels.

ANSI approved the 1982 standard only after its review of the scientific literature uncovered no convincing evidence for health-damaging effects in animals exposed to energy producing absorption rates below .4 watt per kilogram. The committee concluded that the threshold for possible hazards was an absorbed power level of roughly four watts per kilogram of body weight; in other words, the standard had a safety factor of about 10 built into it. The committee further concluded that the proposed standard would exclude heat stress and burns. It would also avoid other reported effects for whose existence and undesirability the evidence was considered reliable (such as serious disruption in the behavior of several kinds of animals at absorption levels of four to eight watts per kilogram).

The standard did not attempt to avoid all reported effects, because the

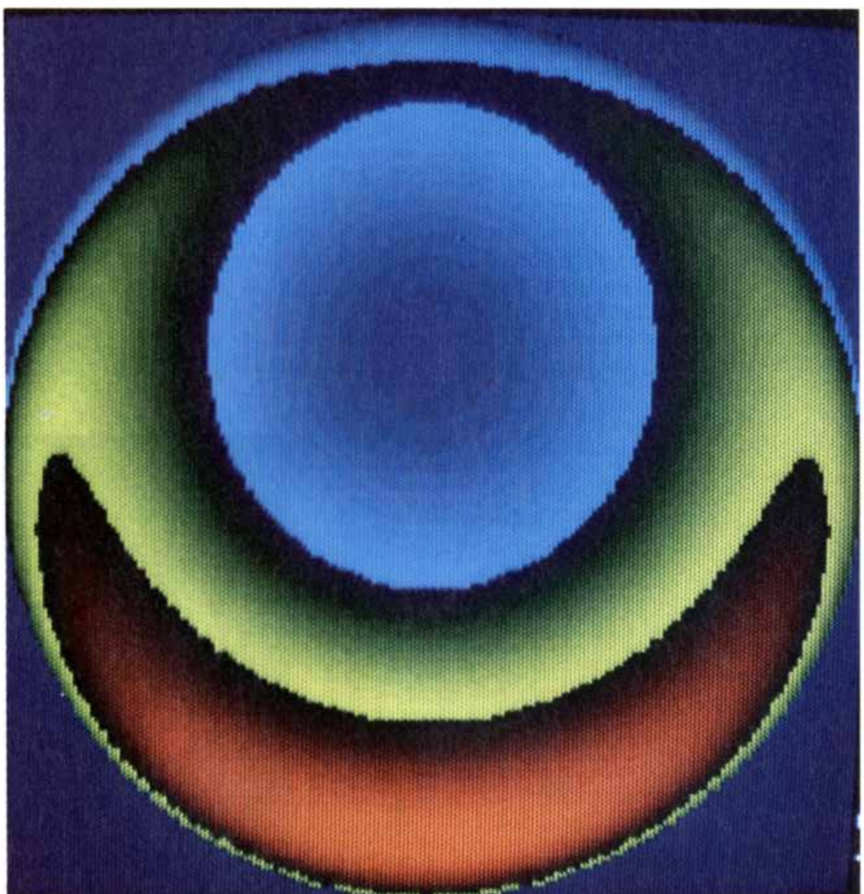


"SPLIT PHANTOM" TECHNIQUE for dosimetry measurement of absorbed energy was introduced by one of the authors (Guy) in 1968. The technique measures the temperature rise in models of human beings or animals and reveals the pattern, depth and amount of energy that would be absorbed by a living creature. Workers hollow out two halves of a Styrofoam block to form a mold of the body and then fill the mold with a gel that has electrical properties resembling those of living tissue. The model is briefly irradiated (*left*) and quickly split apart, and the interior is photographed with an infrared camera to record the amount of heating in various regions. The photograph at the right reveals the pattern of energy deposition produced in one such model, which, for practical reasons, is smaller than life size. The pattern is comparable to that produced by 79-megahertz waves (a frequency resulting in maximum absorption by humans) striking the side of a man weighing 70 kilograms and standing 1.74 meters tall. The bright areas indicate relatively high absorption.

literature relating to many of the effects was—and still is—problematic. Many studies have shown that exposure to high levels of microwaves is clearly hazardous, producing obvious heat stress in animals. Other studies, in which absorption levels are comparable to the rate of heat generated by the body, have observed changes that in part could be normal physiological responses to the added heat, although evidence for this inference is often less clear. Other effects have been reported at quite low power levels; no obvious explanation has been found.

There is also great diversity in the quality of the evidence. Although many studies reporting effects on living systems have apparently been well done, some have had obvious technical flaws (in particular, some have lacked adequate dosimetry) and others have been too briefly described to allow any judgment of their quality. Of the hundreds of biological effects of high or low levels of microwaves that have been reported, a surprising number are examples of the "Cheshire cat" phenomenon: they have not reappeared in follow-up studies.

Part of this disparity arises from the nature of the research itself. An investigator might report an "effect" on the basis of some difference found between control subjects and those exposed to microwaves. The reported effect might well arise from some specific biological activity of the energy. On the other hand, it might also result from a normal physiological response to added heat, from a statistical fluctuation or even from some experimental variable that was not adequately controlled by the investigators.



COMPUTER-GENERATED SPHERE can be employed to study energy-absorption patterns in tissues and is particularly valuable for determining how changes in frequency alter the distribution of energy within irradiated subjects. The sphere, which can represent a person or a part of the body, is evaluated as if it had electrical and heat-transporting properties resembling those of human tissues. With the computer an investigator calculates the energy-deposition patterns, assigning different colors to represent amounts of temperature increase. When a sphere with a 10-centimeter radius is irradiated at a frequency of 100 megahertz, the front surface absorbs the greatest amount of energy (*red*); the region just behind the surface absorbs a moderate amount (*green*) and the center of the sphere absorbs very little energy (*blue*). At an incident power level of 10 watts per square meter, the maximum calculated temperature rise in this sphere is minuscule: .003 degree Celsius. The image shown here was made by Haralambos N. Kritikos of the University of Pennsylvania.

To illustrate the ease with which data can be subject to various interpretations and produce apprehension, we have chosen three case studies. All the reported findings were observed at power levels that are within the ANSI guideline in effect before 1982, and all have played some role in the public controversy about possible hazards of low-level microwave energy.

The first case involves the possible influence of low-level microwave energy on brain function. In 1975 two American teams of workers reported that exposure to power levels from one-third to one-fiftieth of the ANSI standard could increase the rate at which tracer molecules enter the brain from the blood. Although the exact significance of this finding for human beings was difficult to judge, any such effect could be construed as disturbing the blood-brain barrier and would be cause for concern. Predictably, the report fueled the microwave-safety controversy. A dozen research groups followed up on this finding during the next decade. As studies continued and were progressively better controlled at low exposure levels, the effect went away. (Most investigators agree, however, that exposure to energy levels high enough to significantly heat the brain would produce substantial alterations in the barrier.)

The second case study involves the "microwave auditory effect," which has been known since 1947. When a subject's head is exposed to pulses of microwave energy, the person can often hear "clicks" in synchrony with the pulses; these clicks seem to originate from within the head. To be heard, the pulses must be relatively intense (on the order of 10,000 to 500,000 watts per square meter); they can be brief enough (microseconds), however, to result in a rate of energy absorption that, when averaged over time, falls well below maximum safety levels.

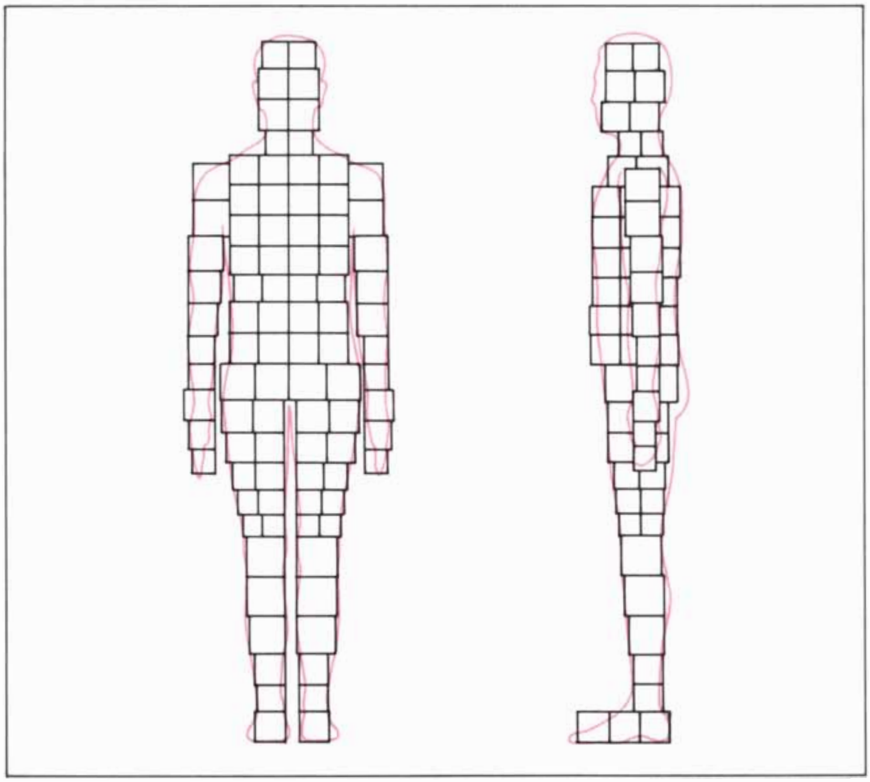
Results of one early study on the effect suggested that the center of the brain was the most sensitive region for producing the clicks. The alarming possibility thus arose that the pulses of microwave energy might somehow act directly on the brain. In 1974 one of us (Foster) proposed that the clicks might result from a benign physical effect accompanying the absorption of the energy by the head. The proposed mechanism was simple: the thermal expansion of tissue, caused by minuscule but abrupt heating (a few millionths of a degree following each pulse), launches sound waves that the subject perceives as clicks. Simple calculations and experiments using water-filled models showed that sound waves generated by the microwave pulses should indeed be audible. Soon after, studies with animals confirmed the hypothesis. The microwave auditory effect is not now regarded as pointing to a hazard.

Our final illustration is an extended study conducted by one of us (Guy) with his colleagues at the University of Washington at Seattle. Its results reveal typical ambiguities that can result from screening studies, which are the source of many reports of microwave effects. The three-year study of the effects of long-term, low-level irradiation was funded by the U.S. Air Force School of Aerospace Medicine. It compared 100 rats that were irradiated for most of their lives with 100 rats that were not exposed to radiation but were otherwise treated identically. The radiation beamed at the experimental group had an average power level of five watts per square meter and a frequency of 2.45 gigahertz. The rats were exposed for 21 hours per day for 25 months. On the average, depending on age, they absorbed from .2 to .4 watt per kilogram of body weight, the latter being the current ANSI exposure limit for human beings.

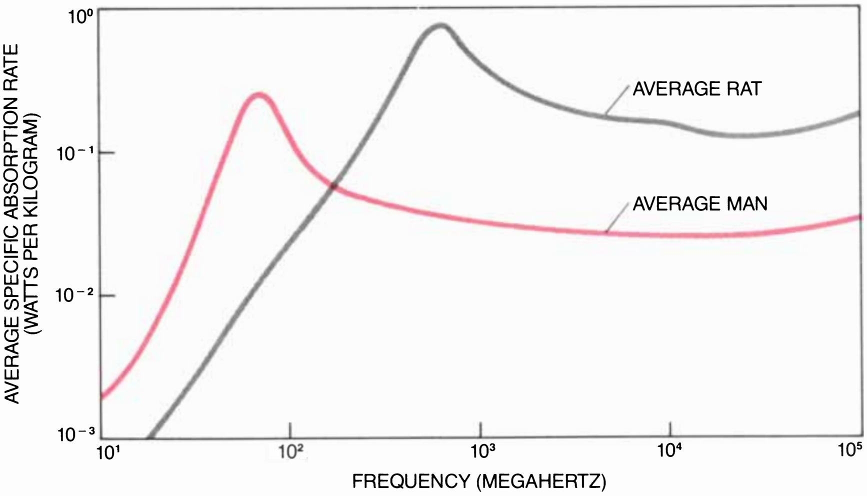
The investigators examined 155 different measures of health and behavior, including blood chemistry, body weight, daily food and water consumption, oxygen consumption, carbon dioxide production and activity level. The results revealed few differences between the exposed and the control rats, and those differences for the most part were either not statistically significant or came and went, suggesting that they might be due to chance. For example, plasma cortisone levels (which indicate the level of arousal) were higher for the exposed group during the first sampling session but were higher for the controls during the third session.

Earlier studies had suggested that microwaves might impair the immune system. The Seattle group therefore evaluated the function of lymphocytes, a cell type fundamental to the immune response. Some but not all of the immunological tests showed a difference between the exposed and the control animals after 13 months. After 25 months the differences were no longer discernible. The mean survival time of the exposed animals was slightly longer than that of the control animals: 688 days v. 663 days—a difference that was probably due to chance alone.

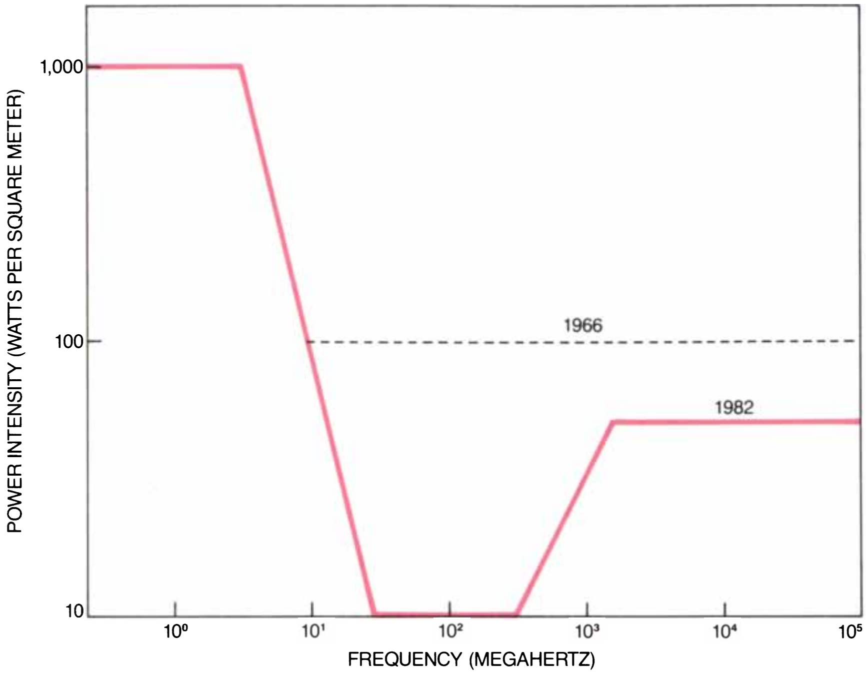
One difference was striking: primary malignant tumors developed in 18 of the exposed animals but in only five of the controls. The probability



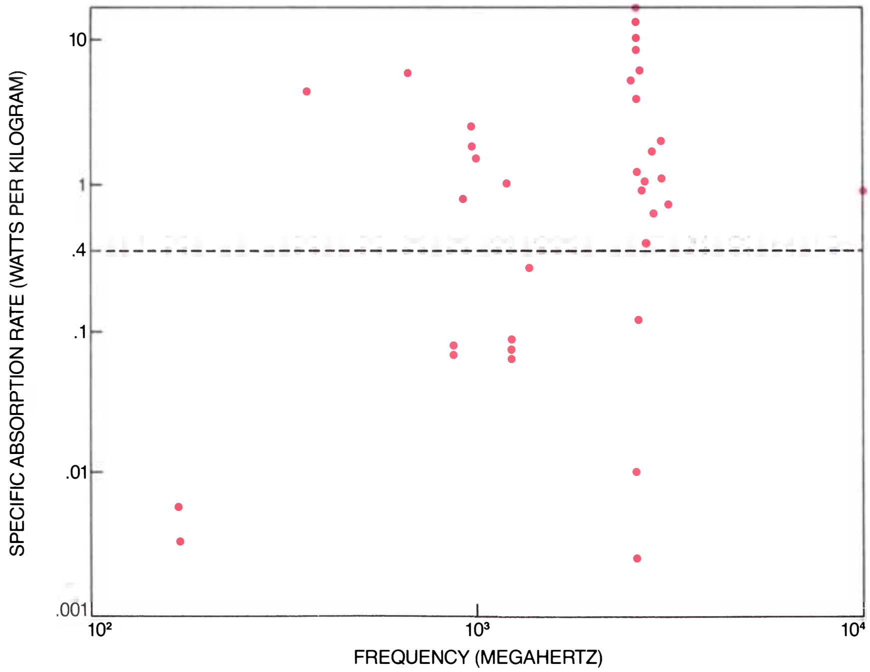
BLOCK FIGURES are the most complex of all theoretical models employed to calculate energy absorption. Workers load the coordinates of each block into a computer. Then, on the basis of presumed electrical properties of the tissues, radiation frequency and other factors, the computer calculates the specific absorption rate (SAR): the watts of energy absorbed per kilogram of weight when a subject is exposed to a power density of 10 watts per square meter. Other theoretical models (not shown) include ellipsoids that are assumed to be filled with homogeneous material representing human tissues and ellipsoids that are assumed to contain separate layers representing fat and muscle. Results from all these theoretical approaches agree well with one another and with data derived from experiments with irradiated models, increasing confidence in the accuracy of the various calculations.



RATES OF ENERGY ABSORPTION by an average-size rat and man differ at most frequencies and also peak at different frequencies, as is shown by comparisons of SAR's in ellipsoid models. Man has a pronounced increase in SAR at approximately 70 megahertz, the rat near 1,000 megahertz. The large differences in SAR's between the species arise from the differing properties of their bodies as antennas. Such differences, together with differing physiological responses to heat, are important sources of uncertainty when investigators attempt to extrapolate from animal data to determine a power threshold above which microwaves pose a hazard to human beings. The curves reflect absorption by subjects exposed to waves propagated toward the body at an angle that most promotes energy absorption. (The electric field is perpendicular to the long axis of the body; the magnetic field is perpendicular to the frontal plane.) The data are from Carl H. Durney of the University of Utah.



WIDELY RECOGNIZED STANDARD for maximum public and occupational exposure to microwaves was established in 1966 by the American National Standards Institute (ANSI) and was significantly revised in 1982. In 1966 the voluntary standard allowed the maximum incident power level to be 100 watts per square meter for frequencies ranging from 10 megahertz to 100 gigahertz (100,000 megahertz). An ANSI committee revised the limit in 1982 to reflect the finding that the human body absorbs more energy at some frequencies than it does at others. The 1982 version makes power densities dependent on frequency, lowering allowed power levels for frequencies in the neighborhood of 100 megahertz. The standard's goal is to limit absorption to .4 watt per kilogram, averaged over the entire body. If only part of the body is exposed, higher power intensities are sometimes permissible.



REPORTS OF BIOLOGICAL EFFECTS were plotted (*dots*) by the ANSI committee according to SAR and frequency before the organization revised its exposure standard in 1982. For effects reported to occur at SAR's above five watts per kilogram, different studies generally agreed on the nature of the effect and the SAR's that would reliably produce it. For effects reported at lower power levels, different studies showed much less agreement. Many of the effects reported at the lower levels were not considered indicative of a hazard. ANSI further determined that reported effects that could be considered hazardous occurred at an entire-body SAR of four watts per kilogram and above; the organization arrived at the 1982 SAR limit of .4 watt per kilogram (*broken line*) by then building in a safety factor of 10.

of such a difference occurring in two samples from an identical population of only 100 animals each is roughly .005, and so the difference is statistically highly significant.

At face value this last finding suggested that low levels of microwave radiation can cause cancer in mice (and by inference in humans). The finding was widely reported by the lay media in 1984 and has been frequently cited in public disputes over proposed microwave facilities. Nevertheless, various considerations militate against drawing a hasty conclusion.

For one thing, the total number of malignant tumors in the control animals was lower than the number expected for the particular strain of rat; the rate of malignancies in the exposed rats was about as expected. Thus the exposed animals had an excess of tumors only in comparison with the controls, not in comparison with the rate of tumor development generally observed in this strain of animal.

Other problems, of a statistical nature, also arose. So few malignancies were found that tumors of all kinds had to be grouped in the statistical analysis. No single type of tumor predominated, nor was the incidence of any single type of tumor unexpected on the basis of previous studies. If some specific type of tumor had predominated, that finding would have made a much stronger case for a carcinogenic effect from low levels of microwave energy.

The comparison of rates of malignancy was just one of 155 different comparisons made in the study. Given such a large number of comparisons, some striking differences would be likely to be found that are in fact merely chance occurrences. The cancer finding may be such a statistical anomaly. In short, the finding of excess cancer is provocative, but whether it reflects a biological activity of microwave radiation is not certain. To demonstrate reliably a connection between microwave irradiation and the development of any one kind of tumor might require a study hundreds of times larger and more expensive than the Seattle study—a size that might be infeasible. Our conclusion from these examples and from the large literature on microwaves is that although some hazard from weak microwave fields might be proved in the future, there is currently little evidence for the presence of such a hazard.

Considering that some uncertainty persists, how should future research proceed, and on what basis should any new standards be set? During the 1980's U.S. Government support of investigation into the possible biological effects of electromagnetic fields, including those at power-line and radio frequencies, has been at the level of about $10 million per year. The resulting work, taken together with earlier studies, has produced a consensus among most investigators that the only strong evidence for the hazards of microwaves is found at high levels of exposure. Beyond this, agreement is harder to come by. Scientists disagree over the meaning of effects that have been reported to occur in animals at exposure levels slightly below the ANSI standard. Such disagreements are not likely to result in drastic changes in exposure standards. Public debate, on the other hand, has often focused on the possibility of hazards at much lower levels of exposure and on the possible need for more stringent standards.

Better coordination of future research should reduce some of the scientific and public confusion. Many Government agencies, including the military, have funded investigations of the biological effects of microwaves. Lack of coordination among these agencies has unfortunately resulted in a scatter-gun approach in which many preliminary studies have been carried out but not followed up.

Before new studies are undertaken the Government should commit itself to supporting independent attempts at duplication and also supporting follow-up studies to explore the significance of new findings. Some agencies appear to be moving in this direction, with the Army, the Air Force and the Navy now funding follow-up studies on controversial recent reports. Conversely, some criteria must be developed for determining when to halt research on a given topic, open questions notwithstanding.

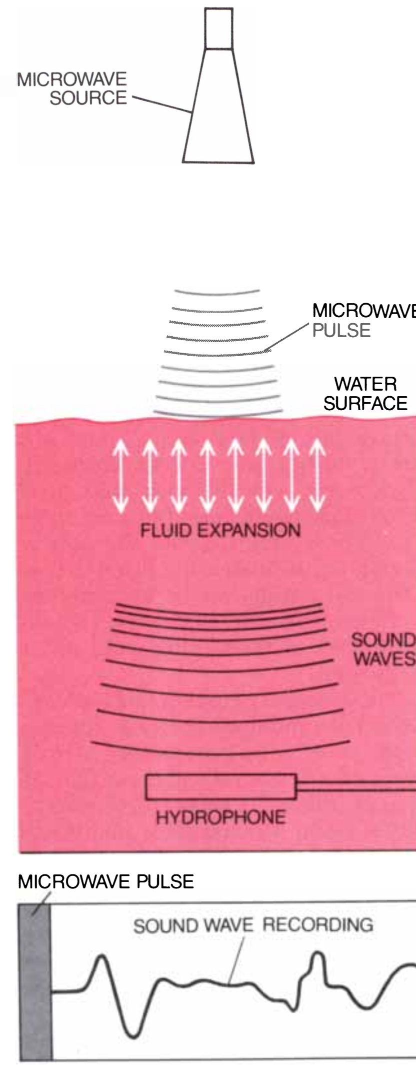
Even with improved coordination of research, it will not be possible to prove the absence of hazard, which means that regulators cannot guarantee total safety when they set standards. Standards are one way by which society balances the benefits of technology against potential risks. Traffic speed limits are set at a level that achieves some balance between the danger of excessive speed and the desire of most people to travel as quickly as possible. The limits offer no promise of zero risk at lower speeds but merely legitimize the speeds at which known hazards are unlikely to be a problem. Similarly, exposure limits for microwaves (and other environmental agents) can only be based on known hazards, with a safety factor built in. The revision of the 1982 ANSI standard is an example of this approach. When improved dosimetry revealed that the human body can absorb more energy at specific frequencies, ANSI in effect built in a wider safety margin for exposure at those frequencies.

The U.S. has no consistent approach to the setting of standards. Individual Federal agencies have been setting their own limits, as have some individual states, counties and nongovernmental organizations. Different regulations apply in different situations and have varying degrees of legal force. Standards can be voluntary or mandatory and can apply to occupational groups or to the general public. A town can impose regulations that differ from those of its county, which itself may impose regulations that differ from those of the relevant Federal agency. Such divergence can force users of the electromagnetic spectrum to abide by differing standards; it also creates burdens on the various governmental agencies.

The U.S. Environmental Protection Agency has begun a process that may eventually result in a more uniform standard for the general public. Earlier this year it published for public comment three possible guidelines: one similar to (but not identical with) the ANSI guideline, another five times lower than the ANSI guideline and one 10 times lower than the ANSI guideline. The EPA is probably more than a year away from promulgating its final standard, but if a regulation is adopted, it is likely to preempt other efforts to set standards for the general population.

It is a curious fact that the Soviet Union and the Warsaw Pact countries have recently relaxed their standards. At least one Soviet commentator has stated that an acceptable level of exposure is .4 watt per kilogram for one hour, comparable to the present ANSI limit of .4 watt per kilogram averaged over six minutes. The current Soviet standards for occupational exposure to microwave energy are two watts per square meter for stationary sources and 20 watts per square meter for mobile sources, both averaged over one hour. The levels for the general public are lower: .1 watt per square meter. Thus the Soviet and the American standards are approaching agreement.

The setting of standards is a surprisingly difficult process. Some hazards of microwave energy, such as burns or heat stress, are well established; effective standards can be set for these. Whether other hazards can arise from exposure to low levels of microwave energy is a matter of conjecture, which depends on the interpretation of a large body of often unreliable reports. How to deal with the uncertainty, balancing the benefits of technology against the costs of possible hazards, is an urgent problem, not only with respect to microwave energy but also for many other environmental agents whose potential hazards have only begun to be explored.



MICROWAVE AUDITORY EFFECT, a clicking noise heard when microwave pulses are beamed at the head, was explained in 1974 by one of the authors (Foster). He proposed that the head responds to microwaves much as water does. When water absorbs pulses of microwaves, it undergoes a rapid but tiny rise in temperature. The resulting expansion of the fluid generates a sound wave that propagates from the surface of the water and can be heard as a click or recorded by a hydrophone. The same process, Foster suggested, occurs in tissues of the head on exposure to pulsed microwaves. The hearing of clicks is one of the few unequivocal effects of microwave energy at average power densities that could be below the ANSI standard. The auditory effect was first reported in 1947 but its mechanism was only explained years later. The effect is not now considered to represent a hazard.

Predicting Chemistry from Topology

Methods relying on the topology of molecules—the geometric patterns in which their atoms are linked—but ignoring their three-dimensional shapes predict a broad range of properties

by Dennis H. Rouvray

Chemists have managed so far to synthesize and measure at least some of the properties of more than seven million different molecules. The huge data bases containing this store of information can be drawn on by the research chemist, providing nearly instant knowledge about any molecule that has ever been isolated. Is there some way this vast amount of knowledge could be put to even fuller use? For example, could the data be drawn on to predict the properties of chemical substances before they are even synthesized—that is, before any molecules of those substances physically exist?

The prospect of making such predictions has tantalized chemists for generations. It is now becoming a reality, thanks to a novel technique that, although still in its early stages, can already claim a remarkable number of successes in a surprisingly broad range of applications. At the heart of the new technique is the topology of individual molecules: the pattern of interconnections among each molecule's atoms, which determines the ultimate architecture of the molecule.

For the sake of topological analysis, the actual three-dimensional shape of a molecule, the nature and lengths of the chemical bonds connecting its atoms, the angles between the bonds—and sometimes even the kinds of atoms that make up the molecule—are all unimportant. What matters are such considerations as how many atoms there are in the molecule, how many other atoms each is connected to within the molecule, and whether the atoms are connected to form a single straight chain, a straight chain with branches, a ring (or several rings) or some combination of rings, straight chains and branches.

The most important tools in the topological method of making chemical predictions are known as indices. They derive from algorithms, or procedures, for converting the topological structure of a molecule into a single, characteristic number. For example, an index might involve adding together the total number of rings in a molecule or the number of atoms that are connected to three or more other atoms.

In making a topology-based chemical prediction the first step is to apply some index to a relatively small number of well-known molecules. In general the value of the index (that is, the number yielded by that particular way of analyzing the molecule) will differ from molecule to molecule. The next step is to construct a plot in which one axis represents the index value of the molecules and the other axis represents some chemical property, such as the molecules' boiling points. Each molecule will be represented by one point in the plot. If there is some line that fits the points reasonably well, that line can serve as the basis for predictions because it establishes a relation between the index and the chemical property: it gives an indication of what boiling point, for example, to expect for a molecule that has a given index value. In this way the properties of well-known molecules serve as a tool for predicting the properties of molecules that do not yet exist. The key to the method lies in finding the index that correlates best with the chemical property being studied.

The topological method has found applications beyond the simple prediction of chemical properties. It has the potential to help in modeling and controlling corrosion, in developing new anesthetics and psychoactive drugs, in predicting the degree to which various pollutants might spread in the environment and the harm they might do once they have spread, in estimating the cancer-causing potential of certain chemicals and even in developing a beer with a well-balanced taste. That the topological method is applicable to all these fields is now well established; the only question is how long it will be before the method is exploited to its full potential.

In the past there have been two basic approaches to predicting the properties of unsynthesized molecules. The first approach is to determine the molecule's precise structure (the exact spatial coordinates of all its atoms) and to apply the rules of quantum mechanics to this structure, thereby determining the shapes and energies of its electron clouds, from which many of the molecule's properties can be derived. The procedure is a laborious one and generally consumes several hours of time on a mainframe computer. Moreover, the results obtained by the procedure apply only to the specific molecule under analysis; they give no indication of the results that could be expected for related molecules.

The second and much easier approach is to determine the properties of basic fragments found in many molecules and then combine these fragmental properties in ways that depend on the makeup of the molecule in question. The properties of each fragment are usually determined by analyzing a set of molecules that are similar to the specific molecule that is being studied.

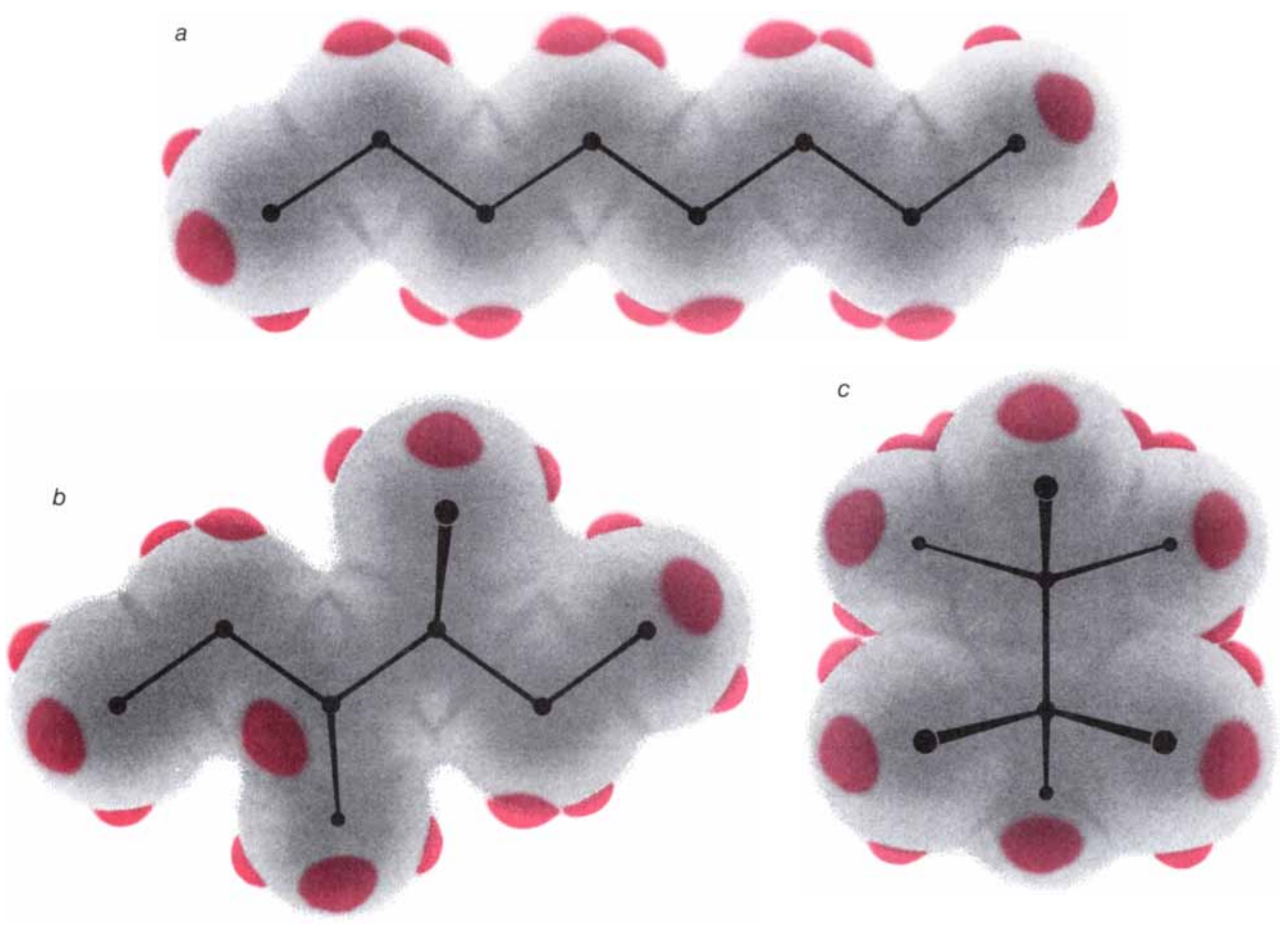
In both approaches the object is to relate a rather ill-defined concept of overall structure to rigorously defined molecular behavior. The new, topological technique is founded on the assumption that molecular structure can be characterized mathematically and precisely, and that the mathematically determined parameters of molecules can be correlated with the molecules' experimentally measured properties. One need not attempt to derive chemical properties from physical structures by first principles; one need only choose a mathematical way of assigning numbers to molecules and then correlate those numbers with the vast data base of known chemical properties. Perfect correlations are impossible: whereas mathematical parameters can be determined precisely, experimental results are always associated with a certain amount of error. Nevertheless, if the experimental data are reasonably reliable, good correlations are possible.

The topological analysis of a molecule begins with a drawing in which the molecule's atoms are depicted as points and the bonds linking them are depicted as straight lines. The length of any line and the angles between the lines do not matter: the existence of a connection between atoms, rather than the precise nature of that connection, is paramount. (Although the loss of information about the relative lengths and orientations of bonds might seem damaging, in practice the information usually turns out to be much less important than might be supposed.)

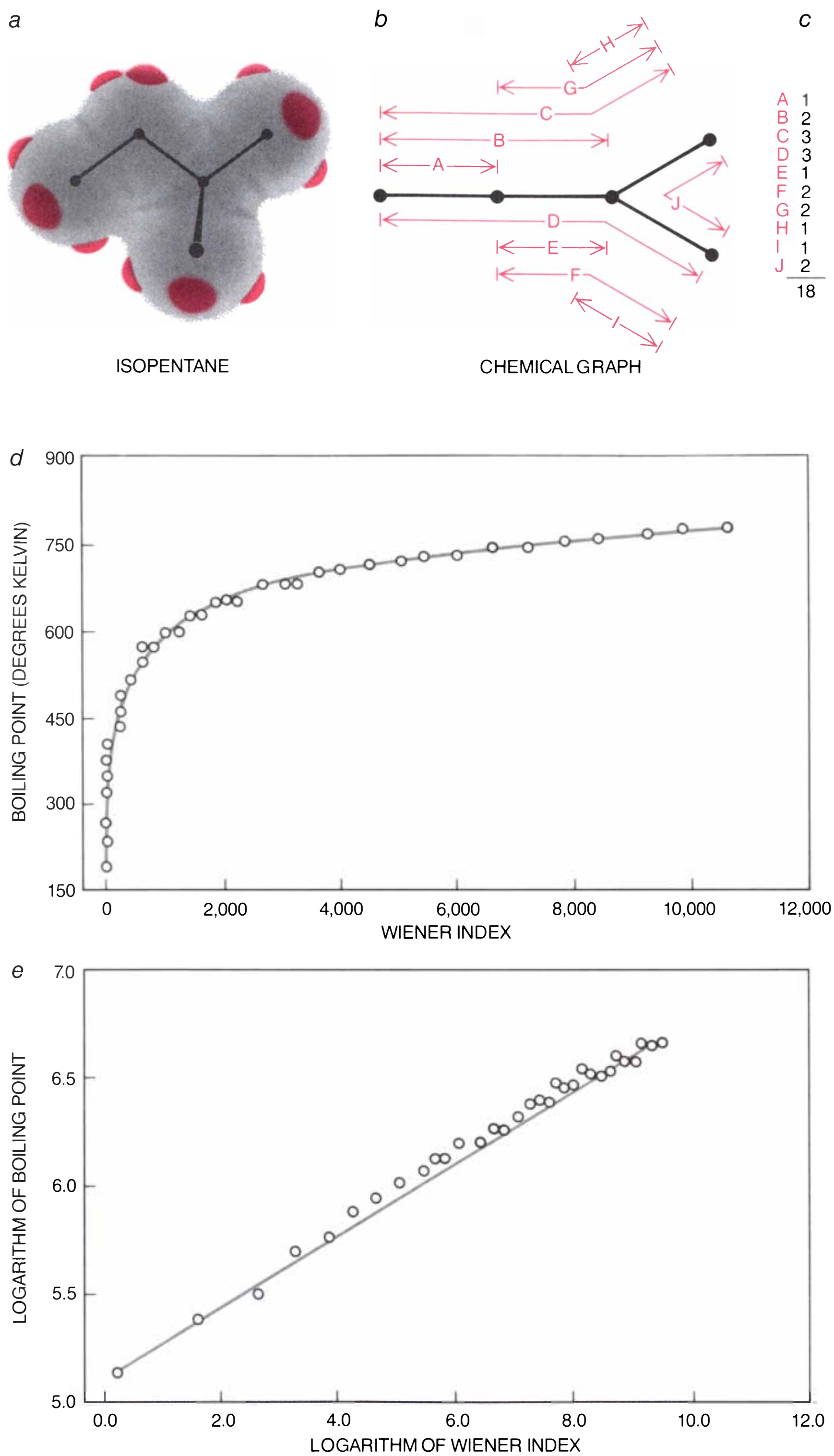
Drawings of this kind are known as chemical graphs. Graphs represent structures in an abstract way. They were first studied more than a century ago by the mathematician Arthur Cayley, and they now form the subject matter of a mathematical discipline known as graph theory. In graph theory points are usually referred to as vertexes and lines are referred to as edges. Chemical graphs are the basic tool used in applying the techniques of abstract mathematical graph theory to the specific problems of chemistry. (In chemical graphs the hydrogen atoms are often omitted, because they normally do not play a major role in determining the structure of a molecule.)

Once the chemical graph of a molecule has been drawn, it is a comparatively simple matter to derive a topological index for the graph and thus obtain a number that characterizes the graph. Obviously the number must have the same value for a given molecule no matter how the molecule's graph is drawn or labeled. A number of this type is referred to by mathematicians as a graph invariant. One of the earliest graph invariants, or topological indices, considered by chemists is known to mathematicians as the vertex number and to chemists as the carbon number. Mathematically it is simply the number of vertexes in the graph; as applied to a hydrocarbon molecule (a molecule that contains only hydrogen and carbon atoms) it is the number of carbon atoms. Chemists have been relying on the carbon number for certain correlations for more than a century without realizing that they were using what is in fact a simple topological index.

The carbon number is an appropriate index for analyzing straight-chain molecules, but it is not well suited to branched molecules, which may have very different shapes from one another even if they have the same number of carbon atoms. Because many different molecules have the same carbon



HYDROCARBON MOLECULES may take on quite different shapes even when they have roughly the same volume. Topological methods for predicting the chemical behavior of molecules emphasize one particular aspect of a molecule's structure: the pattern of interconnections among its atoms. In topological analysis such features as the actual three-dimensional shape of the molecule and the angles between its bonds are ignored; what matters are such considerations as how many other atoms each atom is connected to and whether they are connected in a straight chain, a branched structure or a ring. Even though topological analysis ignores what seem to be important structural considerations, it has proved extremely powerful in predicting a variety of phenomena, from the boiling points of various substances to the likelihood that certain pollutants will spread in the environment. The molecules shown here are *n*-octane (*a*), 3,4-dimethylhexane (*b*) and 2,2,3,3-tetramethylbutane. They consist entirely of hydrogen (*color*) and carbon (*gray*) atoms.



WIENER INDEX, a procedure developed by Harry Wiener of Brooklyn College for converting the topological structure of a molecule into a single descriptive number, serves as the basis for predictions of chemical behavior. First the molecular structure (*a*) is depicted in a "chemical graph" (*b*), in which atoms are shown as points (called vertexes) and bonds are shown as lines (called edges). The molecule graphed here is isopentane. The Wiener index is calculated (*c*) by summing up the number of bonds that would have to be traversed in traveling along the molecule from every atom to every other atom. The Wiener index correlates with many physical properties. For example, when the Wiener indices of a number of known molecules are plotted against the molecules' boiling points (*d*), the points fit a smooth curved line. The line is straighter when the logarithms of boiling point and Wiener index are plotted (*e*). The boiling point of a molecule whose properties are not known can be predicted by calculating the molecule's Wiener index and then finding its place on the line. Similar correlations can be made for many chemical properties and topological indices.

number, the carbon number is said to be an index with low discriminating power. Clearly it is necessary to develop other indices that are able to distinguish effectively between branched and unbranched molecules and among branched molecules having different branching structures.

The first topological index capable of characterizing the "branchedness" of molecules was put forward by Harry Wiener of Brooklyn College in 1947, but it is only in the past few years that chemists have begun to appreciate its great power. The Wiener index is based on the graph-theoretical notion of distance: the distance between any two vertexes is equal to the number of edges one would traverse in taking the shortest route possible through the molecule's graph from one of the vertexes to the other. The Wiener index of a molecule is equal to the sum of the graph-theoretical distances between all pairs of atoms in the molecule [*see illustration at left*].

The Wiener index of a molecule is like the carbon number in that it is generally larger for molecules that have more atoms, but it also provides some measure of a molecule's branching structure. In particular it is larger for extended molecules and smaller for compact ones. After Wiener devised it, several other workers found that for certain types of hydrocarbon molecules the index correlated surprisingly well with such properties as boiling point, viscosity, surface tension and refractive index (which reflects the degree to which a substance bends the path of a light ray).

In the past few years the Wiener index has been shown to have a much broader range of applicability. For example, in 1979 Ovanes Mekenyan, Danail Bonchev and Nenad Trinajstić of the Higher Institute of Chemical Technology in Bulgaria showed that it correlates very well with the energies of the bonding electrons in certain kinds of complex molecules known as spiro-type molecules (molecules having two or more rings that lie in mutually perpendicular planes, with adjacent rings sharing a single vertex) and polycyclic aromatic hydrocarbons (hydrocarbons having two or more fused hexagonal rings that share at least one edge and that all lie in the same plane).

This is a very important result, because the energy levels of a molecule's bonding electrons are directly responsible for a great deal of the molecule's chemical behavior. A correlation between a molecule's Wiener index and the energy of its bonding electrons makes it possible to predict such properties as the molecule's electrical conductivity, its electron affinity (its tendency to attract other electrons and hence to bond to molecules that have an abundance of electrons) and the wavelengths of light it will absorb. By extension the Wiener index can then be used to predict the relative stabilities of unknown compounds made from molecules that have not yet been synthesized.

The utility of the Wiener index is not limited to small molecules; it has been augmented in ways that make it applicable to such potentially infinite systems as polymers (single molecules consisting of many identical pseudomolecular units bonded together in long chains) and crystals (three-dimensional solids built up of identical unit cells). The original formulation of the Wiener index, when applied to an infinite polymer chain, would give an infinite value, because there would be an infinite number of paths connecting atoms along the chain. In 1980 Bonchev and Mekenyan proposed a method of modifying the Wiener index so that it would yield finite values even for infinite systems, as long as the systems are composed of many identical finite units.

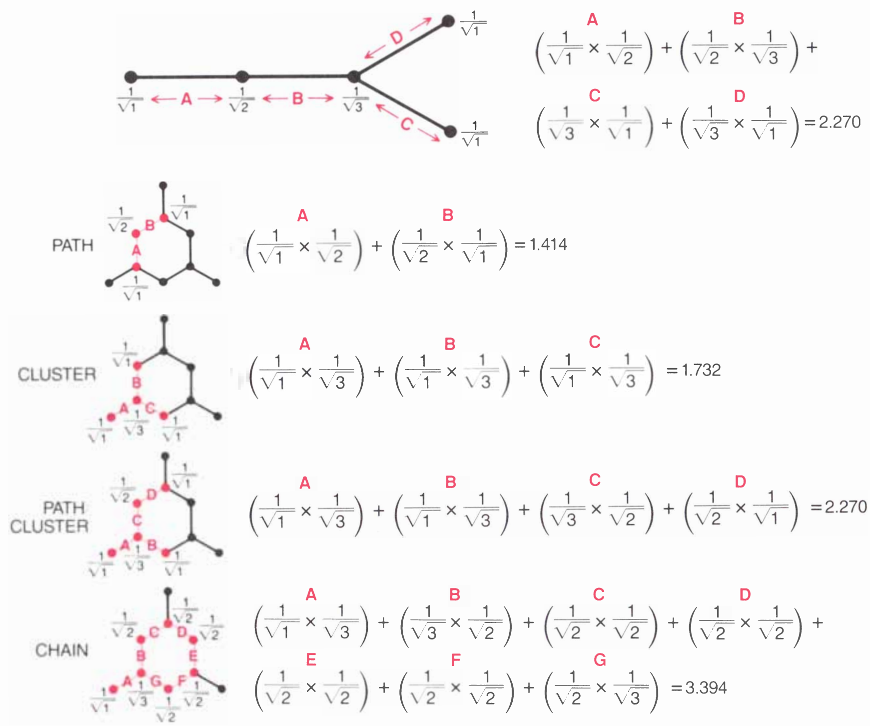
The approach proposed by Bonchev and Mekenyan produced good estimates of certain physical and electrical properties of polymers called polyenes, which have the potential to act as electrical conductors or semiconductors. It also led to predictions of the melting points, boiling points and other physicochemical properties of many other polymers, including polytetrafluoroethylene, polycapramide and polyethyleneterephthalate.

Extensions of the Wiener index also make it possible to model various processes in which atoms of one element fill interstitial areas in a crystal composed of another element. An important question to be answered here is which of the interstitial sites will be filled by foreign atoms. This type of question is applicable to problems in corrosion control, in catalysis and in chemisorption (the binding of one substance on the surface of another). The fundamental idea behind the solution is this: in general a system is in its minimum-energy state when its Wiener index is at a minimum. The procedure, then, is to calculate the three-dimensional Wiener index of each possible configuration of crystal-lattice atoms and foreign atoms. The configurations with the lowest Wiener indices are the ones most likely to be realized.

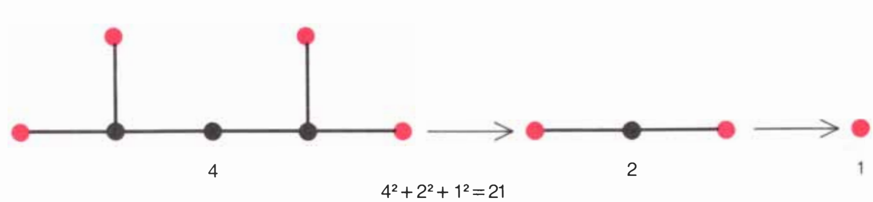
The Wiener index of a molecule offers primarily a measure of the molecule's volume, although it does give some indication of the molecule's shape. An index that is more sensitive to shape was introduced in 1975 by Milan Randić, now at Drake University, and is known today as the molecular-connectivity index. Judging from the number and diversity of its applications, the Randić molecular-connectivity index is by far the most valuable index yet devised.

Just as the Wiener index depends on the topological concept of distance, so the Randić index depends on the topological concept of degree. The degree of any vertex is equal to the number of other vertexes to which it is attached. To calculate the Randić index one assigns to each edge a "value," which depends on the degrees of the vertexes the edge connects. More specifically, the value of an edge is equal to the reciprocal of the square root of the product of the degrees of the two vertexes. The Randić index of a molecule is equal to the sum of the values of all the molecule's edges.

After Randić proposed his molecular-connectivity index, other investigators noted that in some cases better correlations could be obtained by focusing on the component parts of a molecule. For example, a certain



RANDIĆ MOLECULAR-CONNECTIVITY INDEX, named for Milan Randić, now at Drake University, puts more emphasis on structure and less on absolute size than the Wiener index does. Each vertex has a "degree": the number of other vertexes it is linked to. Similarly, each edge has a "value": the product of the reciprocals of the square roots of the degrees of the vertexes it joins. The Randić index of a molecule is equal to the sum of the values of all the molecule's edges. In some cases better correlations with chemical behavior are found when fragments of a molecule are considered independently of the molecule as a whole. Methods for deriving the connectivity indices of four kinds of fragments are shown at the bottom. Connectivity indices have a wider range of applications than any others yet devised. For example, they have potential roles in developing new drugs, in modeling the toxicity and spread of pollutants and in predicting the taste and smell of new substances.



BALABAN CENTRIC INDEX, named for Alexandru T. Balaban of the Polytechnic Institute in Bucharest, emphasizes the degree of branching in a molecule. All vertexes that are linked to just one other vertex are counted and "pruned" from the molecule's graph; the number of vertexes pruned at each step is squared and added to a running total. The process is repeated until every vertex has been counted. The final total is the molecule's index.

property might correlate best with the sum of the Randić indices of all the clusters (groupings of three atoms about a single central atom) in a molecule rather than with a Randić index based on single edges in the graph of the molecule. Connectivity indices have therefore been developed that are applicable to particular molecular substructures that may be of importance in determining specific behavioral characteristics. The Randić molecular-connectivity index is therefore really a set of indices, each applicable in different cases.

Molecular-connectivity indices correlate well with a great variety of physical properties, such as density, solubility in water and heat of vaporization. More important, molecular-connectivity indices have been found to correlate with an even greater variety of biological properties as well. It has been known for some time that many biological responses are triggered when an appropriate stimulating molecule docks with a receptor on the surface of a cell. In many cases the specific shape of the triggering molecule is not as important as its volume or surface area. Molecular-connectivity indices correlate well with surface area and volume, and so they also correlate well with the ability to induce particular biological responses.

Among these correlated properties of molecules are their tendencies to act as anesthetics, narcotics and hallucinogens. It has also been possible to correlate the odor of certain molecules with their connectivity indices; an index can indicate whether the molecule will smell like an etherous chemical, a flower or a carboxylic acid (a main component of human sweat). It is also possible to predict whether a molecule will taste bitter or sweet by determining whether its connectivity index lies above or below a certain threshold value. Ronald Gardner, in association with Harp Lager, Ltd., has exploited this taste-discriminating ability of connectivity indices in the development of new brews of beer.

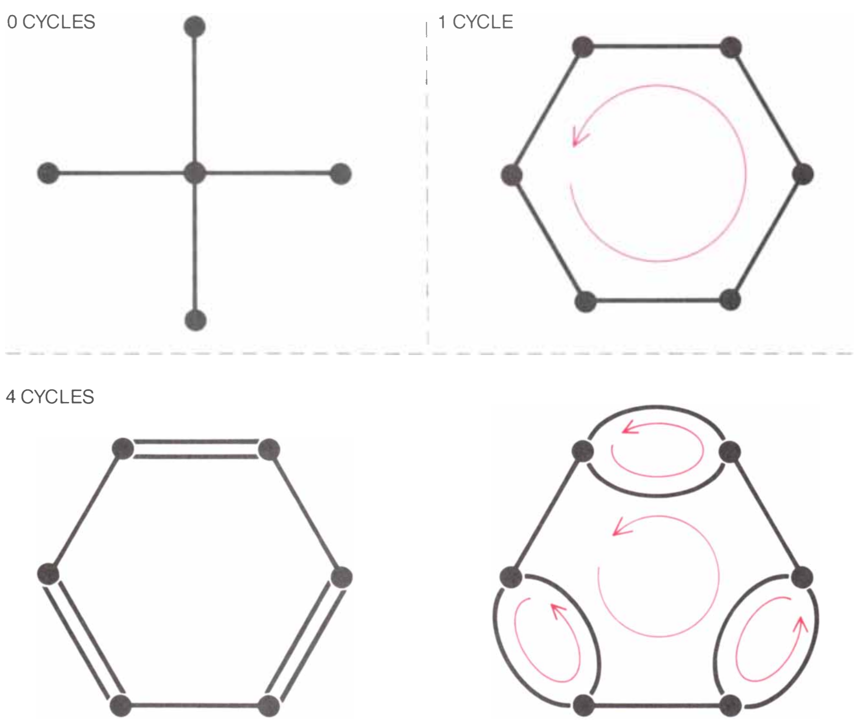
There are manifold other biological applications for connectivity indices. They have been found to correlate well with the power of certain chemicals to inhibit the reproduction of microorganisms such as *Staphylococcus aureus, S. typhosa* and *Mycobacterium tuberculosis.* The ability of certain substances to kill the Lee strain of *B* influenza virus also correlates well with the indices. Connectivity indices can also serve as effective measures of the mutagenicity (the tendency to cause genetic mutations) of chemicals known as nitrosamines, which are found in cigarette smoke, nitrate-pickled meat and smoked fish.

Many of the most exciting applications of molecular-connectivity indices are in environmental studies. The indices have been shown to predict the toxicity of many compounds toward minnows and other freshwater organisms. More important, the indices correlate well with the ability of many pollutants to spread within the air, water or soil, to concentrate within living organisms or to pass from one of these environmental compartments to another. It is often extremely difficult and expensive to test such properties physically, but to do the predicting with indices one usually needs little more equipment than a pocket calculator. The U.S. Environmental Protection Agency has already begun applying the indices to such problems as predicting the toxic potential of unknown or untested pollutants.

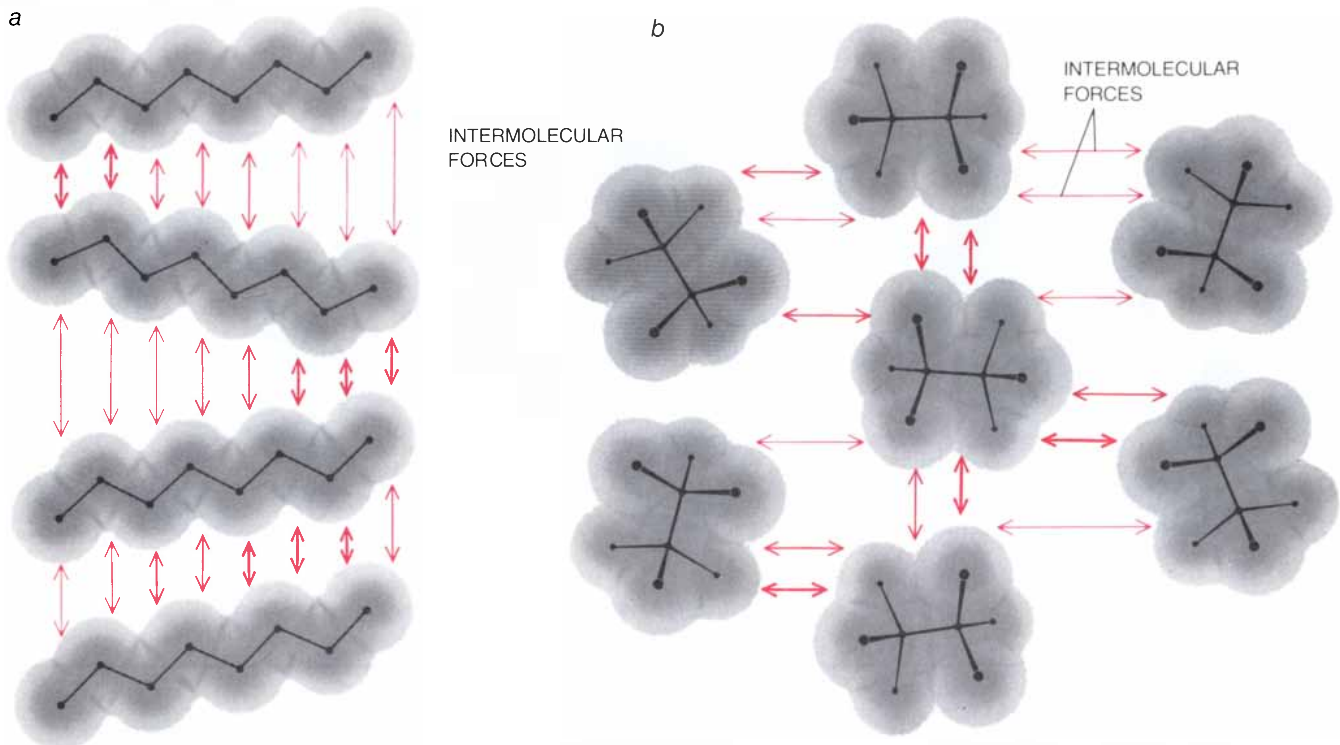
Although the Wiener and Randić indices are applicable to a broad range of problems, there are some situations in which they do not lead to very good correlations. In these cases it is often possible to develop a specialized index, one that is applicable primarily to the problem under immediate consideration. Such specialized indices are called for whenever factors other than the overall size and shape of a molecule play a dominant role in determining its chemical behavior.

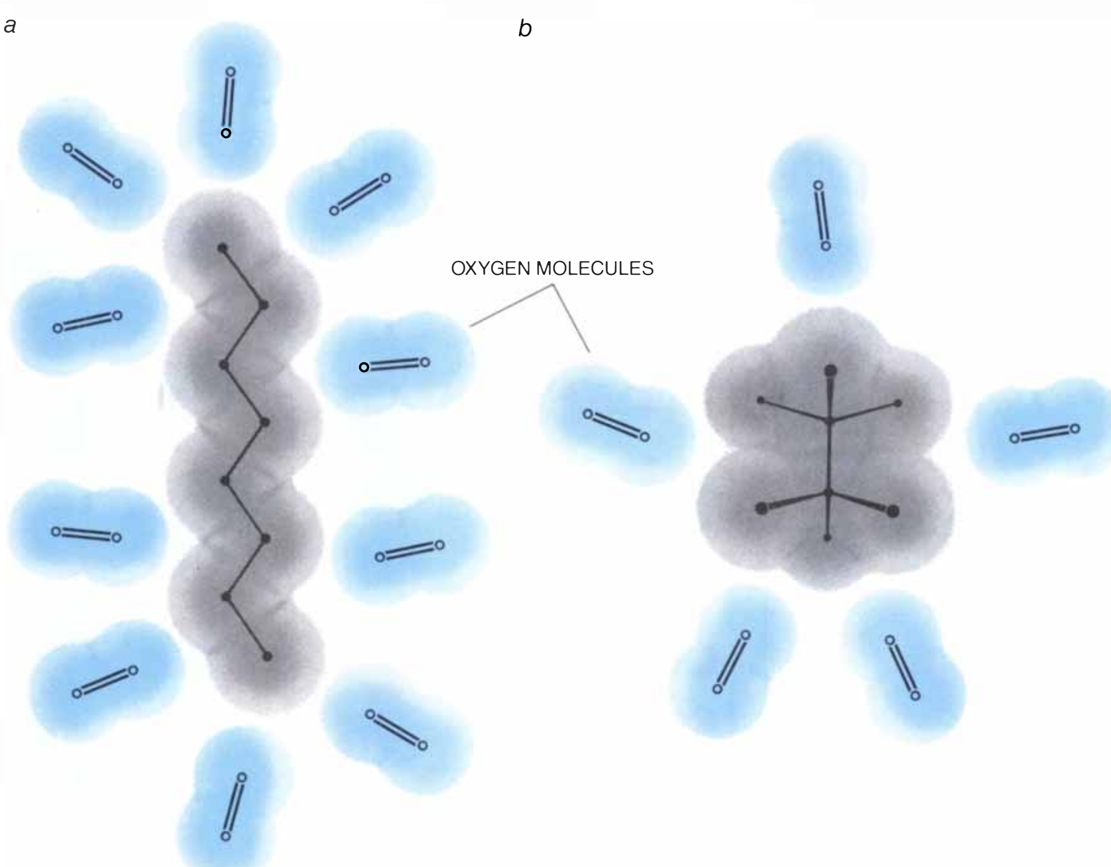
One application for which a specialized index is required is in predicting the octane number of a fuel. The octane number, which is usually determined under standardized conditions in a test engine, provides a measure of the efficiency with which a fuel burns: its tendency not to "knock." Knock is caused when oxygen atoms combine with fuel as it is being compressed and before it has been ignited. Straight-chain molecules are generally more susceptible to knock than branched molecules because the atoms in a straight molecule are all vulnerable to impinging oxygen molecules, whereas in a branched molecule many of the atoms are shielded in crevices formed by the branching pattern.

Because the antiknock tendency of molecules depends heavily on the extent to which they are branched, many attempts have been made to correlate octane numbers with standard indices, such as the Wiener index, that take some account of shape as well as volume. Fair correlations were obtained in this way, but many investigators thought that an index that explicitly emphasized branching might correlate better. In 1979 Alexandru T. Balaban of the Polytechnic Institute in Bucharest introduced an index he called the centric index, which correlated ex-

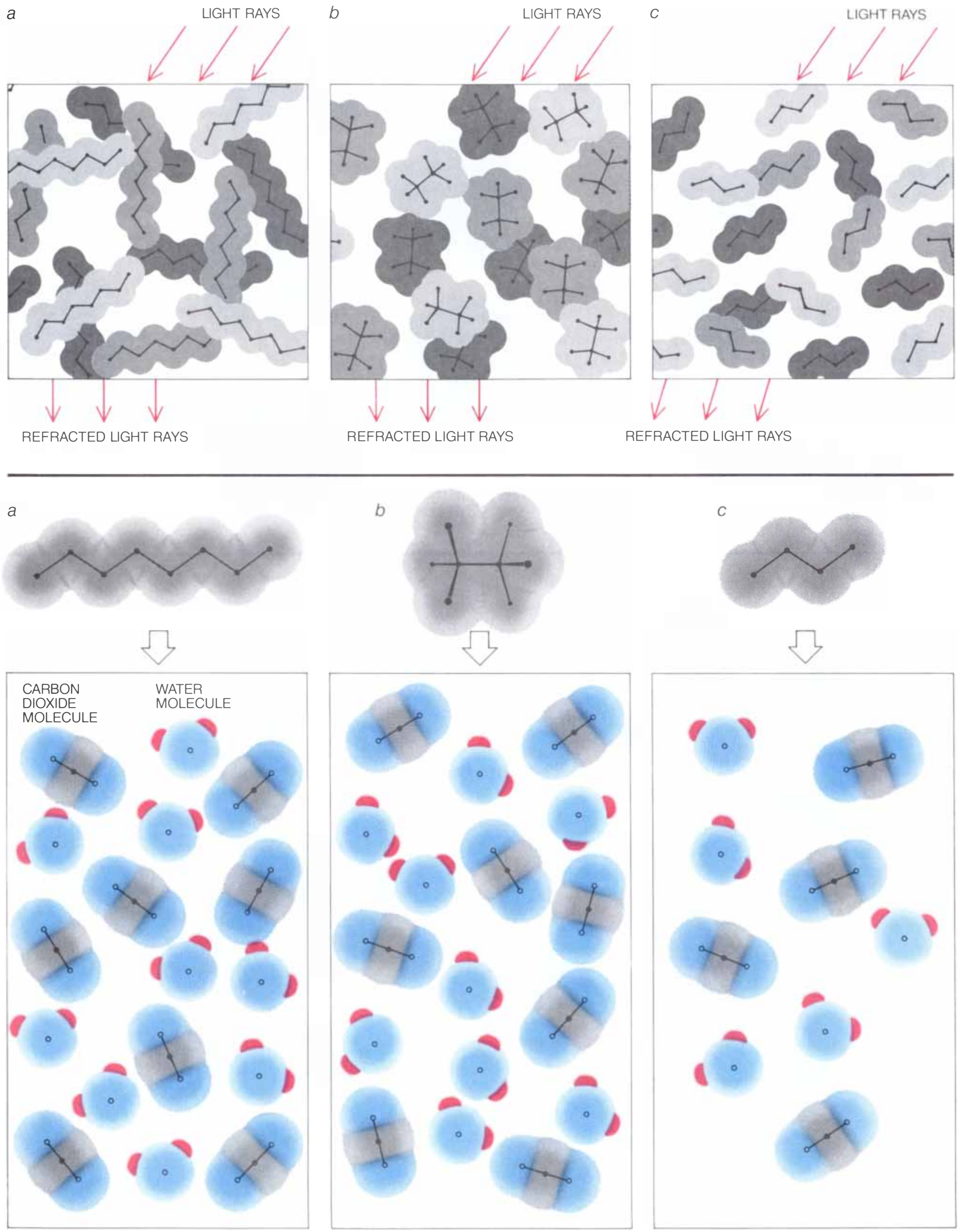


HYDROGEN-DEFICIENCY NUMBER is equal to the number of independent cycles, or rings and double bonds, in a molecule. The hydrogen-deficiency number, when multiplied by an index that measures the relative prevalence of open chains and closed rings, yields an index that correlates with the amount of soot produced by burning hydrocarbon molecules.





PROPERTIES DEPENDING ON SHAPE of a molecule include boiling point and octane number. Boiling point (*top*) depends on shape because it is determined by intermolecular forces: the stronger the attractive forces among molecules of a liquid are, the higher the boiling point will be. Straight-chain molecules (*a*) interact with one another more strongly than branched molecules (*b*) do, because straight molecules have more surface area open to interactions. The octane number of a fuel (*bottom*) is a measure of its tendency not to "knock," or burn (combine with oxygen) while being compressed and before being ignited. Along a straight molecule (*a*) there are many sites where an impinging oxygen molecule could attack. On branched molecules (*b*) some sites are hidden within crevices formed by branching. Branched molecules have better antiknock properties. Indices that are sensitive to shape correlate with such properties.



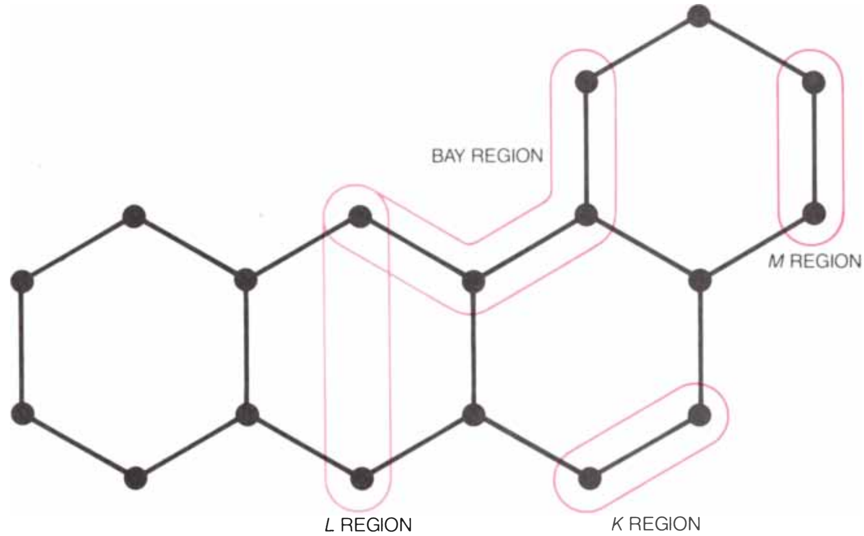
PROPERTIES DEPENDING ON SIZE of a molecule include molar refractivity and heat of combustion. The molar refractivity of a substance is the degree to which the substance bends light rays (*top*). Molar refractivity depends on the number of electrons in each molecule that can interact with incoming light rays. Larger molecules (*a*, *b*) have more electrons and bend light more than smaller molecules do (*c*). The heat of combustion of a hydrocarbon is the amount of heat energy given off when the molecule combines with oxygen to form water and carbon dioxide (*bottom*). A certain amount of heat is given off for every water or carbon dioxide molecule formed, and so larger hydrocarbon molecules, which form more product molecules when they are burned (*a*, *b*), have greater heats of combustion than smaller hydrocarbon molecules do (*c*). Indices sensitive to the size of a molecule correlate with these properties.

tremely well with the octane number of fuel hydrocarbon molecules.

The Balaban centric index is derived in an iterative "pruning" process that is carried out on the graph of the molecule being studied [*see bottom illustration on page 43*]. In the first step of the process all the vertexes in the graph that are connected to just one other vertex are counted, and the graph is then redrawn without them. The number of vertexes that were pruned, or eliminated, is squared, and the square is added to a running total. The process is repeated until the graph has been reduced to nothing by repeated pruning. The final total (that is, the sum of the squares of the number of vertexes pruned at each step) is the Balaban centric index for that molecule.

Another specialized index was designed to predict the amount of soot produced when a given hydrocarbon molecule is burned. The correlation of molecular structure with the production of soot has proved to be a particularly difficult task. Simple indices, such as the carbon number, yield good correlations for molecules within the same family, that is, for those having similar structures and bonding behavior, but they are unreliable in comparing the soot production of molecules from different families.

Earlier this year Milton P. Hanson and I were trying to devise an index that would correlate soot production from family to family as well as within any single family. It soon became evident that this task required an index able to distinguish among molecules that do or do not have atoms held together by double bonds or rings of atoms. We eventually found a solution by combining two relatively simple indices. One index, known as the hydrogen-deficiency number, measures the number of independent cycles, or rings, in the molecule's graph [*see illustration on page 44*]. The other, called the averaged-distance-sum connectivity index, measures the relative prevalence of open chains and closed rings within the molecule. Our index is the product of these two indices, and it yields a single, straight-line correlation for the soot production of nearly 100 hydrocarbon molecules.



CARCINOGENICITY (cancer-causing tendency) of a molecule can be predicted topologically in some cases. In hydrocarbons consisting of several fused hydrogen rings lying in the same plane, certain regions, known as bay, *K*, *L* and *M* regions, are important for the chemical reactions occurring in the series of steps leading to carcinogenesis. An equation that includes many indices, formulated by William C. Herndon and László von Szentpály of the University of Texas at El Paso, takes account of the prevalence of structural features that can be defined in terms of these regions. The correlation obtained also depends on the size of the molecule studied; molecules above or below certain sizes are not carcinogenic.

One of the most ambitious efforts to exploit topological prediction is the recent attempt to correlate the carcinogenic behavior of various molecules with indices. Because the causation of cancer involves a multistep process, an appropriate index would have to be sufficiently subtle to reflect not only the chemical interactions of the original molecule but also those of molecules formed as products of the initial reactions. Although the Randić molecular-connectivity index can be used to predict whether a particular molecule will be carcinogenic or not, in predicting the degree of carcinogenicity of a molecule the best results so far have been achieved by William C. Herndon and László von Szentpály of the University of Texas at El Paso. They devised an ingenious combination of simple indices to model the carcinogenicity of polycyclic aromatic hydrocarbons.

One important consideration in predicting the carcinogenicity of these molecules has been that molecules falling above or below certain size limits tend not to be carcinogenic. In their model Herndon and Szentpály have allowed for these limitations by including, as indices in the correlation, the carbon number and the square of the carbon number. Another consideration is that certain regions of a polycyclic aromatic hydrocarbon are more important in carcinogenesis than other regions. These regions, known as bay, *K*, *L* and *M* regions, are important in at least the initial reactions in the chain of carcinogenesis. The bay, *K* and *M* regions must be fairly active chemically, and the *L* region relatively inactive, if the molecule is to be carcinogenic. Herndon and Szentpály's model takes the form of an equation containing four topological indices, two of which relate to the size of the molecule and two to the occurrence of certain structural features (which are related to bay, *K*, *L* and *M* regions) in the molecule. Their equation has yielded excellent correlation with experimental measurements of carcinogenicity, even though the experimental measurements themselves have been subject to a good deal of uncertainty.

Topological indices can model an incredibly wide range of physical, chemical and biological phenomena. They can model the behavior of gases, liquids and solids and of both inorganic and organic species. Although it has not yet been possible to tell in advance with complete certainty which index will be most suited to any particular application, various indices are already becoming known for their applicability in modeling size, shape, branching, reactivity and a large number of other characteristics. Two areas in particular where indices appear to have a bright future are in designing new drugs and in tracking the fate of various pollutants in the environment and modeling their likely effects on living organisms.

It is only in recent years that the great power of topological indices has been widely recognized; in many fields they are on the threshold of general acceptance. We now know that indices work; it remains only to put them to work. When that is done, a new kind of chemical paradigm will become established. It will then be possible to make valid and useful chemical predictions routinely by referring to one of the simplest, most fundamental and yet often most neglected molecular parameters available. Molecular topology will have arrived.

Superstrings

If all elementary particles are treated as strings, a consistent quantum theory emerges that accounts for all four fundamental forces. The theory may transform our ideas about space and time

by Michael B. Green

The central paradox of the contemporary physics of elementary particles is the apparent incompatibility of its two main theoretical foundations. The first foundation is Einstein's general theory of relativity, which relates the force of gravity to the structure of space and time. This view of gravity has led to models of phenomena on a cosmic scale and to an understanding of the evolution of the universe. The second theoretical foundation is quantum mechanics, which can account for the atomic and subatomic world. Quantum theories have been formulated for three of the four known forces of nature: the strong, weak and electromagnetic interactions. Until recently there seemed to be little hope that Einstein's theory of gravity—the fourth fundamental force—could be united with the precepts of quantum mechanics. The basic difficulty is that such a unification seems to call for a radically new formulation of the laws of physics at the smallest distance scales; in such a reformulation the idea that space and time are continuous sets of points would have to be abandoned. Without a quantum theory of gravity and the conceptual revisions such a theory implies, a comprehensive description of all the forces of nature could not be realized.

In the past two years elementary-particle physicists have become optimistic that the theoretical impasse might be resolved. The optimism is based on striking developments in a new kind of theory known as superstring theory. In superstring theory, as in any other string theory, elementary particles can be thought of as strings. String theories thereby differ from all familiar quantum-mechanical field theories, such as the quantum theory of electromagnetism, whose quanta, or constituent particles, are pointlike. Since a string has extension, it can vibrate much like an ordinary violin string. The harmonic, or normal, modes of vibration are determined by the tension of the string. In quantum mechanics waves and particles are dual aspects of the same phenomenon, and so each vibrational mode of a string corresponds to a particle. The vibrational frequency of the mode determines the energy of the particle and hence its mass. The familiar elementary particles are understood as different modes of a single string.

Superstring theory combines string theory with a mathematical structure called supersymmetry [see "Is Nature Supersymmetric?" by Howard E. Haber and Gordon L. Kane; SCIENTIFIC AMERICAN, June]. Not only does superstring theory avoid the problems previously encountered in combining gravity with quantum mechanics, but also, in the process, the theory makes it possible to consider all four fundamental forces as various aspects of a single underlying principle. Furthermore, the unification of the forces is accomplished in a way determined almost uniquely by the logical requirement that the theory be internally consistent. These developments have led to an extraordinary revitalization of the interplay between mathematics and physics. Many of the deepest discoveries in modern mathematics are contributing to the understanding of the theory; in return string theories raise new issues in mathematics.

According to superstring theories, the standard laws of physics are approximate versions of a much richer theory that takes account of structure at an inconceivably small distance scale. The strings postulated by the theory are about 10^{-35} meter long, or some 10^{20} times smaller than the diameter of the proton. The differences between superstring theories and more conventional theories at such minute scales are essential to the consistency and predictive power of the theory.

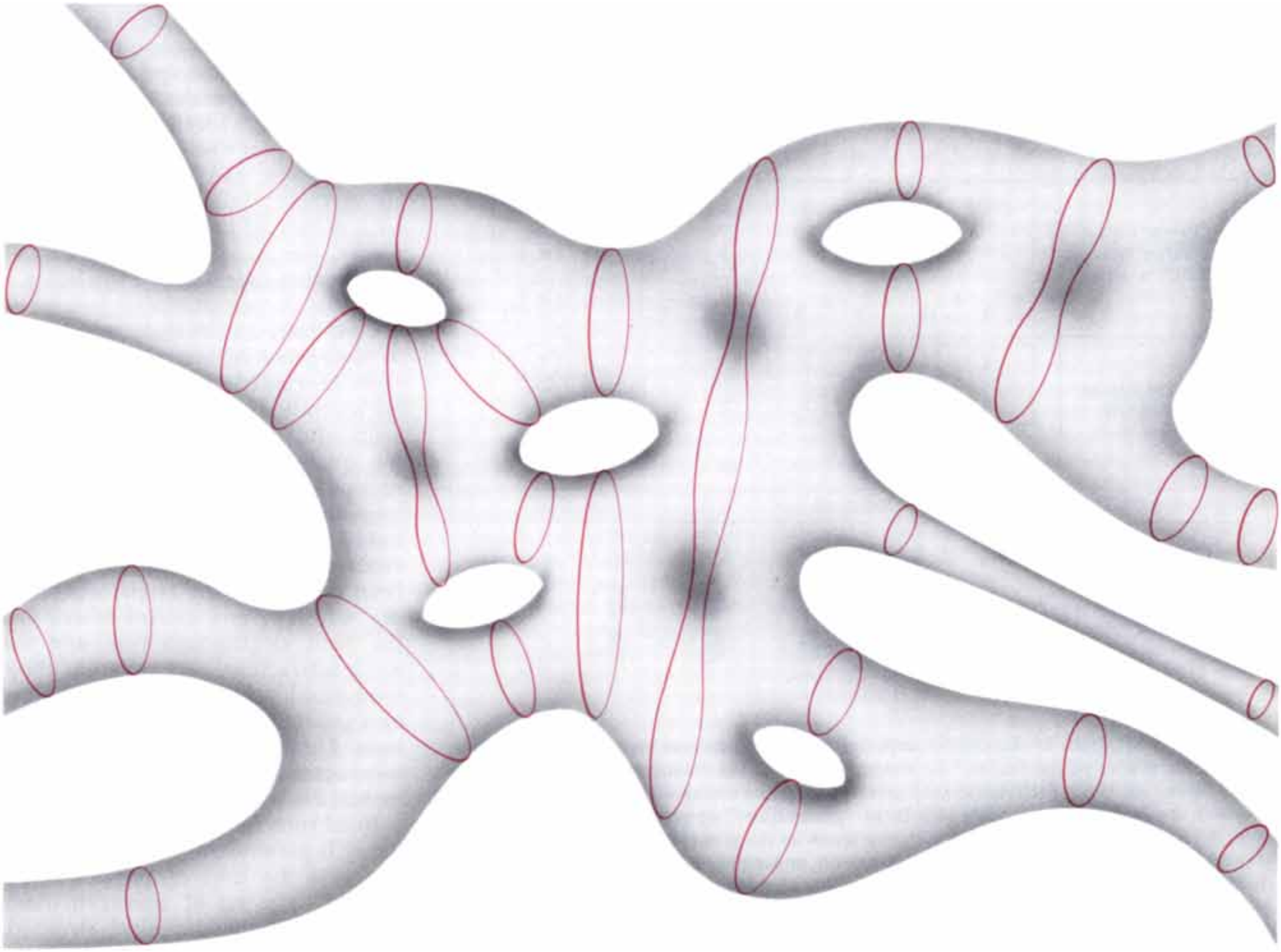
For example, if one disregards gravity, it is possible to construct a unified picture of the strong, weak and electromagnetic forces in an ordinary field theory having pointlike quanta. The unified picture is the outcome of some underlying symmetry built into the theory, but there are many possible underlying symmetries. No theoretical reason is known for preferring one such symmetry to another. In contrast, in superstring theories gravity cannot be excluded, and the kind of symmetry needed for its inclusion in the theory leads to a natural prediction about the underlying symmetry that unifies the other three forces.

Since new concepts of space and time have long been expected from a quantum theory of gravity, it is worth mentioning how superstring theory could change our ideas about the geometry of the universe. It is not correct, strictly speaking, to regard the strings as independent particles moving in some fixed background space. In Einstein's theory of gravity, which superstring theory must approximate, space and time are unified in a four-dimensional continuum called spacetime. The influence of the gravitational force is determined by the so-called curvature of spacetime, which is analogous to the curvature of a two-dimensional surface such as the surface of a sphere. A particle moves along a geodesic, or shortest path, in the curved spacetime; on the sphere the analogue to such a path is the great-circle route between two points. The particle exerts a reciprocal influence on spacetime, causing gravitational waves that can disturb the very geodesics along which the particle is moving. The equations of general relativity determine not only the paths of particles but also the structure of the spacetime in which they are moving.

In superstring theory gravity is defined in a world expanded to nine spatial dimensions and time, making 10 dimensions in all. Again motion proceeds along geodesics, but the geodesics are surfaces of minimum area in 10 dimensions. Evidently six of the 10 dimensions must be hidden from view, thereby leaving only the four familiar dimensions of spacetime to be observed. The six extra dimensions must be curled up to form a structure so small that it cannot directly be seen. The idea of unobservably small dimensions can readily be understood by considering a simple, two-dimensional analogy. A hose is a two-dimensional surface that appears to be one-dimensional when it is observed at scales too coarse to resolve its thickness. In superstring theory it is likely that the size of the six curled-up dimensions is approximately the same as the length of the string. The world appears to have three spatial dimensions in the same sense that the string acts like a point particle.

The expansion of the idea of geometry is not limited to adding six spatial dimensions. In ordinary general relativity a gravitational field is defined at every point in spacetime. The equivalence of waves and particles in quantum mechanics requires that a gravitational wave, or disturbance in a gravitational field, be identified with a particle; the particle is called the graviton. Similarly, in string theory there should be a field that depends on the configurations of a string; such a field is called a string field. The number of possible configurations of a string in space is vastly greater than the number of points in the space; a string field would therefore be related to a new kind of geometry in an enormous extension of the idea of space, defined by all the possible configurations of a string. A stringlike particle should then be thought of as a "wavelike" disturbance in this huge space, just as a graviton is a wave in ordinary space.

Superstring theory can be historically traced to a theory called the dual-resonance model, which was developed in the late 1960's to explain the observed features of hadrons, or particles subject to the strong force [see "Dual-Resonance Models of Elementary Particles," by John H. Schwarz; SCIENTIFIC AMERICAN, February, 1975]. Although at the time an enormously successful quantum field theory of electromagnetism had been constructed, many theorists were becoming disillusioned with the general approach taken by quantum field theory. No such theory seemed able to ac-



WORLD SHEET is swept out by stringlike particles as they move and interact in spacetime. A stringlike particle, like an ordinary string, has extension in only one dimension. In superstring theory the strings are exceedingly small and move through a 10-dimensional analogue of space and time. In the diagram time is the horizontal axis; closed strings, or loops (*color*), enter from the left and leave on the right. The closed strings sweep out world sheets that are deformed cylinders, that is, they are topologically equivalent to a cylinder. The corresponding diagrams for processes involving open strings are somewhat more complicated because the surfaces have boundaries traced out by the endpoints of the string. When two strings collide, they join to form a third string: two cylinders form a third cylinder. When strings split apart and rejoin, a hole is left in the world sheet. In quantum calculations all possible splittings and joinings between an initial state of strings and a final state must be considered. The topological structure of a world sheet describing these quantum-mechanical interactions is like that of a doughnut with an arbitrary number of holes. Properties of superstring theory for such world sheets are under intensive investigation and are closely related to several important topics in modern mathematics.

count for the behavior of strongly interacting hadrons having a large spin, or quantized angular momentum.

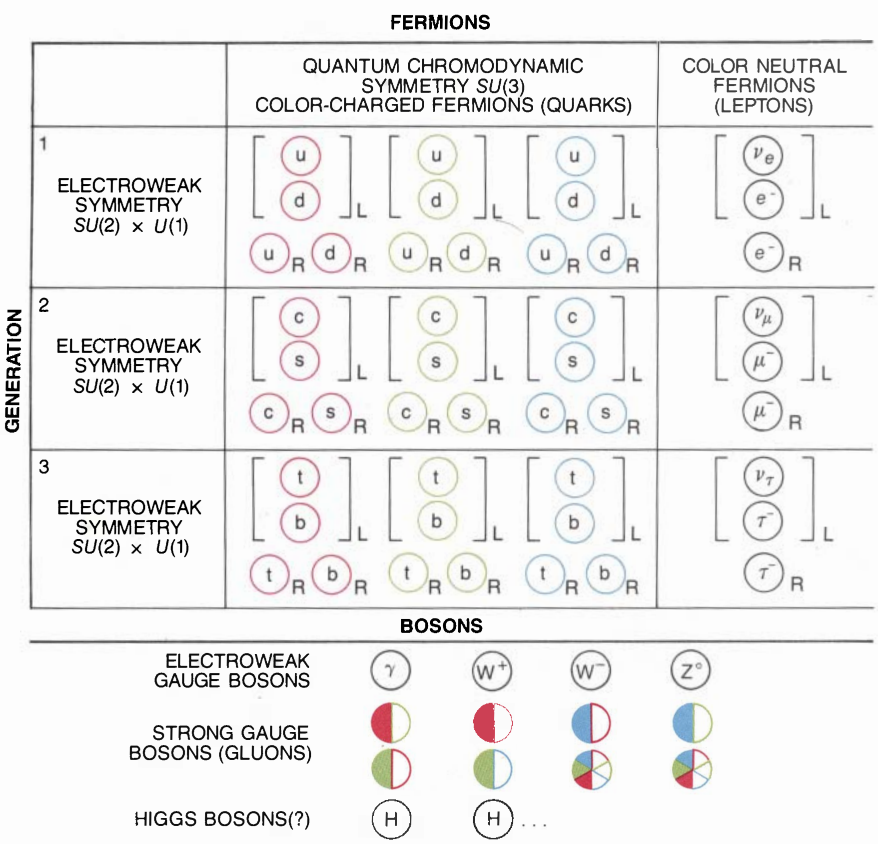
It was in this context that Gabriele Veneziano, now at CERN, the European laboratory for particle physics, simply guessed a formula, unrelated to the formulas of quantum field theory, that expressed many features of hadron interactions. Subsequently Yoichiro Nambu of the University of Chicago, the late T. Goto, Holger B. Nielsen of the Niels Bohr Institute in Copenhagen and Leonard Susskind of Stanford University showed that applying Veneziano's formula is equivalent to describing the hadrons as strings. The harmonics of the string vibrations were supposed to correspond to the observed hadrons. Roughly speaking, the strings served to bind together the quarks that make up the proton, the neutron and other hadrons.

The original dual-resonance model could account only for particles, such as the pi meson, whose spin is an integer in fundamental units. Such particles are called bosons, and in quantum mechanics they are sharply distinguished from the fermions, such as the electron and the proton, whose spin is one-half times an odd integer. In 1971 a variant of the original theory that included fermions was developed by Pierre M. Ramond of the University of Florida, André Neveu of the École Normale Supérieure in Paris and John H. Schwarz of the California Institute of Technology. The variant, known as the spinning-string theory, was the precursor of supersymmetric theories.

Unfortunately both early string theories turned out to have several features that were considered serious liabilities at the time. First, the quantum-mechanical behavior of the original string theory for bosons makes sense only if spacetime has 26 dimensions! For the spinning-string theory and indeed for the current superstring theories the corresponding number of dimensions is 10. Furthermore, the theories were plagued by the fact that the states of lowest energy of the string must be tachyons, or particles that travel faster than light. A relativistic quantum theory with tachyons is inconsistent. Finally, the theories required the existence of massless spin-1 and spin-2 particles, which did not correspond to observed hadrons. Their properties were more like those of the photon, the graviton and the so-called weak gauge bosons that carry the weak force. Although Joël Scherk, an exceptional physicist who died at a tragically young age, and Schwarz suggested string theory might be reinterpreted as a theory of gravity and the other forces, the inconsistencies lurking in such a theory seemed overwhelming.

In the early 1970's there was a great resurgence of interest in quantum field theories based on point particles. As recently as 20 years ago the only successful quantum field theory was the quantum theory of electromagnetism I mentioned above, known as quantum electrodynamics, or QED. Yet only a few years later two more highly successful quantum field theories had been developed: the electroweak theory, which gives a unified description for both electromagnetism and the weak interaction, and quantum chromodynamics, or QCD, which describes how quarks bind together to form hadrons. These theoretical achievements were confirmed by remarkable experimental discoveries.

In all such theories the role of symmetry is paramount. The idea of symmetry in the laws of physics is expressed by a set of transformations that form a mathematical structure called a group. For example, the physical laws governing the behavior of an apparatus do not depend on its orientation in space; the laws are said to be symmetric under rotations about any of the three independent spatial axes. All such rotations belong to the three-dimensional rotation group designated $O(3)$; since rotations are specified by continuous angles, the group is a con-



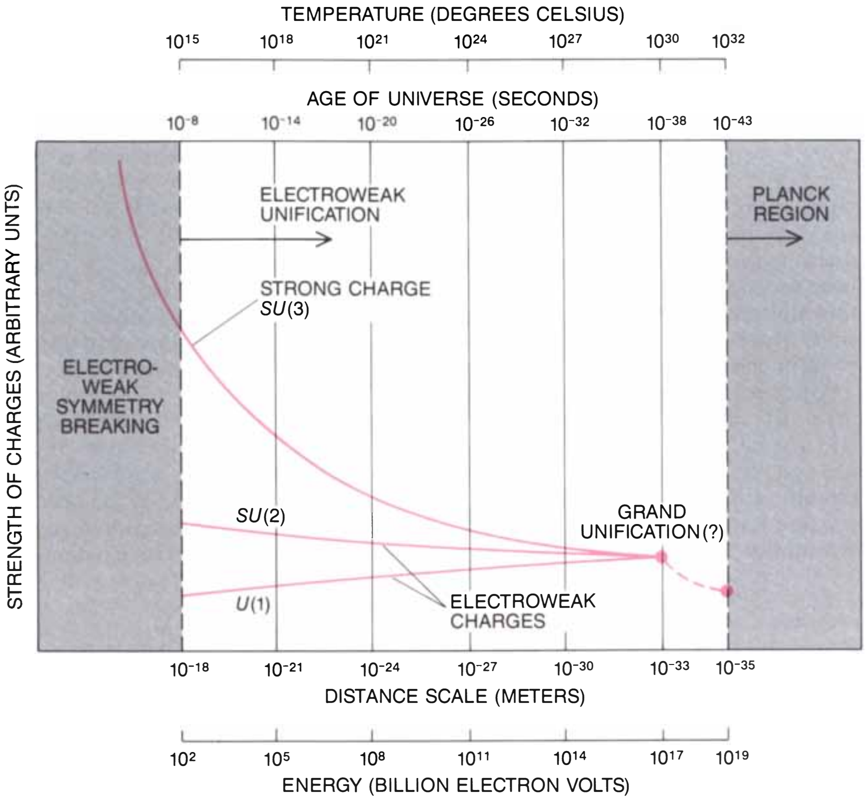
STANDARD MODEL of elementary particles combines quantum chromodynamics (QCD), which is the theory of the "color," or strong, force, and the electroweak theory. In superstring theory all these particles should arise as massless states of a string vibration. The fermions are grouped in the table into three "generations" of particles. The grouping displays the symmetry of the underlying theory much as the grouping of chemical elements in Mendeleev's periodic table displays the symmetry underlying atomic physics. Fermions include the quarks, which carry one of three color charges associated with the symmetry group $SU(3)$ of QCD, and the leptons, which carry no color charge. The color charges are represented here as red, green and blue. Quarks also carry electroweak "flavor," which is associated with the symmetry group $SU(2) \times U(1)$; six flavors are known and are indicated by the letters u, d, s, c, b and t. Leptons are subject only to the electroweak force. The leptons include the electron neutrino (ν_e), the electron (e^-), the muon neutrino (ν_μ), the muon (μ^-), the tau neutrino (ν_τ) and the tau (τ^-). Gauge bosons transmit the forces. The electroweak gauge bosons include the photon (γ) and the three massive particles W^+, W^- and Z^0. Eight strong gauge bosons, or gluons, carry color charges and anticolor charges, shown as solid color or as a color outline. There may also be Higgs bosons, which are responsible for the nonzero masses of particles. The subscripts L and R (for left and right) indicate handedness. For every particle except the neutrinos there is an antiparticle of the opposite handedness. The left-right asymmetry, most notable in the absence of right-handed neutrinos, is a signal that the weak interaction distinguishes among particles on the basis of handedness.

tinuous symmetry group. It turns out that for every continuous symmetry in physics there is a conserved quantity or charge. The continuous symmetry of rotations in space gives rise to the conservation of angular momentum.

Many symmetries in particle physics are not related to ordinary space; instead they can be thought of as symmetries related to some so-called internal space. For example, in Maxwell's electromagnetic theory the internal space is regarded as a circle. Physical phenomena are independent of rotations about the circle just as they are independent of rotations in space. The conserved quantity associated with the symmetry is the electric charge.

The richest and most interesting field theories are the ones in which there can be independent symmetry transformations at each point in space and time. Such symmetries are called gauge symmetries, and the theories are called gauge theories. The earliest and simplest example of a gauge theory is again Maxwell's theory of electromagnetism. A circle is associated with each point in spacetime, and the laws of electromagnetism are not altered by independent rotations about each of the circles in this infinite set. This gauge symmetry is expressed by a symmetry group designated $U(1)$.

In any gauge theory there is a set of gauge particles that transmit the force between particles bearing a charge. In electrodynamics the gauge particle is a massless spin-1 particle called the photon. More general gauge theories with greater internal symmetry, such as the electroweak theory and QCD, also include massless spin-1 gauge particles. Such theories are known generically as Yang-Mills theories, after C. N. Yang of the State University of New York at Stony Brook and Robert L. Mills of Ohio State University. The forces described by Yang-Mills theories are nongravitational; they are called Yang-Mills forces. Finally, even Einstein's theory of gravity is a kind of gauge theory, but the spin of its gauge particle, the graviton, is 2.



CHARGES associated with the electroweak force and the strong force depend on the separation between the particles carrying the charges. Quantum theory leads to fluctuations in energy, which are manifest as a sea of "virtual" particles throughout space. The virtual particles carry charges that can shield the separated particles, thereby causing the variation in the strength of the charges. The other three horizontal coordinates indicate equivalent ways of understanding the variation. If one extrapolates from the measured values of the charges at an energy of 100 GeV (100 billion electron volts), the three charges appear to be equal in strength at about 10^{17} GeV. The extrapolation suggests a larger, "grand unified" symmetry prevails above that energy, in which the electroweak and strong forces become indistinguishable. The grand-unified energy, however, may not be distinct from the energy at the Planck scale (10^{19} GeV), in which case the force of gravity cannot be neglected.

One important aspect of a symmetry is that it can appear to be broken; the spontaneous appearance of a broken symmetry signals a change of state called a phase transition of a system of particles. For example, a lump of iron is not magnetized at high temperatures because the magnetic moments of the atoms point randomly in all directions. Since no direction is picked out, the system has rotational symmetry. When the iron is cooled, there is a change of phase: the atomic moments line up and the iron becomes magnetized in a particular direction. The rotational symmetry appears to be broken. Note that the rotational symmetry of the laws governing the forces on a microscopic scale remains intact; the appearance of a broken symmetry is characteristic of a system (the lump of iron) made up of a large number of particles.

Similarly, many theories of elementary particles, including superstring theories, require large gauge symmetry groups in order to give a unified account of diverse phenomena. Typically such large symmetry is apparent only at extremely high temperatures; the symmetry must appear to be broken at ordinary terrestrial temperatures if the theory is to be consistent with observations. For example, the electroweak theory describes a unified version of the electromagnetic and weak forces called the electroweak force, whose gauge symmetry is based on a group called $SU(2) \times U(1)$; the group is an extension of the group $U(1)$ associated with electromagnetism. At ordinary temperatures, however, one observes two forces, the electromagnetic and weak forces, which are entirely distinct. The symmetry associated with the unification of the two forces becomes apparent only at temperatures much higher than 10^{15} degrees Celsius.

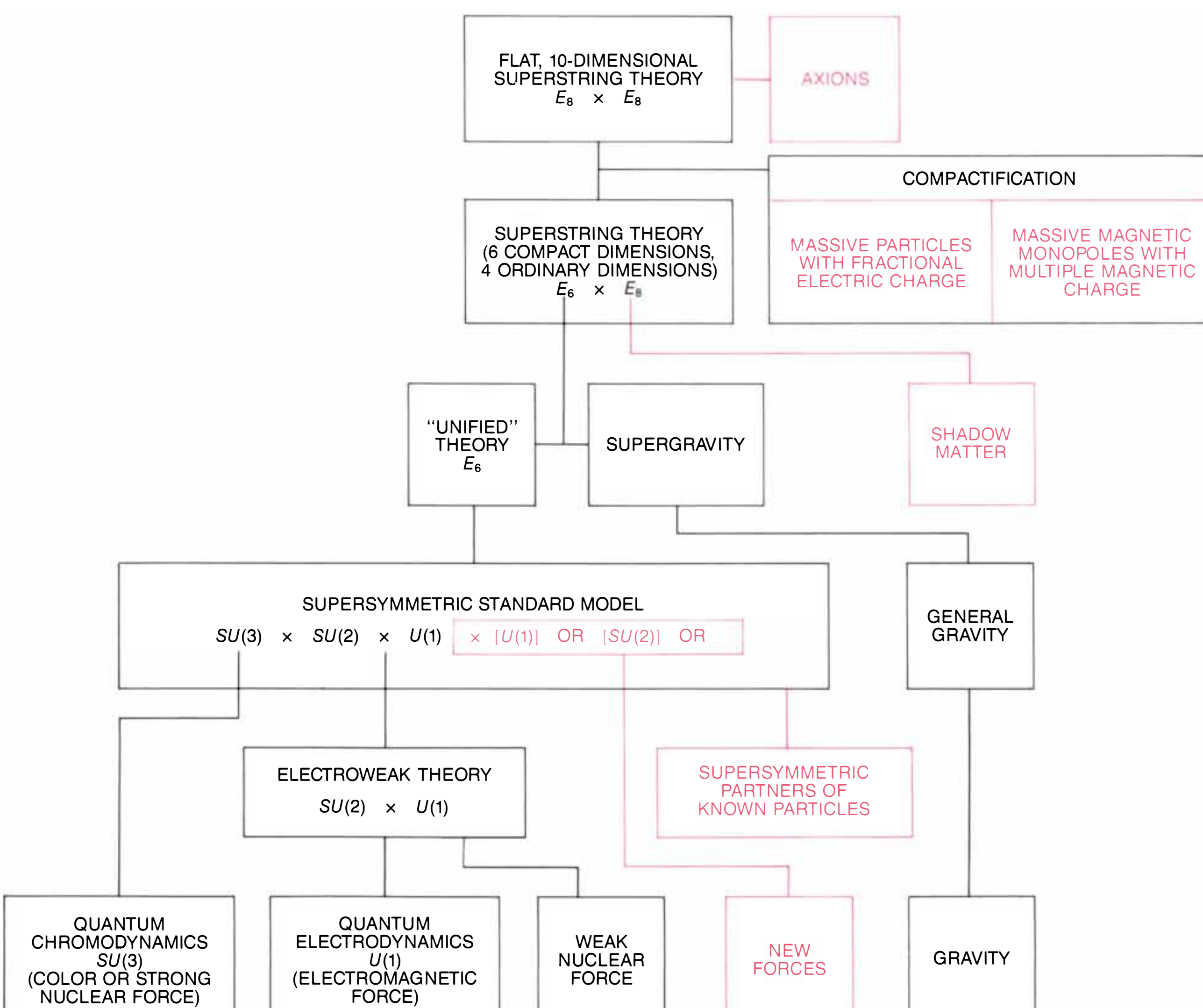
The success of the quantum field theories of point particles gave a new lease on life to quantum field theory, and many physicists turned their attention to more ambitious schemes for unification. Such schemes were almost invariably based not on string theories but on more elaborate, so-called grand-unified symmetries built into quantum field theories with point particles. The grand-unified schemes, which ignore gravity, were associated with symmetry groups called $SU(5)$, $SO(10)$ or E_6. These large symmetries can break into smaller symmetries associated with the group $SU(3)$ of QCD and the group $SU(2) \times U(1)$ of the electroweak theory.

The temperatures at which grand-unified symmetries—and indeed the effects of quantum gravity—might become important are extraordinary: between 10^{30} and 10^{32} degrees C. According to current thinking about the origins of the universe, these temperatures were realized only between 10^{-43} and 10^{-38} second after the big bang. In spite of the brevity of this period, its implications for the subsequent evolution of the universe have been profound. In this way it turns out that the physics of the incomparably small is a key to the understanding of phenomena on a cosmic scale.

Many of the new grand-unified theories also incorporated supersymmetry, a symmetry that transforms bosons and fermions into one another and thereby unifies particles having integer and half-integer spin. In a supersymmetric theory there are equal numbers of bosons and fermions for any given mass. Several recent attempts have been made to combine Einstein's theory of gravity with supersymmetry. The resulting so-called supergravity theories belong to a new kind of gauge theory in which the gauge particle responsible for supersymmetry is called the gravitino; its spin is 3/2. The most popular supergravity theory for a time was formulated in 11 spacetime dimensions: the four ordinary spacetime dimensions and seven additional spatial ones.

It is ironic from the current theoretical standpoint that one of the most important early discoveries in string theory was made in the process of studying higher-dimensional supergravity theories. In 1976 Ferdinando Gliozzi of the University of Turin, Scherk and David A. Olive of the Imperial College of Science and Technology in London suggested that the spinning-string theory could be made supersymmetric. The implications of this argument, however, were largely unrecognized, and work on string theory was virtually abandoned. It then lay dormant until 1980, when Schwarz and I (together with contributions from Lars Brink of the Chalmers Institute of Technology in Göteborg) began constructing and investigating the properties of string theories having spacetime supersymmetry.

In order to understand why a solution to the problems of reconciling gravity with quantum theory has proved so elusive one must consider the implications of Heisenberg's un-



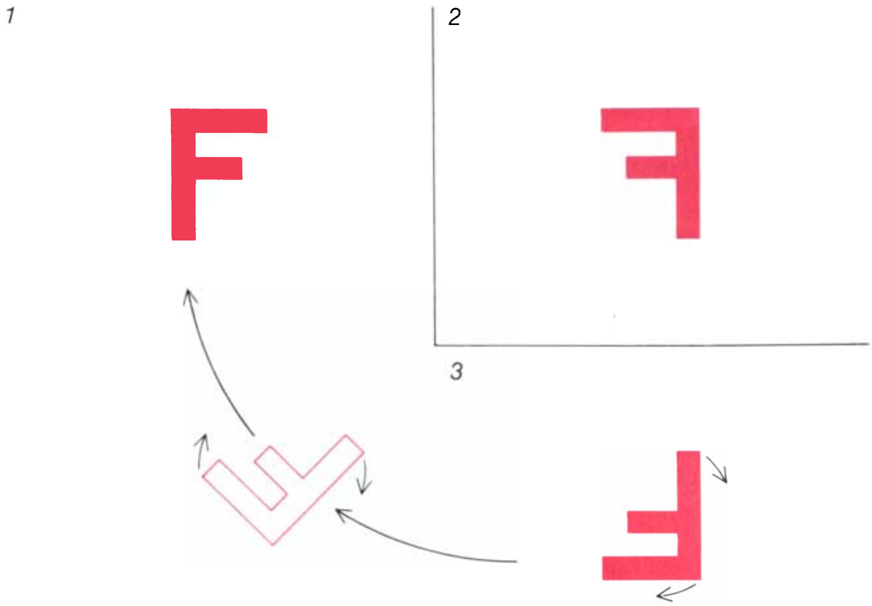
SUPERSTRING THEORY that is based on the huge symmetry group $E_8 \times E_8$ might make contact with observed physical phenomena by way of the connections shown in the logical diagram. When the effects of curvature in the six curled-up dimensions are considered, the theory resembles a supersymmetric grand-unified theory linked to supergravity, whose symmetry is associated with the group $E_6 \times E_8$. That symmetry can break down to give the standard model of elementary particles and forces, which accounts for the three nongravitational forces. New particles and forces that may be required by certain interpretations of the theory are shown in color.

certainty principle at distance scales of less than about 10^{-15} meter. According to that principle, the more precisely a spatial measurement is made, the less precisely the momentum or the energy of the system being measured can be known. The uncertainties of energy are realized as fluctuations at short distances; because energy and mass are equivalent, the energy fluctuations can be manifest as the creation of "virtual" particles. Virtual particles and antiparticles can materialize out of the vacuum for a short time before they annihilate one another. The sea of virtual particles gives rise to multiparticle effects similar to the ones that arise for a substance such as a ferromagnet made up of many atoms.

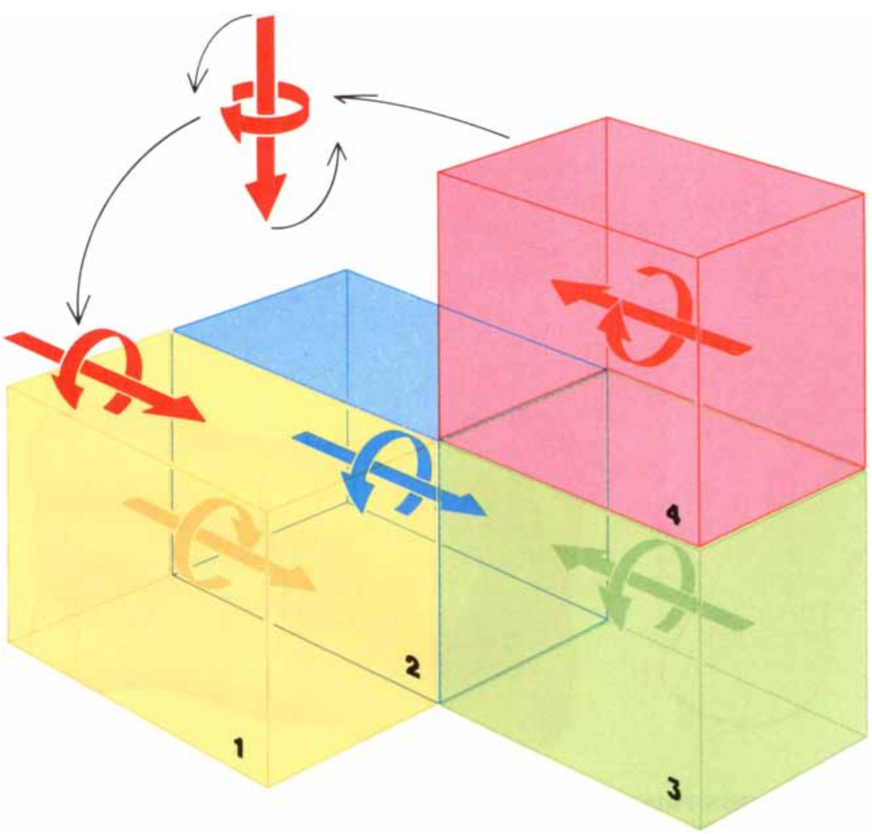
When the spatial resolution is less than about 10^{-35} meter, the energy fluctuations become so huge that, according to general relativity, virtual black holes form. The energy of the fluctuations, about 10^{19} GeV (a GeV is one billion electron volts), is called the Planck energy, and the distance scale is called the Planck distance. Spacetime must therefore be considered highly curved on small scales; in a sense it is foamy. This conclusion has disastrous consequences for the usual computational techniques of quantum field theory because it invalidates the notion of spacetime as a smooth collection of points. In all such calculations one assumes the curvature of spacetime is small; if the assumption does not hold, the calculations give nonsensical infinite results. This argument suggests that at short distances or, equivalently, at high energies either general relativity or quantum mechanics (or both) must be revised.

The corresponding argument cannot be made for strings, since superstring theory radically modifies the assumptions of general relativity at short distances. Indeed, as I outlined above, superstring theories may require a revision of general relativity at its most fundamental level: the idea of a curved spacetime, which is central to Einstein's theory, may have to be extended to the infinitely richer space of string configurations. The current understanding of the theory, however, begins with a more primitive model, in which the strings are treated as independent particles moving in some fixed spacetime background. Even with this restriction the quantum-mechanical treatment of a string leads to strong constraints on any superstring theory of gravity in higher dimensions.

There are two kinds of strings, open and closed. Open strings have endpoints, and conserved charges, such as the electric charge, that are associated with the Yang-Mills forces are



CHIRALITY, or handedness, cannot be defined in an even number of spatial dimensions (or in odd-dimensional spacetime). For example, in two dimensions successive reflections of the asymmetric letter F through both coordinate axes can be undone by a rotation.



MIRROR REFLECTION through all the coordinate axes in a space with an odd number of dimensions (or an even-dimensional spacetime) gives rise to the possibility of finding chiral particles in the space. For example, a neutrino (*1*) travels at the speed of light and spins in a left-handed sense about the direction of its motion. Its mirror image after reflection through all three axes (*4*) spins in a right-handed sense; in other words, it is not equivalent to any rotated version of the original neutrino. Only the left-handed neutrino occurs in nature, which shows the laws of physics are asymmetric with respect to handedness.

tied to the endpoints. Particles associated with the vibrational states of an open string include the massless spin-1 gauge particles, but they do not include the graviton.

When open strings collide, they can interact by touching and joining at their endpoints to form a third string; the third string can then split apart to form two final strings. Similarly, the two endpoints of an open string can join to form a closed string. The vibrational states of a closed string include the massless spin-2 graviton. Thus in any theory with open strings there are also closed strings, and in any string theory with closed strings it is inconsistent to neglect the force of gravity. Accordingly, if the Yang-Mills forces, such as electromagnetism, are included in a string theory, they must be unified with gravity in an intimate way.

A kind of theory in which the Yang-Mills forces can be associated with closed strings was formulated by David J. Gross, Jeffrey A. Harvey, Emil J. Martinec and Ryan Rohm of Princeton University. Such a theory is known as heterotic, and it is the most promising kind of superstring theory developed so far. Its construction is quite strange. The charges of the Yang-Mills forces are included by smearing them out over the entire heterotic string. Waves can travel around any closed string in two directions, but on a heterotic closed string the waves traveling clockwise are waves of a 10-dimensional superstring theory; the waves traveling counterclockwise are waves of the original, 26-dimensional string theory. The extra 16 dimensions are interpreted as internal dimensions responsible for the symmetries of the Yang-Mills forces.

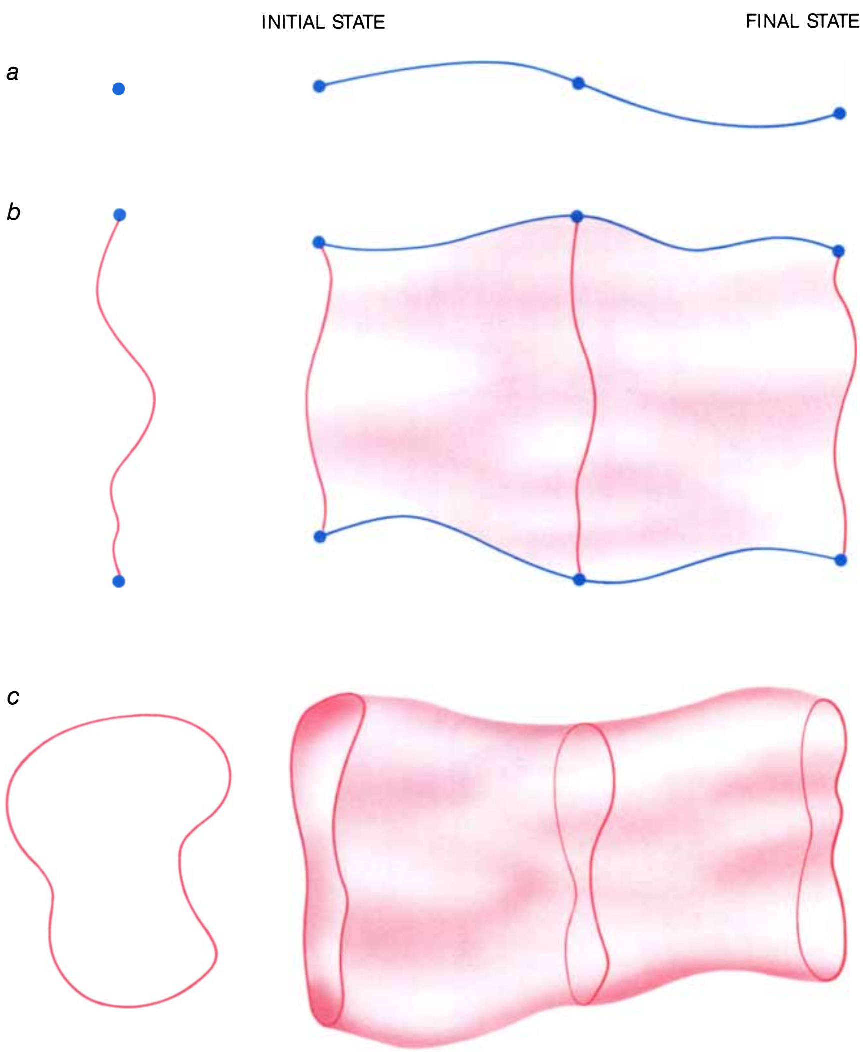
As a string moves it sweeps out a two-dimensional surface in spacetime called a world sheet, just as a point particle sweeps out a world line [*see illustration on this page*]. In classical, or non-quantum-mechanical, general relativity a particle moves along the world line that minimizes the so-called action of the particle: its energy as it moves through time. The action is proportional to the length of the world line, and so a path of least action is a geodesic, or the shortest distance between two points in spacetime.

The motion of a string is treated in an analogous way. In a non-quantum-mechanical approximation the string also moves in a way that minimizes its action. The action is proportional to the area swept out by the string, and so the world sheet must be a surface of minimum area. If time is regarded as a spatial dimension, the world sheet swept out by a closed string can be thought of as a kind of soap film that joins the string at its starting point and at the end of its path in spacetime.

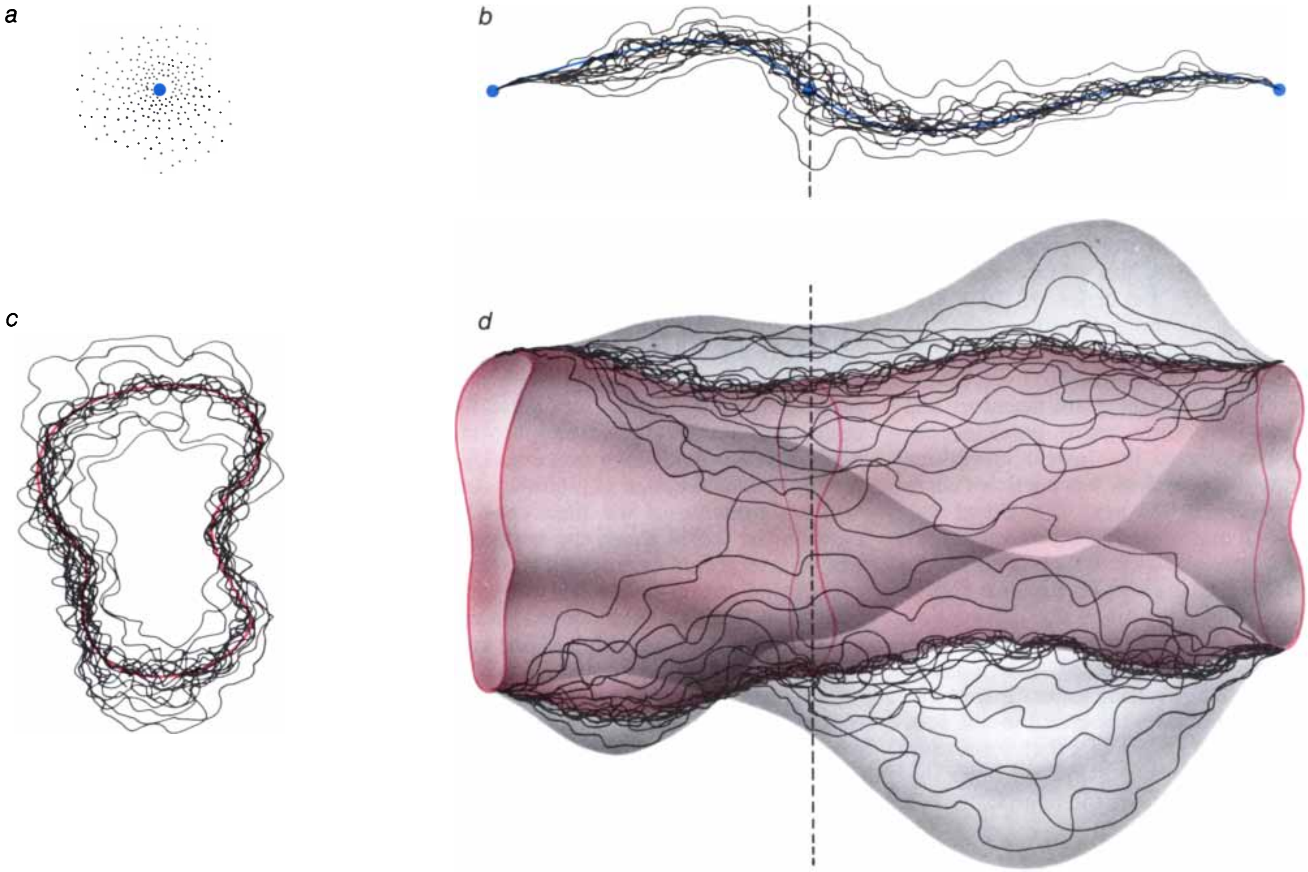
There is an enormous symmetry embodied in the condition that the motion of the string is determined by minimizing the area of its world sheet. The area is a geometric quantity that is independent of how the points on the two-dimensional sheet are labeled. No amount of distortion of the labeling can change the underlying geometry of the sheet, and physics is said to be symmetric under arbitrary relabelings of the world-sheet coordinates.

Because of this symmetry, there is no physical significance to distortions of the world sheet that lie in the two directions tangent to the surface of the sheet at any point. The only meaningful vibrations of the string are undulations of the world sheet perpendicular to its surface. Hence if the string is moving in d dimensions, two directions of vibration have no physical reality. All real vibrations are transverse vibrations in $d - 2$ dimensions.

It is worth noting that such a constraint on the vibrational modes of photons, or waves in an electromagnetic field, is a hallmark of Maxwell's theory. There is no physical meaning to vibrations of the photon in the time direction, and longitudinal vibrations, which take place in the same direction as the wave is moving, are not possi-



LEAST-ACTION PRINCIPLE determines the classical, or non-quantum-mechanical, trajectory of a particle: if time is treated as a fourth spatial dimension, the trajectory is the line of minimum length in spacetime joining the initial state and the final one (*top*). Similarly, according to the least-action principle, a string sweeps out a world sheet of minimum area in spacetime. Cross sections are shown at the left. Open strings carry charges at their endpoints (*blue*) that define the boundaries of the sheet (*middle*). Closed strings are loops, and so they have no endpoints (*bottom*). A string satisfying the least-action principle can vibrate in any combination of harmonic frequencies. Because deformations of the world sheet in the two directions tangential to its surface at any point are not distinct configurations of the sheet, only the vibrations perpendicular to its surface have physical reality.



"SUMMING OVER HISTORIES" is one way to take account of the probabilistic interpretation of the path of a particle in quantum mechanics. A point particle is pictured as moving simultaneously along all possible world lines in spacetime; a weight, or statistical probability, is assigned to each path in such a way that the shortest paths in spacetime are by far the most probable. The result is a collection of world lines that are densest along the classical trajectory (*upper right*). The cross section of the world lines through time, indicated by the broken vertical line, is shown at the upper left. The density of the world lines through an arbitrary region of the cross section represents the probability of finding the string in that region at the corresponding time. Similarly, the quantum mechanics of a string is obtained by considering all possible world sheets with the same initial state and final state, weighted in such a way that the ones of smallest area are by far the most probable (*lower right*). The density of the closed loops in a given region of the temporal cross section of the world sheet (*lower left*) represents the probability of finding the complete string in that region at the corresponding time.

ble for a wave traveling at the speed of light. The gauge symmetry of electromagnetism ensures the absence of such nonphysical vibrations.

From this point of view one might expect insurmountable problems in a string theory, since a string is made up of an infinite number of points. Each point along the string vibrates, and so the potential for nonphysical vibrations of a string is infinitely greater than it is for a point particle such as the photon. The absence of such vibrations is guaranteed by the symmetry under coordinate relabelings. When the quantum mechanics of the string is considered, the same symmetry also introduces powerful constraints on possible string theories. It is in this sense that superstring theory is immensely elegant.

Up to now I have discussed the string as a classical particle. In quantum mechanics the motion of a particle is not precisely defined. When it moves through space, any particle tries, in effect, to take all possible paths between its initial state and its final state. The probability of each path is weighted so that a higher probability is assigned to paths of lower action; the classical path of least action is the most probable one. This formulation of the quantum behavior of a particle is called the method of summing over histories, and it was devised by Richard P. Feynman of Caltech. Its application to string theory has been most completely developed by Stanley Mandelstam of the University of California at Berkeley and by Alexander M. Polyakov of the Landau Institute for Theoretical Physics near Moscow.

In superstring theory summing over histories requires summing over all the possible surfaces that join the initial and final states of a string or a set of interacting strings. The different paths can be thought of as fluctuations of the world sheet, which are just like the random, shimmering motions of a soap film at any temperature above absolute zero [*see illustration above*].

An important quantum-mechanical constraint on string theories was formulated in 1972 by Richard Brower of Boston University, Peter Goddard of the University of Cambridge and Charles B. Thorn of the University of Florida at Gainesville: Requiring symmetry under relabelings of the world-sheet coordinates of a free string undergoing quantum fluctuations is sufficient to determine the number of dimensions of the background spacetime in which the string is moving. Remember that in the original dual-resonance theory the number of spacetime dimensions is 26; in superstring theory 10 dimensions are required. This result had been intimated a year earlier in work by Claud W. V. Lovelace of Rutgers University. The constraint is a striking instance of how the formulation of the theory in terms of the two-dimensional world sheet leads to a rich and highly determined structure for the space in which the string is moving.

The vibrational frequencies of a superstring are determined by its tension, which is measured in units of energy per unit length, or mass squared in fundamental units. Since the theory is to describe gravity, the string tension must be closely related to the Planck energy, which is the only dimensional parameter in a gravitational theory. The tension of the string must therefore be on the order of 10^{19} GeV squared; in more familiar units that is equal to a force of 10^{39} tons. The frequencies of the normal vibrational modes of the string are therefore separated by huge gaps: particles corresponding to the lowest vibrational state are massless, but particles corresponding to the next vibrational state have a mass roughly equal to the mass of a speck of dust, which is enormous for an elementary particle. For higher-frequency vibrations the corresponding masses increase without limit.

The significance of the massless states of superstring theory is that they include not only the graviton, the spin-1 gauge particles of other forces and the spin-0 and spin-1/2 particles, but also the gravitino, the spin-3/2 gauge particle associated with supergravity. Thus for energies below the Planck energy the massless particles of superstring theories are the same ones found in supergravity theories.

A quantum string theory is different from the quantum theory of a point particle in another important respect. Again consider the sum over histories of a single closed string: the sum includes all possible connected surfaces that can be stretched, twisted or otherwise smoothly deformed into a cylinder without tearing. All such surfaces are said to be topologically equivalent to the cylinder, and they include surfaces with long tentacles [*see top illustration on page 59*]. Under certain assignments of the time coordinate on the world sheet, the tentacles can be interpreted as motions in which two closed strings join to form a new one or in which a new closed string breaks away from the original one and disappears into the vacuum.

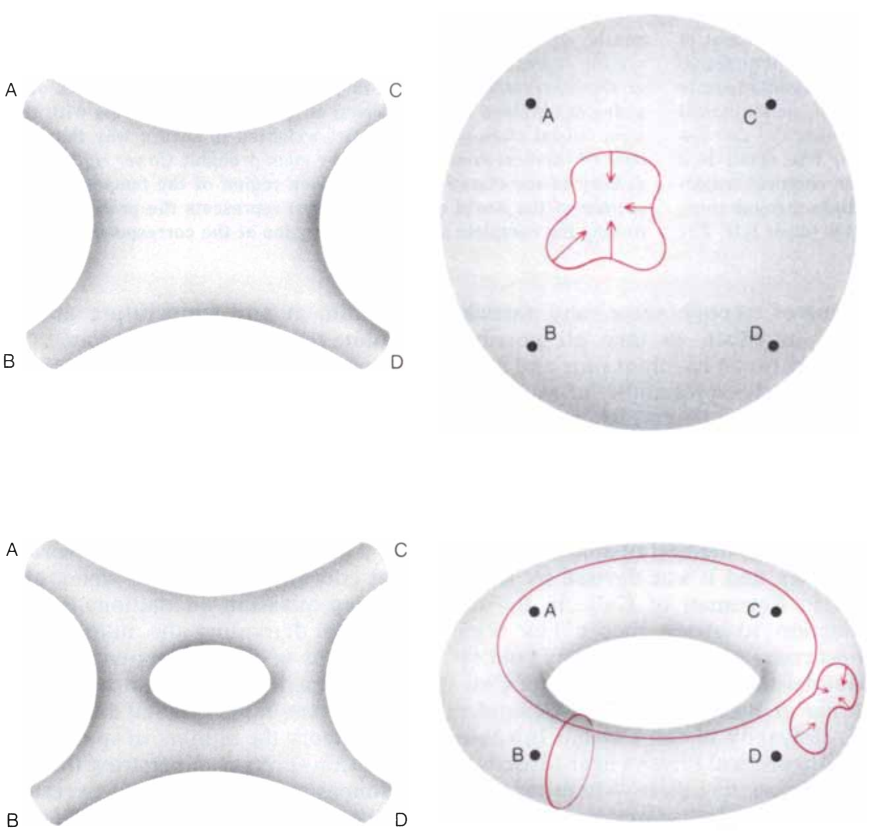
Thus the string automatically interacts with the background space in which it moves, even though interactions are not explicitly included in the sum over histories. In contrast, the sum over histories of a single point particle does not include any information about interactions with the background space. Because of the nature of the interaction of a string with its background, even the motion of a single string can be consistently described only if the curvature of the background space is severely constrained. This result was originally implied by the work of Daniel Friedan of the University of Chicago, done in 1979.

To appreciate the significance of the finding one must understand that the 10 dimensions required in a consistent superstring theory were not initially assumed to be curved at all. The theory was first formulated under the simplifying assumption that all 10 dimensions are equivalent, that is, all of them are "flat." If any superstring theory is to make sense of physical observations, however, six spatial dimensions must be highly curved. If the sum over the histories of the world sheet is to be consistent, the six dimensions must be curled up in one of a few special ways. They are said to form a kind of space called a Calabi-Yau space (after Eugenio Calabi of the University of Pennsylvania and Shing-Tung Yau of the University of California at San Diego); they may also form a generalization of such a space called an orbifold. These spaces lead to a promising scheme for explaining the physics of the four observable dimensions.

Much of the interest in superstring theories follows from the rich structure that results by requiring the theory to be consistent. If the theory is to give a realistic quantum account of the Yang-Mills forces, there is an empirical constraint in addition to the requirement that there be only four observable dimensions: the theory must lead to the observed chirality, or handedness, of the weak force.

The weak force is responsible for radioactive decay such as beta decay, which is an important reaction in the sun. The force is chiral (from the Greek word *cheir,* meaning hand) in the sense that it gives rise to effects whose mirror-reflected counterparts do not exist in nature. Unfortunately the quantum-mechanical version of any chiral gauge theory is likely to violate conservation laws such as the conservation of electric charge. Such a violation is called a chiral anomaly; it signals a breakdown of gauge symmetry, which renders the theory inconsistent. Devising a theory that is chiral and yet avoids chiral anomalies is a delicate matter in four dimensions, and until recently it was thought to be impossible in 10 dimensions.

It is only when space has an odd

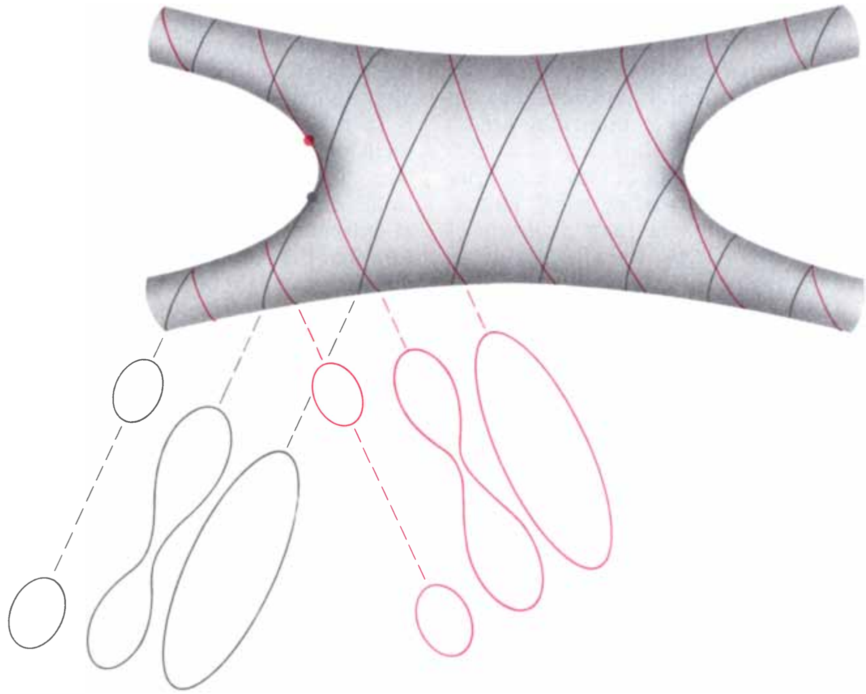


INTERACTIONS AMONG STRINGS are represented by world sheets topologically equivalent to a sphere or to a doughnut with an arbitrary number of holes. The surfaces are continuous except at the points that represent the incoming or outgoing strings (*A, B, C, D*). The sphere is distinct in that any closed curve on its surface can be continuously shrunk to a point. Two kinds of closed curves on the torus cannot be continuously shrunk to a point.

number of dimensions (or when spacetime has an even number) that the concept of chirality can be defined at all. In any number of dimensions chirality depends on the outcome of successive mirror reflections along all the spatial axes. When space has an odd number of dimensions, mirror reflection along each spatial axis gives an odd number of reflections, and so a left-handed shape is transformed into a right-handed one. When space has an even number of dimensions, a reflection along each spatial axis leaves any shape unaltered: the reflected image can be rotated into the original shape. For example, the popular 11-dimensional theory of supergravity cannot lead to a chiral theory because it is formulated in 10 spatial dimensions (an even number). With nine spatial dimensions superstring theory can be chiral.

Even when the higher-dimensional theory is chiral, the process of curling up the extra dimensions generally washes out chirality. It is now thought that the observed chirality can be explained only if the chiral higher-dimensional theory is initially formulated to include the gauge symmetry group of a Yang-Mills force, in addition to the gravitational force. The symmetry group must be present in the flat, 10-dimensional theory before the possibility that the extra dimensions curl up can be considered.

In August, 1984, to the surprise of many physicists, Schwarz and I showed that a chiral theory free of anomalies can be formulated in 10 dimensions, provided the symmetry group of the Yang-Mills force is one of two special groups. The groups must be either $SO(32)$, a generalization to 32 internal dimensions of the symmetry of spacetime, or the group $E_8 \times E_8$, the product of two exceptional, continuous groups that were first discovered by the French mathematician Elie Cartan. A third group known as $O(16) \times O(16)$, which also leads to freedom from anomalies, has recently been noted; its symmetry is a subsymmetry of the other two possible groups. The fact that the quantum consistency of a theory including gravity leads to an almost unique prediction of the unifying symmetry group was an exciting development. It has led to the current wave of enthusiasm for superstring theory.



PRECISE INSTANT at which two strings join cannot be objectively specified because the definition of the time coordinate of a point on the world sheet depends on the observer. The colored coordinate lines indicate contours of equal time as defined by one observer. To this observer the two incoming strings appear to join at the colored point. To another observer the black lines are contours of equal time; the strings appear to join at the black point.

The quantum mechanics of a single string is only an approximation to a full theory, not yet developed, in which the interactions of arbitrary numbers of strings would be described. Interactions arise when strings join or split. The probability that a given set of incoming strings leads to a given set of outgoing strings is determined by an infinite sequence of so-called Feynman diagrams. The diagrams describe all possible joinings and splittings of the world sheets, summed over histories for each case.

The simplicity of the sequence of possible Feynman diagrams for strings contrasts sharply with the complexity of the possible Feynman diagrams for point particles. Consider the possible diagrams for two interacting closed strings. First the two strings can join and then split apart. The Feynman diagram for the process is topologically equivalent to the surface of a sphere [*see illustration on opposite page*]. In the next term in the sequence the intermediate string splits into two strings, which then rejoin. The diagram is topologically equivalent to a torus, or the surface of a doughnut. The sequence of diagrams is continued simply by adding holes to the doughnuts: two holes, three holes and so on.

Requiring symmetry under relabelings of coordinates on the torus introduces strong, new constraints to superstring theory. If the torus is cut in either one of two ways, twisted and glued back together, its topology is unchanged; the twisted coordinate system, however, cannot be continuously returned to the original one without cutting the surface again [*see illustration on next page*]. The invariance of the torus under such changes in coordinates leads to an important constraint in the heterotic superstring theory. A heterotic string can move through flat, 10-dimensional spacetime only if the symmetry of the nongravitational forces in the theory is described either by the group $E_8 \times E_8$, the group $spin\ (32)/Z_2$, a variant of $SO(32)$, or the group $O(16) \times O(16)$.

The restriction to these groups is striking. Remember that Schwarz and I had earlier singled out the same groups by requiring the absence of chiral anomalies in any consistent 10-dimensional chiral gauge theory. The constraint that leads to these groups in the heterotic theory is related to remarkable mathematical properties of certain 16-dimensional lattices of points. Such lattices are constructed from the 16 internal dimensions I mentioned above.

There is another and almost equally striking feature of the torus diagram that indicates the profound difference between string theories and the theories of point particles. Among the many kinds of one-loop Feynman diagrams for point particles that are the analogues of the torus diagram, there are diagrams that give infinite answers when the sum over histories is calculated. Such infinities arise because the diagrams for point particles include the sum over histories when two interaction points are arbitrarily close in spacetime. The contributions to the entire process of each one of an indefinitely large number of fluctuating paths must be taken into account, and that leads to infinite answers. In contrast, on a string Feynman diagram no one point can be identified as a splitting or joining point of two strings [*see illustration on preceding page*]. Hence, at least in the one-loop diagram of string interactions, the concept of an interaction point does not arise. The sum over histories leads to finite answers in precisely those superstring theories whose Yang-Mills forces are associated with the symmetry groups $E_8 \times E_8$ or $SO(32)$.

An unfortunate feature of any quantum theory of gravity, whose natural energy scale is necessarily enormous, is the difficulty of extracting testable or observable predictions at more modest energies. For example, although superstring theory is initially formulated in 10 flat spacetime dimensions with huge unifying symmetries, it is only at an inconceivably small distance scale or, equivalently, at extremely high energy or temperature that the curvature of the extra six dimensions might be negligible and the full symmetry of the theory would be realized. Thus any attempt to derive low-energy consequences of the theory must be considered speculative, and there are severe problems in obtaining concrete predictions. Nevertheless, it is now possible to give a plausible picture of how superstring theory might make contact with the phenomena observed in accelerators.

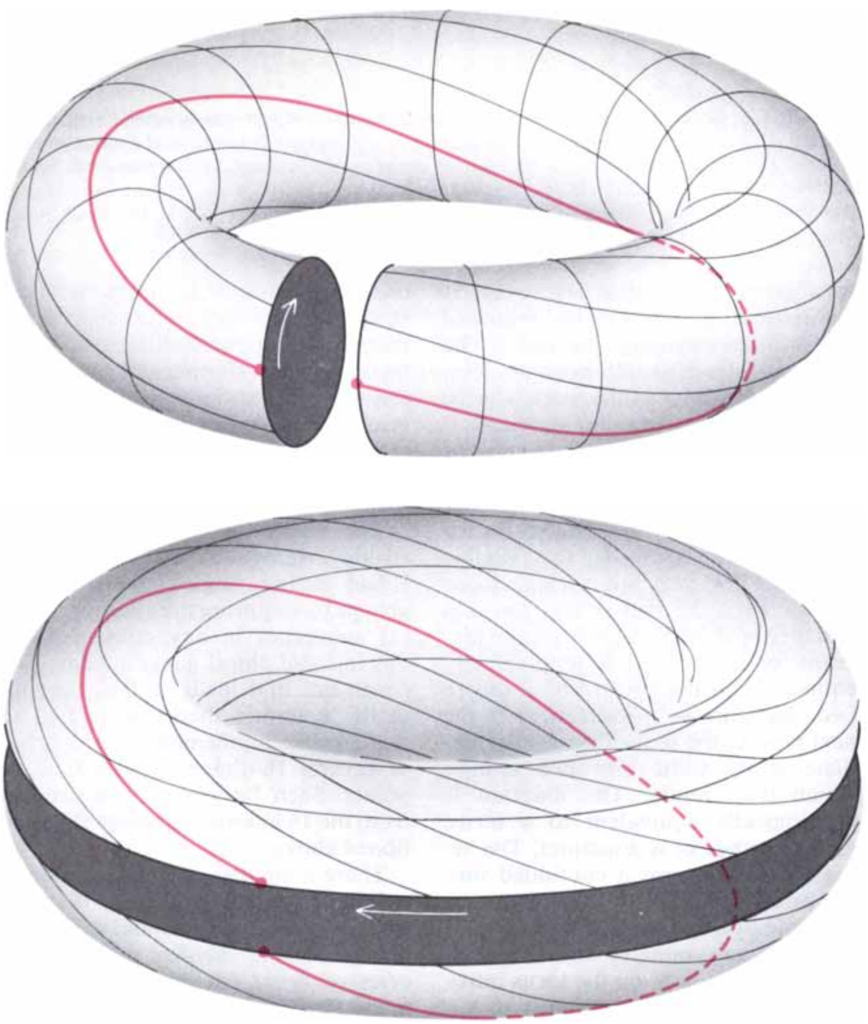
Many aspects of this picture are based on topological arguments, whose application to physics has been pioneered by Edward Witten of Princeton. Witten noted that the curling up of the extra dimensions and symmetry breaking go hand in hand: there must be a strong correlation between the curvature of the extra dimensions and the way in which the huge symmetry associated with the Yang-Mills forces breaks down to a smaller symmetry. In flat 10-dimensional spacetime the average values of the fields associated with the $E_8 \times E_8$ or $SO(32)$ Yang-Mills forces are zero. Witten showed that if the curvature of the space is nonzero in certain dimensions, the average values of these fields in the curved space must be nonzero in the same dimensions.

A nonzero average value for a field signals a transition to a phase with less symmetry, just as the nonzero magnetization of a ferromagnet signals the transition to the less symmetric magnetized phase. Thus if six dimensions of the original flat spacetime become highly curved, the Yang-Mills forces in the theory of the resulting curved space are unified by a subsymmetry of $E_8 \times E_8$ or $SO(32)$. This result is just what one would want.

So far it has not been proved that the six spatial dimensions must be curled up in superstring theory. As I stated above, however, the requirement that the theory retain its large symmetry under relabelings of the coordinates of the world sheet forces the curled-up dimensions to form a Calabi-Yau space or perhaps an orbifold. Assuming six dimensions are curled up in this way, Philip Candelas of the University of Texas at Austin, Gary T. Horowitz of the University of California at Santa Barbara, Andrew E. Strominger of the Institute for Advanced Study in Princeton and Witten have devised an imaginative scheme for understanding how superstring theory might relate to phenomena measured at the energies of present experiments.

In the scheme of Candelas and his coworkers the average value of the electric field associated with the group $E_8 \times E_8$ is set equal to the nonzero value of the curvature of the six-dimensional space. Because of special features of Calabi-Yau spaces (or the associated orbifolds), the symmetry breaks down in a special way to the group $E_6 \times E_8$, while still preserving its supersymmetry. It is reassuring to derive the group E_6 in this way because it is one of the groups studied in efforts to unify the Yang-Mills forces.

The extra group factor E_8 enlarges



INVARIANCE of the labeling of the space and time coordinates on a world sheet with the topology of the torus is an even stronger requirement than it is on the sphere. A torus can be cut like a sausage (*upper diagram*) or like a yo-yo (*lower diagram*). One side of the cut can be given a full twist with respect to the other side, and the two boundaries can then be "glued" back together. The points on the surface are thereby labeled according to a set of coordinates (*black and colored curves*) that cannot be continuously deformed to the original set of coordinates. The requirement that superstring theory not be sensitive to such relabelings leads to strong conditions on the symmetry groups associated with the nongravitational forces of the theory. Symmetry groups must be either $SO(32)$, $E_8 \times E_8$ or $O(16) \times O(16)$.

the symmetry enormously, but it does not directly affect the observed particles. Particles carrying the forces associated with the extra E_8 symmetry are neutral with respect to the observed Yang-Mills forces, but they should exert gravitational attraction. It has been speculated that matter made up of such particles, which has been called shadow matter, might account for part of the unseen mass known to be present in the universe.

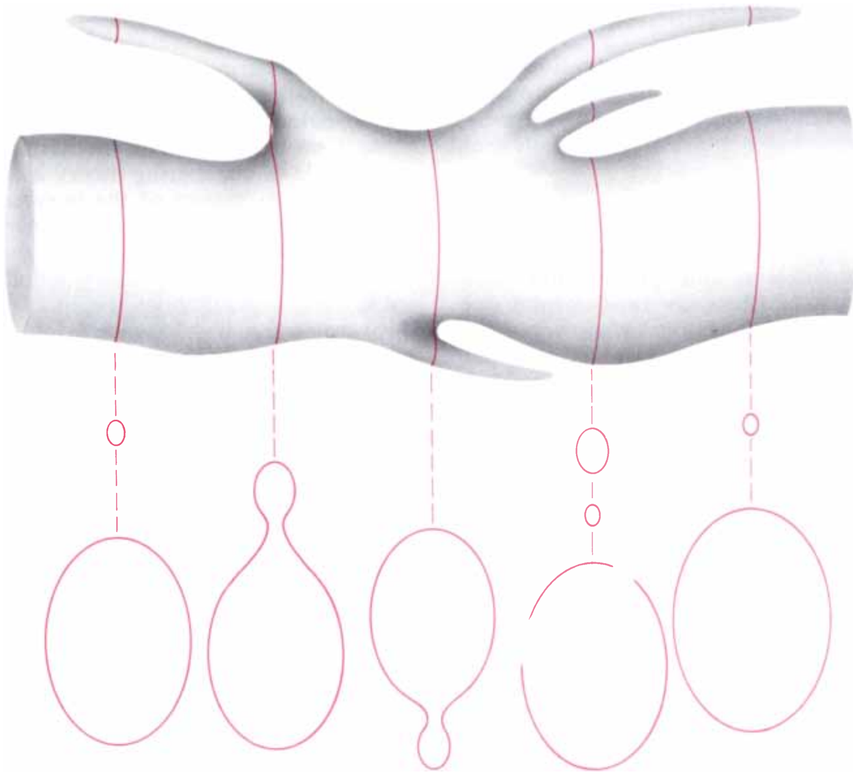
One immediate prediction of this interpretation of superstring theory is the existence of a particle called the axion. Such a particle has been postulated in order to avoid a violation of an important symmetry in QCD, the theory of the strong interaction. Although the axion has not yet been found, several groups of investigators are planning experiments that could detect it.

Many other physical consequences follow from purely topological properties of the six curled-up dimensions. For example, one of the most basic topological properties of a space is its Euler number. If the curled-up space were two-dimensional, the Euler number would be 2 minus twice the genus of the surface, where the genus is the number of holes. In six dimensions the Euler number is less easy to describe.

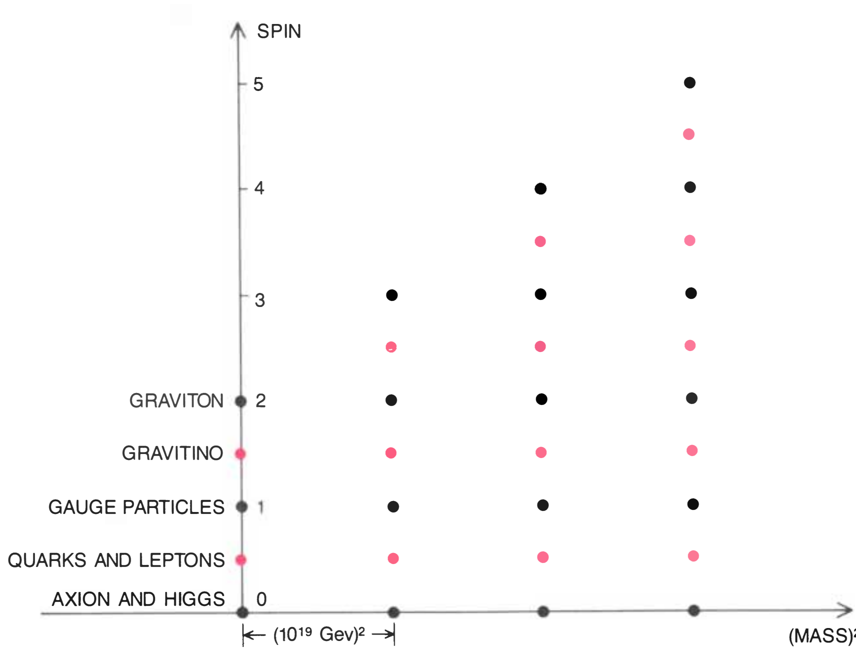
The observed fermions are naturally grouped into "generations" of quarks and leptons [*see illustration on page 50*]. (The leptons are fermions that do not strongly interact.) Three generations have been discovered so far, and there is probably at most one more. The predicted number of generations is half the Euler number of the six-dimensional compact space. Because only a few of the known Calabi-Yau spaces and orbifold spaces have a small Euler number, only a few such spaces are candidates for describing observed physical phenomena.

The six-dimensional spaces known to have a small Euler number are spaces that also have "holes" in them [*see illustration on page 60*]. The presence of the holes allows the fields associated with the symmetry group E_6 to be trapped. The trapping of the fields causes the symmetry to break down further, without the need for the massive so-called Higgs particles usually associated with this symmetry breaking. Quantum chromodynamics and the electroweak theory of standard elementary-particle physics, whose symmetries are described by the group product $SU(3) \times SU(2) \times U(1)$, can readily be obtained from this latest symmetry breaking. It is likely that the same symmetry breaking will also give rise to additional forces associated with remnants of E_6.

There are many ways a closed string



QUANTUM-MECHANICAL ANALYSIS of a single world sheet leads to strong constraints on the kind of curled-up background space in which the string is moving. The quantum-mechanical sum over histories includes configurations of the world sheet that describe strings materializing from the vacuum (*left*) or breaking away and disappearing into the vacuum (*right*). Thus string interactions with the background are automatically included in the quantum treatment of one string. The only background spaces known to be consistent with such effects are the so-called Calabi-Yau spaces or related spaces called orbifolds.



SPECTRUM OF STRING STATES is plotted for the heterotic string theory in which the extra six dimensions of spacetime have been curled up. Each black dot represents a set of bosons and each colored dot represents a set of fermions. All string states that correspond to known particles are massless states; the states with nonzero mass form an infinite series whose masses are a whole number times the square of the Planck mass, which is 10^{19} GeV. For each mass the number of fermion states is equal to the number of boson states. If each possible spin direction is counted as a different state, there are 8,064 massless states, and 18,883,584 states at the first mass level; the number increases exponentially thereafter.

can get trapped in the curled-up space. For example, the string can wind through a hole; its vibrations might then be manifest as massive particles with fractional electric charge or massive magnetic monopoles with multiple magnetic charges.

Although superstring theory has already opened many lines of investigation in both physics and mathematics, there are fundamental questions that cannot be answered until more is understood about the structure of the theory. For example, why is observed spacetime approximately flat and four-dimensional? Can one prove that six dimensions are curled up? More to the point, can one explain why the four familiar dimensions are so large? Just after the big bang, when the size of the observed universe was on the order of the Planck distance, 10^{-35} meter, all 10 dimensions must have been curled up. During the subsequent expansion of the universe all but six of the dimensions must have begun to unfold and expand. A consistent account of the process might lead to the prediction of an observable cosmic remnant.

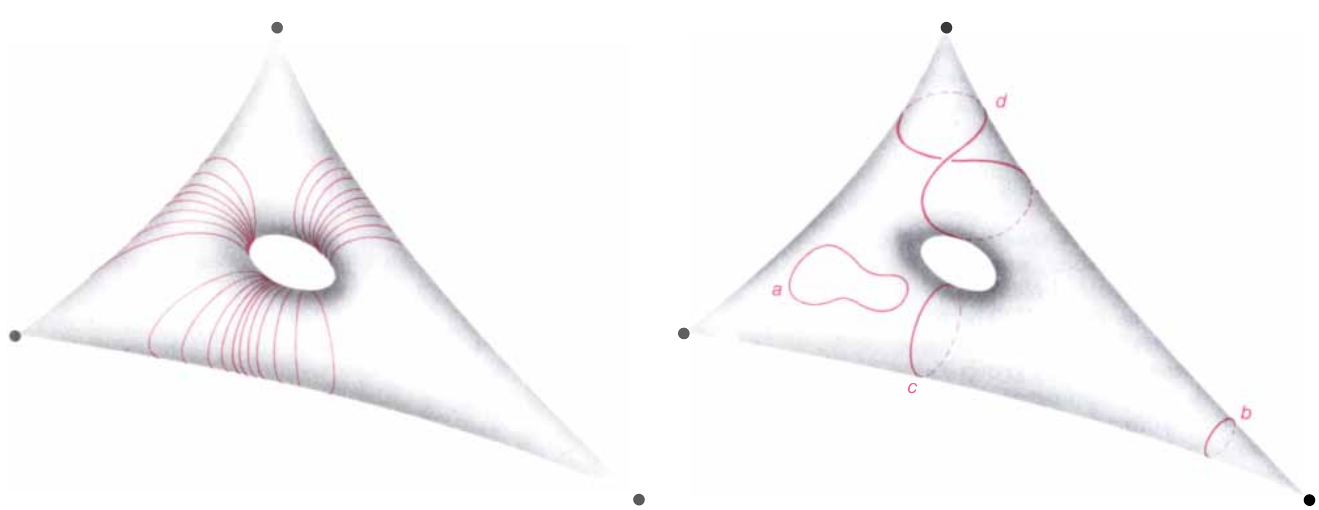
These issues are closely related to another fundamental question: Why is the cosmological constant so close to zero? This constant describes the part of the curvature of the universe that is not caused by matter; its value has been determined to be zero to within one part in 10^{120}, which makes it the most accurate measurement in all science. If superstring theory can account for the value, the explanation would be a convincing test of the theory.

As I mentioned above, the natural mass scale of superstring theory is so large that the masses of the heaviest particles observable in current accelerators are infinitesimal by comparison. As long as the supersymmetry included in superstring theories remains unbroken, the masses of the particles we observe are zero. The small, nonzero values of particle masses are therefore correlated with the breaking of supersymmetry at the relatively low temperatures of the world we observe. Nevertheless, there are strong arguments that the breaking of supersymmetry, and hence the precise values for the masses of particles observed in high-energy accelerators, cannot be understood in superstring theory as it is currently formulated.

In order to resolve such issues it is presumably necessary to understand the deep principles on which the theory is based. In a sense the development of superstring theories is in sharp contrast to the development of general relativity. In general relativity the detailed structure of the theory follows from Einstein's penetrating insights into the logic of the laws of physics. In superstring theories certain details have come first; we are still groping for a unifying insight into the logic of the theory. For example, the occurrence of the massless graviton and gauge particles that emerge from superstring theories appears accidental and somewhat mysterious; one would like them to emerge naturally in a theory after the unifying principles are well established.

How can the logic of superstring theory be discovered? The principles of general relativity must be a special case of the more general principles of superstring theory, and so in a sense general relativity can serve as a guide. For example, by developing an analogy with general relativity I explained above how superstring theories are expected to extend the idea of ordinary spacetime to the space of all possible configurations of a string. This idea is under intensive investigation. Even more radical is the suggestion that the theory should be studied solely in its two-dimensional formulation; no reference at all would then be made to the coordinates of the space and time in which we live.

Whatever the resolution of these possibilities may be, developing a deep understanding of the logical status of the theory will undoubtedly lead to profound mathematical and physical problems. It should also lead to a better understanding of the predictions of superstring theory. The prospect is for a period of intellectual ferment and rapid advance.



ORBIFOLD SPACE may form the highly curved, six-dimensional background space in which strings move. The diagrams give an impressionistic rendering of a two-dimensional version of an orbifold. A closed, two-dimensional surface has been stretched out to form three sharp points, which are called conical singularities. The surface is considered flat everywhere except at the singularities in the sense that it can be cut and laid out on a flat plane. Similarly, a six-dimensional orbifold is flat everywhere except at isolated conical singularities; there the curvature is infinite. The hole in the center of the surface in the diagram indicates that the orbifold is not "simply connected": it can be cut through without falling into separate pieces. A closed path looping through a hole in the six-dimensional orbifolds of interest can be unwound after a fixed number of loops. Such holes cause lines of electromagnetic potential associated with the symmetry group E_6 to become trapped (*colored curves at left*). The entrapment causes E_6 to break down to a smaller symmetry group, similar to the one associated with the standard model. A closed string moving in an orbifold (*right*) can move freely (*a*), but it can also become trapped in various ways. If it loops around a conical singularity, the curvature prevents it from unlooping (*b*). It can also loop through the hole (*c*) or wind first through the hole and then around a conical singularity (*d*). A string looping through the hole can lead to massive particles with fractional electric charge or to massive magnetic monopoles with multiple magnetic charges.

Photographed in the Lunar Odyssey

IT DOESN'T TAKE THE RIGHT STUFF TO TRY THE REAL STUFF.

All it takes is a trip to The Space and Rocket Center in Huntsville, Alabama.

At earth's largest space museum, you can climb inside a real space capsule. Fire a live rocket engine. Strap yourself in for a simulated mission. And even pilot a spacecraft by computer.

Witness the development of NASA's Space Station next door at Marshall Space Flight Center. Or get a closeup look at our full-size Space Shuttle, the only one on permanent display anywhere in the world.

Then, see where thousands of young people come each year to learn what training to be an astronaut is like — the incredible U.S. Space Camp.

During your visit, you can catch your breath or kick up your heels at the Huntsville Marriott, located on the grounds.

So come see our stuff soon. There's just no place on earth where you'll learn more about the space program.

Get the right stuff for planning a visit, plus information on U.S. Space Camp. Call toll-free for a copy of our free brochure.

NASA

THE SPACE AND ROCKET CENTER.
Huntsville, Alabama
800/633-7280 or 205/837-3400

SCIENCE AND THE CITIZEN

Radon and Its Daughters

The health hazards faced by citizens of industrialized nations seem almost universally to be manmade. Well-publicized risks are associated with transportation, medical care, manufacturing and energy production—above all, with nuclear power plants. Yet the source of the largest dose of ionizing radiation most people are likely to receive in their lifetime is not manmade. It is radon, a radioactive gas that emanates from rock formations as the uranium they contain decays. Many epidemiologists and environmental scientists think the gas can cause lung cancer in people who have done nothing more hazardous than live in houses built over such formations. Because the risks are potentially so great, these scientists argue for immediate remedial action. Other investigators urge that more information be collected before any far-reaching measure is taken.

Through the swirl of debate a few generally accepted facts can be discerned. Radon is a link in a radioactive-decay chain that begins with the transmutation of uranium into thorium and ends when a stable isotope of lead is reached. The gas issues continuously from the ground since uranium is present in virtually all rocks and soils. In the open air radon dissipates quickly. Inside a building, however, radon that has seeped through small cracks in the foundation, up drainage pipes or along the base of walls can accumulate in concentrations tens, hundreds or even thousands of times higher than those outdoors.

Radon itself does not pose a significant direct radiation hazard. Although the gas does emit radiation (in the form of alpha particles, or helium nuclei) that can transform a normal cell into a cancerous one, it does not remain in dangerous proximity to susceptible lung tissues for long: because it does not react chemically, inhaled radon is promptly exhaled or carried away from the lungs by the blood. Rather, it is certain "daughters," or radioactive-decay products, of the radon (primarily isotopes of polonium) that contribute the major radiation dose to lung tissue. These radioactive isotopes are chemically reactive. They can stick, either in elemental form or adsorbed onto minute airborne particles, to the lining of the bronchial passageways, whence they irradiate the surrounding tissue.

The first inkling that high, potentially dangerous radon concentrations could percolate up from the ground into buildings came from measurements made during the mid-1970's in recently built, well-insulated houses in Sweden. Prompted by the findings, the U.S. Department of Energy began to fund research on indoor radon, concerned that energy-conservation measures such as weather stripping might have led many U.S. homeowners inadvertently to construct "radon traps." Because energy-efficient houses are designed to minimize the ventilation rate (the rate of exchange of indoor and outdoor air), it was expected that high radon levels were more likely to be found in such houses.

Some energy-efficient houses were indeed found to have high radon levels, but soon many old, drafty houses exhibiting equally high levels were discovered. Most investigators now discount the importance of ventilation. Merril Eisenbud of the New York University Medical Center maintains that the erroneous association of ventilation rate with radon concentration probably came about because "household energy-conservation programs coincided with the development of on-site instrumentation for radon detection." Anthony V. Nero of the Lawrence Berkeley Laboratory says all the evidence indicates that the variable best able to account for the wide range of radon levels (which can differ among houses by as much as a factor of 1,000) is not the ventilation rate but the entry rate: the rate at which the gas infiltrates the house. What drives the radon indoors is mostly the difference in air pressure between the exterior and the interior of a building; the pressure in the soil is often slightly greater than the pressure inside the building. Such a pressure gradient can be caused by outside wind currents and by an indoor temperature higher than the outdoor one.

Armed with this knowledge, investigators have come up with a number of possible ways to cut back the amount of radon flowing into a house. The measures range from sealing the understructure to reducing the air pressure in the underlying soil or gravel by means of suction fans. Yet before Government agencies can even consider an active role in promoting remedial measures, they need to know whether the problem is serious enough to warrant such action. Here uncertainty begins to cloud the issue.

The primary evidence establishing the relation between exposure to radon (including its radioactive-decay products) and lung cancer comes from epidemiological data on uranium miners who worked between 20 and 30 years ago and whose higher-than-normal exposure to the gas can only be estimated. To analyze such data it is necessary to identify and allow for other factors that could contribute to mortality or morbidity rates in the population studied. In some cases compensating for smoking and exposure to other carcinogens in the working environment leaves only a small percentage of lung cancers that can be attributed to radon exposure.

Further complicating these studies is evidence that the combined effect of smoking and breathing air that has a high radon concentration is considerably worse than the effect of either one alone. Eisenbud believes the role of such synergisms in the onset of lung cancer has been neglected but agrees that the epidemiological evidence is "a pretty good starting point."

Given the limitations of the epidemiological data, investigators have resorted to other studies. Results of experiments on animals seem to agree with the epidemiological data in showing a linear relation between radon dose and the incidence of lung cancer. If one starts with the excess-lung-cancer rates and high radon exposure levels derived from studies of miners and extrapolates linearly down to the radon levels found in houses, one gets a "ball park" estimate of about 10,000 radon-related cases of lung cancer per year in the U.S. This amounts to about 10 percent of the total U.S. incidence of lung cancer.

Yet some investigators question the validity of such extrapolation. Bernard L. Cohen, who heads a program at the University of Pittsburgh to measure residential radon levels, has reached a contradictory, albeit preliminary, conclusion. He maintains that in the range of radon levels found in most houses, which is considerably less than the range normally found in uranium mines, the linear dose response is not applicable. Cohen, taking issue with most radiobiologists, hypothesizes that below a certain level the incidence of lung cancer is largely independent of exposure to radon.

Nero doubts that it will ever be possible to demonstrate radon's effect on public health conclusively from epidemiological studies alone. He thinks that at the radon levels found indoors any excess cancers are likely to be buried among the numerous "back-

ground" cancers brought on by smoking and exposure to other environmental carcinogens. Nevertheless, he says radon is "easily the best-established case of environmental exposure that is probably causing cancer in the population—including chemical carcinogens that have already been brought under Government regulation."

The other area of uncertainty concerns the extent of the problem: the number of households affected by whatever radon concentration is taken to be "unsafe." A report from the General Accounting Office points out that high indoor radon levels have been found in nearly every state. Up to now, however, data on indoor levels have not been collected in a systematic way; the houses selected for sampling may not reflect the real distribution of radon levels in the U.S. To remedy the situation the U.S. Environmental Protection Agency is about to conduct what officials hope will be the most comprehensive survey of domestic radon levels yet undertaken. It will measure radon concentrations in a random nationwide sampling of residential buildings in the course of one year, applying consistent instrumentation techniques.

Nuclear Autumn

Those who first contemplated the likely climatic effects of nuclear war, commonly referred to as nuclear winter, have been criticized for manipulating uncertainties in such a way as to ultimately exaggerate the effects. Yet even the updated, "milder" nuclear-winter hypothesis recently proposed by Starley L. Thompson and Stephen H. Schneider of the National Center for Atmospheric Research does not appear to have departed significantly from the hypothesis originally presented in the so-called TTAPS report (named for the combined initals of the authors, Richard P. Turco, Owen B. Toon, Thomas P. Ackerman, James B. Pollack and Carl Sagan) published in *Science* some three years ago.

The TTAPS report held that smoke and dust generated by a large-scale nuclear war could block the sunlight that normally reaches the surface of the earth and thereby produce prolonged periods of darkness and abnormally low temperatures on the earth's landmasses. Such global environmental perturbations would devastate the plant life that constitutes the primary link in the human food cycle, jeopardizing global civilization and possibly the human species. The report further indicated that the transition from slight attenuation of sunlight to virtually total attenuation would take place within a narrow "threshold" range of values for the quantity of matter injected into the atmosphere, and that such a threshold quantity might be generated by the detonation of 100 megatons (the explosive power of 100 million tons of TNT) over 100 cities.

Thompson and Schneider, writing in *Foreign Affairs,* come to the conclusion that the climatic effects of nuclear war—in particular the worldwide decreases in temperature—would be less extreme than the TTAPS report predicts. Although they agree that climatic effects serious enough to disrupt the global environment are quite probable, they dismiss the notion of a quantifiable nuclear-winter threshold, and they relegate the possibility of human extinction to a "vanishingly low level of probability."

The TTAPS authors point out that the revised temperatures represent no significant contradiction. Although their report cited the possibility of a drop of from 20 to 40 degrees Celsius in their baseline case (a 5,000-megaton nuclear war) if the war were to take place in the summer, that prediction was based on a one-dimensional atmosphere model (in which atmospheric conditions are averaged across all latitudes and longitudes) applied to a hypothetical land-covered planet. Because oceans constitute vast reserves of heat, their exclusion from the model resulted in overestimation of the temperature drop. The TTAPS report itself suggested that including oceans in a more sophisticated model would reduce the predicted temperature decreases by about a factor of two.

Thompson and Schneider base their findings precisely on such a model: a three-dimensional model of the atmosphere that takes into account the ameliorating effect of oceans, the geographic distribution of smoke in the atmosphere and seasonal variations in climate. The estimates of temperature drops cited by Thompson and Schneider, which range from five to 15 degrees, seem to be in keeping with what the TTAPS report predicted.

Sagan argues that it is difficult to equate any given quantity of atmospheric smoke and dust with a specific number of nuclear explosions on a specific set of targets. He nonetheless maintains that it is meaningful to express the sunlight-attenuation threshold in terms of the number of nuclear weapons, albeit roughly, particularly when formulating nuclear policy or targeting doctrine. Sagan also maintains that in addition to the destructive "prompt" effects of nuclear explosions due to blast, radiation, fires and radioactive fallout (which could mean the death of perhaps a billion people worldwide, according to a World Health Organization report), the climatic effects would indeed bring about billions of other fatalities, pushing the total number of fatalities perilously close to the total number of people in the world.

The climatic effects have such apocalyptic dimensions because agricultural systems are exquisitely sensitive to the reduced temperatures, light levels and precipitation levels that could be expected even in a nuclear "autumn," particularly during the growing season. Such perturbations could result in low-yield harvests worldwide and perhaps the complete loss of crops in the Northern Hemisphere. Such crop loss, compounded by disruptions of trade, could in time result in a global famine of inconceivable magnitude.

Seeing the Forest

A recent report by the World Resources Institute estimates that 80,000 square kilometers of tropical forest—an area the size of Maine—are razed each year. The estimate, as the report acknowledges, is one among many. The rate at which tropical regions are being deforested is still vigorously debated, because reliable data are scarce. Governments may understate the problem for political reasons, and even disinterested surveys are limited in their accuracy by the vastness of the terrain to be covered. Satellite imagery may offer a solution: a new technique measures deforestation rates by comparing satellite images made at different times.

Among the many reasons for being concerned about trees falling in the Tropics, two deserve special mention. The first is that tropical forests are home to most terrestrial species of plants and animals; deforestation could therefore mean a mass extinction and a corresponding reduction in the earth's genetic diversity. The second is that tropical forests are also a sizable reservoir of carbon. Some workers, including George M. Woodwell of the Woods Hole Research Center, argue that deforestation is a major source of the increasing amount of carbon dioxide in the atmosphere. If that is the case, it may have a severe impact on the earth's climate through the greenhouse effect.

The technique for measuring deforestation devised by Woodwell (along with Richard A. Houghton and Thomas A. Stone of the Marine Biological Laboratory and Archibald B. Park, an independent consultant) is basically simple. Healthy vegetation absorbs red light and reflects near-infrared radiation. Bare ground is brighter than vegetation in the red spectral band and less bright in the near infrared. When a patch of forest is cut down, or when cleared land is cultivated, the change is visible in Landsat images. By superposing two images and subtracting the later one from the earlier one pixel by pixel, the change can be measured with considerable accuracy. (Each pixel, or picture element, in a Landsat image covers an area 59 meters by 79 meters.)

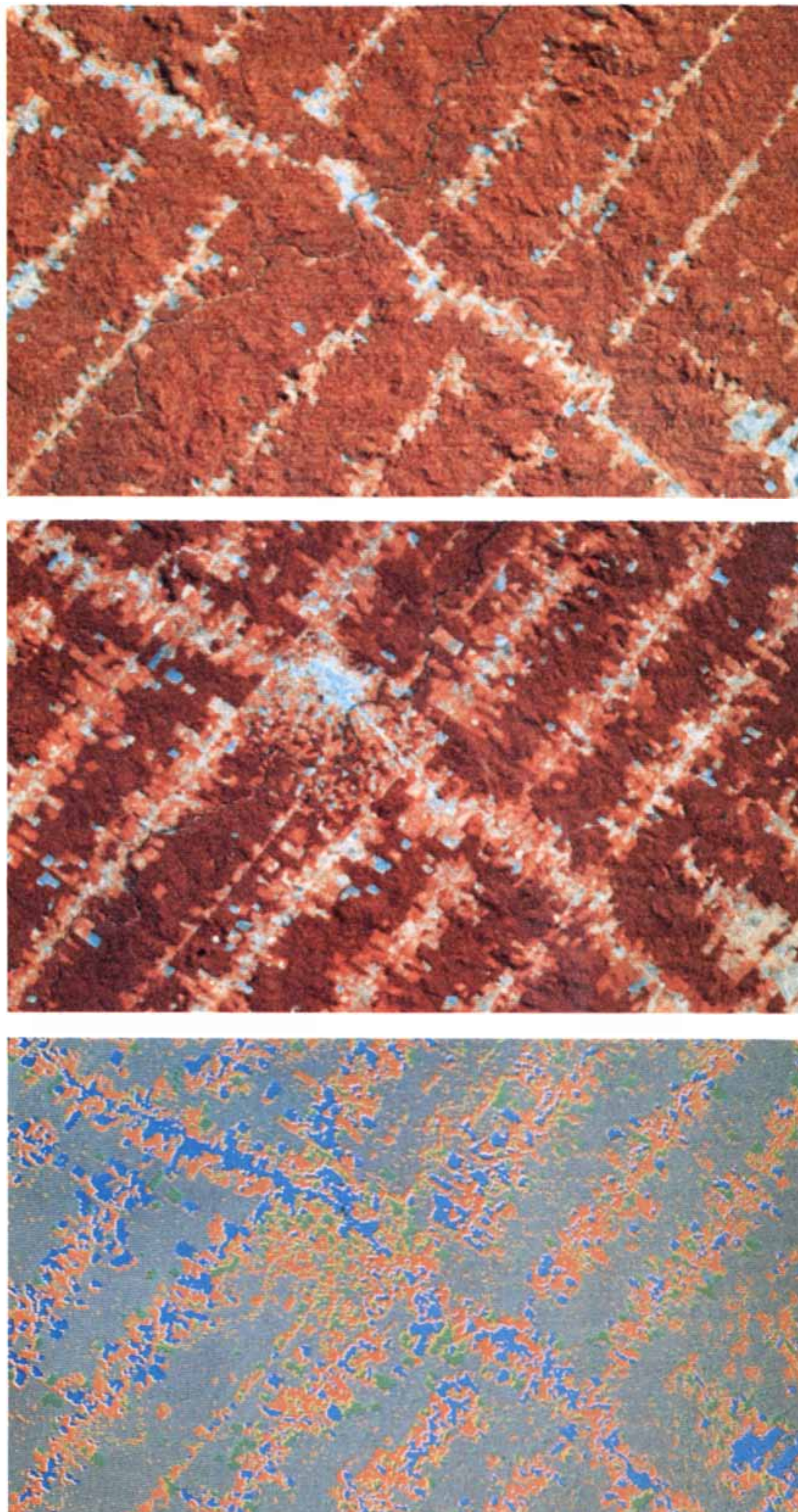
Woodwell and his colleagues have applied this technique to the Brazilian state of Rondônia in the southwestern Amazon basin. Until 1960 the region was largely untouched tropical forest, but in recent years it has undergone rapid development as the Brazilian government has given away 100-hectare land plots to settlers from the crowded southern part of the country. By 1982, according to the Woods Hole group's calculations, the settlers had cleared at least 11,400 square kilometers of forest.

Rondônia was just a test case for the technique; the ultimate goal is a global survey of deforestation. Landsat images covering the entire globe are available going back to 1972. Woodwell's group cut the cost of the Rondônia survey by combining selected Landsat data with images from a National Oceanic and Atmospheric Administration satellite that cover broader areas at lower resolution. The same approach, Woodwell says, could yield a global deforestation survey at a cost of only between $5 and $10 million.

Yellow Rain Stops Falling

Since 1981 the U.S. Department of State has asserted that the U.S.S.R. and its allies have waged chemical warfare in Southeast Asia. The department's case rests mainly on traces of a toxin (T-2, a trichothecene produced by the mold *Fusarium*) detected in yellow spots on leaves and rocks collected near battle sites in Kampuchea and Laos, and on the finding of the toxin in blood and other specimens taken from 20 people said to have been victims of chemical attacks. Several investigators, including Matthew Meselson of Harvard University, challenge the State Department's evidence. They say the "yellow rain" consists primarily of digested pollen and is in fact mass defecation by bees. Recent reports from Canada and the U.K. would appear to have left the State Department's case in disarray.

The Canadian Defense Research Establishment studied blood samples from 280 people in Thailand, 270 of whom had not been in battle areas. Five of the 270 had trichothecenes in



LANDSAT IMAGES illustrate the progress of deforestation between 1976 (*top*) and 1981 (*middle*) in Rondônia, a Brazilian state on the Bolivian border. The images show a region 30 kilometers wide. The red areas are tropical forest; the blue-and-white diagonal lines are roads along which settlers have cleared plots of land. The extent of deforestation can be measured by subtracting the spectral data of the later image from those of the earlier image. On the subtraction image (*bottom*) the green and orange areas are clearings where bare ground has been exposed, and the blue areas are cleared land that has been converted to agriculture. The Landsat data were provided by the INPE, the Brazilian space agency.

their blood; the 10 alleged victims of chemical warfare had none. The Canadians noted that the blood levels of trichothecene are comparable to those found in food infected with *Fusarium* fungus. The Canadians also examined plant samples and found that the yellow spots they bore consisted mainly of pollen. On a plastic bag presented by a resident of a village that was said to have been attacked in 1982 the Canadians did find trichothecenes. The investigators declined to draw a conclusion about the significance of the finding because, they stated, "the authenticity of the source of the bag cannot be verified."

In the U.K., John Stanley, minister of state for the armed forces, told the House of Commons that the British Chemical Defence Research Establishment "has analysed a number of environmental and biomedical samples from alleged chemical warfare attacks in south-east Asia to see whether traces of trichothecene mycotoxins could be detected." The tests, he said, "have yielded negative results."

In the light of these findings, Meselson comments, "the yellow-rain allegation has now crumbled away to almost nothing."

Home Guard

The 120 million privately owned firearms in the U.S. are distributed so widely that every second household has at least one. Nearly half of the weapons are handguns. The commonest reason given for keeping a handgun in the house is protection against intruders. A study done in King County, Wash., shows that such a strategy is likely to backfire.

The investigators found that for each instance in which a gun in a home was actually involved in a "self-protection" homicide there were 1.3 accidental gunshot deaths, 4.6 criminal homicides (mostly of relatives or friends) and 37 suicides. The study was made by Arthur L. Kellermann, who was then at the University of Washington School of Medicine, and Donald T. Reay, the King County medical examiner; it is reported in *The New England Journal of Medicine.*

To gather their data the investigators reviewed all the deaths by gunshot that had taken place during a period of six years (1978–83) in King County. (The 1980 census recorded a county population of 1,270,000, which was 88.4 percent white and 92 percent urban; the county includes Seattle.) There were 743 such deaths, 398 (54 percent) of which occurred in the residence where the firearm was kept. It turned out that only two of the 398 victims were intruders shot while they were attempting entry. Excluding suicides, guns in homes were involved in the death of a member of the household 18 times more often than they were in the death of a stranger; such weapons brought about the death of a friend or an acquaintance 12 times more than they did the death of a stranger. More than 80 percent of the homicides occurred during arguments or altercations.

Kellermann and Reay conclude that "keeping firearms in the home carries associated risks," including accidental injury or death from gunshot wounds, homicide during domestic quarrels and "the ready availability of an immediate, highly lethal means of suicide." In the light of their findings, the investigators say, "it may reasonably be asked whether keeping firearms in the home increases a family's protection or places [the family] in greater danger."

ASTROPHYSICS

Far Out

A team of British and American astronomers has found the most distant object in the universe ever observed. They discovered a quasar—a galaxy with a compact core that might be a black hole—at a distance of 12.4 billion light-years from the earth. The object, which is about 10 million light-years beyond the most distant quasar previously found, occupies a position 90 percent of the way to the edge of the observable universe.

Cyril Hazard of the University of Pittsburgh, R. G. McMahon of the University of Cambridge and Wallace L. W. Sargent of the California Institute of Technology describe in *Nature* how they measured the distance to the quasar. Their method is based on the prevailing model of the universe: the big-bang theory, which holds that all galaxies are rushing uniformly away from one another as the result of a primordial explosion that took place some 15 to 20 billion years ago. According to that theory, the farther a stellar object is from the earth, the higher is its speed. In other words, distance is proportional to velocity.

Although direct measurements of distance are difficult to make, estimates of velocity are relatively easy. The wavelength of light emitted by a receding object grows longer by an amount proportional to the speed of the object, just as the pitch of a receding siren appears to fall. "In a sense, light expands as the universe expands," comments Sargent. Since the longer wavelengths shift the light toward the red end of the optical spectrum, the phenomenon is called red shift. By measuring the amount of red shift of a stellar object one can determine its speed and distance.

To measure the distance of the quasar, dubbed QSO1208 + 1011, Sargent and his colleagues employed a special photographic emulsion sensitive to infrared light. Working with the five-meter Hale telescope at Palomar Observatory, they determined that the quasar has a red shift of 3.8, which is .02 unit greater than the second most distant quasar, PKS2000 − 330. Using the same equipment, the astronomers expect that they will be able to detect yet more distant quasars with red shifts as high as 4.2.

ANTHROPOLOGY

Oral History

When did the first Americans arrive and where did they come from? Christy G. Turner II of Arizona State University has sought to solve that long-standing puzzle with a new technique called dentochronology: dating the genetic separation of human populations on the basis of the detailed form of their teeth. The method depends on statistical analysis of many thousands of measurements. It is one of a number of new anthropological techniques that have been made feasible by the computer.

Turner reports in *Science* that the first step was working out a battery of 34 tooth traits (such as the number of cusps in molars) that are preserved well in skeletal populations, are easy to observe and seem to be heritable. He then gathered data on these traits from more than 10,000 human skeletal remains. For each pair of populations whose divergence was to be studied, the data were reduced to a single number: the multivariate mean measure of divergence, or MMD.

Because the MMD is a measure of the relative similarity between two populations, dentochronology yields information about their lineage. For example, it is generally agreed that New World Indians came from Asia; because the MMD for the New World Indians/northeastern Asians pair is smaller than the MMD for the Indians/southeastern Asians pair, Turner was able to show more specifically that the parent population inhabited northern China and Mongolia.

The constant evolution of tooth characteristics in all populations results in a match between the magnitude of a pair's MMD value and the length of time independent estimates suggest the groups have been separated. American Indians, for example, have been separated for a longer time from their parent population (MMD .14) than Pacific Polynesians have been separated from theirs (MMD .04).

In order to determine absolute dates for these genetic separations it is necessary to find a means of calibration: an independent clock. Turner does that by calculating the average rate at which tooth characteristics have evolved everywhere in the world except in the two populations being compared. He then divides that world rate of change into the MMD for the two populations—the measure of how different their tooth traits are. The result is the number of years the pair must have evolved separately.

Turner's dentochronological study has led him to report in *National Geographic Research* that New World Indians can be divided into three population clusters. All three groups originated in northern China in a single population that came to differ dentally in the course of time. They left China about 20,000 years ago as geographically isolated bands of hunters and fishermen and made their separate ways to Alaska across the Bering Strait land bridge. Driven by environmental changes, some groups continued to move southward, with the first Paleo-Indians appearing in the southwestern U.S. some 12,000 years ago—a date compatible with many independent estimates.

A number of findings suggest earlier dates for New World settlement. Most recently a French team working in a Brazilian cave has reported in *Nature* that radiocarbon dating of charcoal associated with stone artifacts indicates that humans were living in South America at least 32,000 years ago. Turner questions the source of both the carbon and the artifacts. Before accepting the early date he would like to see an explanation of how the material in the French study got into the cave. "Cave archaeology is really tricky stuff," he says. The artifacts could be stream-battered rocks, the charcoal the result of a forest fire. The charcoal may indeed be 32,000 years old, in other words, but Turner finds no proof that it and the stones reached the cave by human hand.

Bones of Contention

Whether members of human societies have ever made it a regular practice to eat other humans as food has been a subject of much dispute among anthropologists. Most reports of such "dietary" cannibalism among primitive peoples have been dismissed as unreliable and vague. Archaeological proof that prehistoric peoples engaged in the practice has been equally unconvincing. Recent excavations in Fontbrégoua Cave in southeastern France have now provided the strongest evidence to date: a socially sanctioned cannibalism seems indeed to have been practiced by the cave's Stone Age inhabitants.

The evidence uncovered by an international team of anthropologists and archaeologists led by Paola Villa of the University of Colorado at Boulder consists of clusters of animal and human bones that appear to have been chipped, cut and broken in much the same way. Because the animals were presumably butchered and processed for food, Villa and her colleagues infer that the humans met a fate similar to that of the animals.

As Villa's group reports in *Science,* the human and animal bones generally are from comparable anatomical parts (implying a similar culling and discarding procedure); they display an abundance of marks that could only have been made by stone tools and not teeth (implying that the flesh had been removed from the bones in the same way, and that animal scavengers had not chewed the bones), and long bones of both types had been broken (implying that the bone marrow in them was extracted).

Although this is not the first time claims of prehistoric cannibalism have been made on the ground that animal and human remains appear to have been similarly treated, earlier findings were often in dubious archaeological condition. Moreover, it was difficult to prove that the marks on human bones were not the traces of some other activity unrelated to food processing. For example, it has been suggested that flesh may have been removed from human bones as part of a funerary rite. The bones uncovered by Villa and her co-workers, however, have been well preserved and apparently have lain undisturbed for nearly 6,000 years. In addition the investigators think the extraction of marrow is unlikely to have been a step in preparing a corpse for burial.

Anthropologists who have dimissed the notion of social cannibalism, such as William Arens of the State University of New York at Stony Brook, grant that Villa's research was carefully done but stress that other interpretations of the evidence are possible. Arens points out that the case for cannibalism would have been much stronger if the bones had shown signs of cooking. What speaks most against Villa's interpretation, however, is the lack of corroborating evidence from other archaeological sites. Until further excavations uncover other such bone clusters, Arens says, the Fontbrégoua findings should be regarded as an unusual case.

MEDICINE

Habitually Unhealthy

A study done some years ago among residents of Alameda County in California established that seven habits appear to be directly associated with good health and reduced mortality: not smoking, not drinking too much, maintaining the right weight, exercising, sleeping seven to eight hours a night, eating breakfast and avoiding snacks. Americans have not been accepting the implied advice. A countrywide survey done by the National Center for Health Statistics has shown that they drank more, weighed more, exercised less and slept less in 1983 than they did in 1977.

A 10 percent decline in smoking was the only significant favorable trend. Nearly all groups surveyed smoked less than they had in 1977, although the reduction was less for women than for men and was strongly correlated with income and education.

The most unfavorable trend was a 28 percent increase in the proportion of people who reported having taken five or more drinks on any one day in the past year. In 1983 alcohol consumption was higher for higher-income groups and for people with higher education.

About 16 percent of the population surveyed in 1983 reported that they weighed at least 30 percent more than what is considered to be their desirable weight. This represented a 10 percent increase in the prevalence of obesity. People in the lowest income and education groups were twice as likely to be overweight as those at the highest levels.

Surprisingly, there was a 12 percent increase in the proportion of people who think they are less physically active than their contemporaries. The greatest trends toward inactivity were reported by women, by people from 20 to 34 years old and by people in high-income and high-education groups. The report cautions that people who were surveyed may have been influenced by the increasing number of joggers on the streets and the burgeoning of health clubs.

Insufficient sleep is also on the increase, with 17 percent more women and 7 percent more men reporting less than six hours a night. Sleeping habits are also related to income and education: fewer people in the higher brackets reported that they get less than six hours' sleep a night.

BIOLOGY

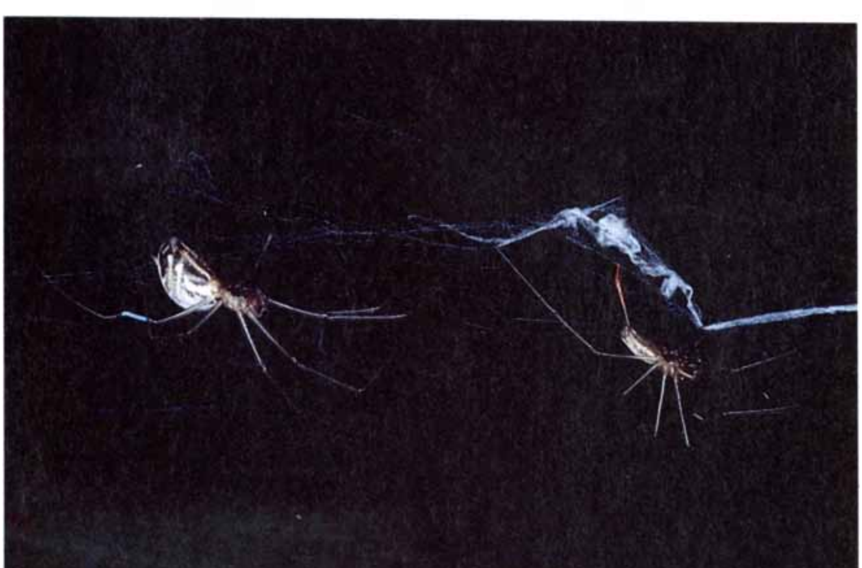
Tangled Web

Before a male Sierra dome spider mates he bundles the female's delicate dome-shaped web into a tight little ball, destroying it. Paul J. Watson of Cornell University has found the reason for this seemingly wanton vandalism: the male thereby prevents interruption by other males during the lengthy mating session. Watson's discovery stemmed from his observation that a mature female, having remained virgin for from seven to 10 days, laces her web with a volatile pheromone, a hormone that attracts males. By reducing the web to a small ball the first male to be attracted significantly decreases the pheromone's evaporation into the ambient air and so lowers the likelihood that other males will come on the scene.

The Sierra dome spider lives in the mountainous areas of western North America. Mature males leave their webs and live by scavenging from those of females. They compete strenuously for the chance to mate with the relatively scarce mature, virgin females because ordinarily only the first male's sperm fertilize her eggs.

In low-density populations an unmated mature female needs to advertise her presence. Perfuming the web with the pheromone, Watson writes in *Science,* increases the rate of male visitation by an order of magnitude. The first thing a male does on entering the web is therefore to reduce it. Only then does the five-hour copulation begin.

What is good for the male is bad for the female. She has to rebuild her destroyed web, which represents a nonrecoverable protein investment and is her only means of foraging and defense. The female also loses what would be implicit in an open, fair fight among male spiders: the chance to mate with the strongest male available. Such conflict between the sexes is common in animals, but the source of a female's signals is not usually so vulnerable. If the pheromone were released from the female's body, for instance, the most the male could do would be to mask it with his own signal, a strategy that is indeed adopted by some other animals.



MALE SIERRA DOME SPIDER ***(right)*** **reduces a female spider's web; the female waits** ***(left)*****. Destruction of the web makes it unlikely that other males will interrupt copulation.**

Genetics: Forward in Reverse

The great contribution of molecular biology to the understanding of inherited diseases has come largely from studies that began with defective proteins and moved from there to the corresponding genes. Some inherited disorders, however, are not well enough understood for a key protein to have been identified. For those syndromes another approach has long been envisioned: find the gene first and use it to identify the defective protein. The first successful application of "reverse genetics" has been reported by Stuart H. Orkin of Children's Hospital in Boston and his co-workers. Without knowing the relevant protein, they have isolated the gene that causes chronic granulomatous disease (CGD).

CGD is an uncommon syndrome that disarms the immune system by crippling the phagocytes, white blood cells that kill microorganisms. In response to recurrent bacterial infections the crippled phagocytes clump together in tumorlike masses called granulomas, which give the disease its name. The biochemistry of CGD has resisted full understanding since the disease was identified two decades ago. It is clear, however, that the genetic defect somehow subverts the NADPH-oxidase system, a group of linked enzymes that help phagocytes to kill invading microorganisms.

Because it is not clear which of the proteins of the NADPH-oxidase system cause CGD, Orkin's group resorted to the reverse-genetics approach. Their first job was to locate the genetic defect approximately. To simplify the task they worked with a form of CGD that is X-linked, or carried on the X chromosome (of which females have two and males one). Preliminary work narrowed the location of the defect to a region of several hundred kilobases (thousands of nucleotide bases). Since such a region generally includes many genes, the next step was to winnow out the CGD gene.

Orkin and his co-workers reasoned that if a gene in the suspect region were expressed in phagocytes but not in other kinds of cells, it would be a likely candidate for the cause of CGD. To obtain the set of genes that are specifically expressed in phagocytes, they performed an elegant sleight called subtractive hybridization. The messenger RNA's (mRNA's) from a tumor-cell line derived from phagocytes were converted into DNA sequences, which were mixed with the mRNA's from a nonphagocytic line. Sequences the two cell lines have in common hybridized (bound) to each other, leaving as a remainder the sequences specific to phagocytes.

It was then necessary to find out whether any of the phagocyte genes falls in the region where the CGD gene was thought to lie. By coincidence Orkin's colleague Louis M. Kunkel had been working on another inherited disorder—Duchenne's muscular dystrophy—whose gene is on the X chromosome somewhere near the CGD gene. Kunkel had prepared a set of DNA fragments covering part of the relevant stretch of the X chromosome. When those fragments were paired with the genes specific to phagocytes, there was one match: a gene, corresponding to a protein of about 500 amino acids, that Orkin's group has dubbed X-CGD.

To show that X-CGD causes the disease, the gene and its protein were examined in CGD patients. All showed an abnormality. Either they lacked X-CGD altogether or they had small aberrations that could lead to the production of a defective protein. Those findings clinched the identification of X-CGD as the cause of the inherited disorder. Will reverse genetics now become a standard approach? According to Orkin, it will probably be used for studying diseases whose biochemistry is "intractable, inaccessible, confusing or complex." Prime candidates are Duchenne's muscular dystrophy, cystic fibrosis and Huntington's disease.

TECHNOLOGY

Early Harvest

Three pioneering products of biotechnology have been approved by the U.S. Food and Drug Administration for human use. They are a genetically engineered vaccine against hepatitis *B*, a recombinant-DNA alpha-interferon for treating a rare leukemia and a monoclonal antibody designed to prevent the rejection of a transplanted kidney.

Hepatitis *B* is a potentially serious liver infection caused by a virus transmitted primarily through the blood. The groups at highest risk are health workers, drug users and homosexuals. A vaccine prepared from a viral protein found in the blood of people who have recovered from the disease is available and has been shown to be safe, but it has not been widely accepted by people at risk, mainly because they think it may be contaminated with infectious hepatitis virus, the AIDS virus or agents of other diseases.

That objection, however unfounded, cannot be raised against the new vaccine, which is not derived from human blood. Instead a stretch of DNA, the isolated gene for a viral surface protein, is inserted, along with the proper regulatory sequences, into yeast cells in culture. The cells manufacture the protein, which then serves as the antigenic agent in the vaccine, stimulating the injected person's immune system to synthesize antibodies against the virus. The vaccine was developed by the Chiron Corporation and will be produced and marketed by Merck Sharp & Dohme.

Like the vaccine antigen, the alpha-interferon molecule is a genetically engineered protein. It has been approved specifically for the treatment of hairy-cell leukemia, a rare cancer that currently affects some 1,000 Americans. Hairy cells, named for their threadlike projections, are immune-system cells, called *B* lymphocytes, that have somehow been triggered to multiply. Symptoms include an enlarged spleen and inhibited development of both red and white blood cells. The result is extreme weakness and susceptibility to infection, and death within several years after onset.

In patients treated with alpha-interferon the cancerous hairy cells disappear from the blood and bone marrow, and normal blood cells proliferate to acceptable levels. More than 90 percent of the hairy-cell-leukemia patients treated with the interferon have survived for two years, and in from 75 to 90 percent of them the disease appears to be in partial or complete remission.

Interferon, a natural cell product, has long been considered a potential antiviral and perhaps anticancer agent, but the extreme difficulty of purifying the natural product from white cells made it difficult to accumulate enough of it even for effective clinical trials. The approved alpha-interferon is the first interferon made available for cancer therapy. It is produced not from white cells but by recombinant-DNA technology. The gene encoding alpha-interferon is isolated and inserted into the bacterium *Escherichia coli*, which synthesizes large amounts of the interferon. The FDA approved two alpha-interferons. One was developed by Biogen Research and has been licensed to the Schering-Plough Corporation for production and marketing; the other was developed, with collaboration from Genentech Inc., by Hoffmann–La Roche, Inc., which makes and markets the product.

A different kind of biotechnology is exploited to prepare a monoclonal antibody. Two types of mouse cell, one that makes the desired antibody and one that can be maintained indefinitely in culture, are fused. The resulting "hybridoma" cell gives rise to a clone (a line of identical cells) that can be grown in culture indefinitely to make large amounts of a single "monoclonal" antibody.

The antibody approved by the FDA is the first to be made available for treatment, rather than diagnosis, in humans. It is directed against *T* cells, different types of which have various roles in the immune system. More than half of the 7,000 or so kidneys transplanted in the U.S. in 1984 were perceived as being foreign by the recipient's *T* cells and were attacked by the immune system in an acute rejection episode. Various drugs can mitigate the rejection and often save the transplant, but in the process they suppress the entire immune system, leaving the patient open to infections.

The newly approved antibody was developed and is to be marketed by the Ortho Pharmaceutical Corporation. Binding to the surface of *T* cells, it proves to be a deadly cloak. The immune-system cells called monocytes recognize a region of the antibody as being foreign and engulf all *T* cells within an hour; in the absence of *T* cells the attack on the transplant is halted. In one large clinical trial the antibody reversed initial rejection episodes in 94 percent of the treated patients compared with a 75 percent success rate for conventional antirejection drugs. In a trial involving patients for whom other treatments had failed, the antibody saved two-thirds of the transplanted kidneys.

A short course of the antibody treatment is effective, after which the *T* cells reappear within 48 hours. Moreover, the fact that the monoclonal antibody does not affect other immune-system cells means that the immune response as a whole is not even temporarily disabled.

A Question of Balance

Mass is a fundamental property of all matter. The mass of an object is usually determined by weighing the object in gravity with a spring balance or an analytical balance; to determine mass one simply divides the measured weight by the strength of gravity. Recently a device has been developed that measures the mass of tiny particles without relying on gravity. The device is particularly well suited to measuring the mass of gas particles.

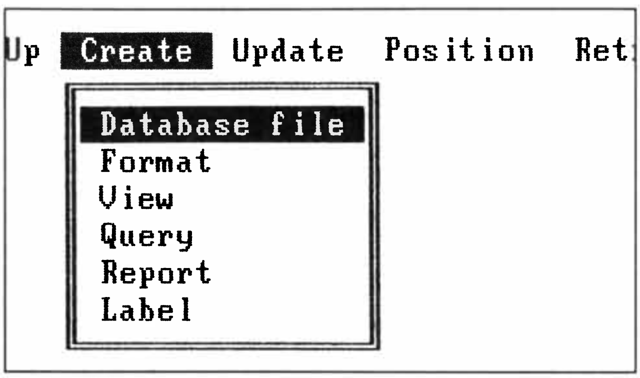
The instrument consists of a tapered, hollow tube made from a special glass. The wide end of the tube is attached to a vacuum pump; a filter sits on the narrow end. In order to measure the mass of particles the narrow end of the tube is made to vibrate, like a tuning fork, at a known frequency. The vacuum pump is turned on, sucking particles into the filter, where they lodge. As the sample accumulates it damps the vibration. Having determined the change in frequency precisely, one can calculate the mass of the collected particles.

The device, which was developed by Harvey Patashnick and Georg Rupprecht and is being marketed by their company, Rupprecht & Patashnick Co., Inc., offers a promising array of applications. It has already monitored the exhaust from both diesel and turbine engines and smoke emitted from burning materials. In the future it could measure the levels of carbon released by coal-burning power plants and determine with unprecedented accuracy the mass of particles in the tail of a comet.

The database used now be used

Introducing dBASE III® PLUS.

The PLUS stands for all the improvements we've made to the world's number one selling database management software.



The Assistant helps beginning users accomplish day-to-day data management tasks without programming.

Mind you, dBASE III PLUS still has the powerful dBASE programming language, dot prompt, and all the features that have made dBASE III the standard of the industry.

We've simply raised the standard.

And just as dBASE III introduced more power to the people, our new dBASE III PLUS introduces more people to the power.

People who aren't all that crazy about programming, for example.

The Assistant feature in dBASE III PLUS now provides them with new easy-to-use pull-down menus for creating, using and modifying multiple databases.

So now anyone who can manage a simple cursor can manage day-to-day data management tasks. Without programming.

And by using our new Screen Painter, anyone can create custom screens. Without programming.

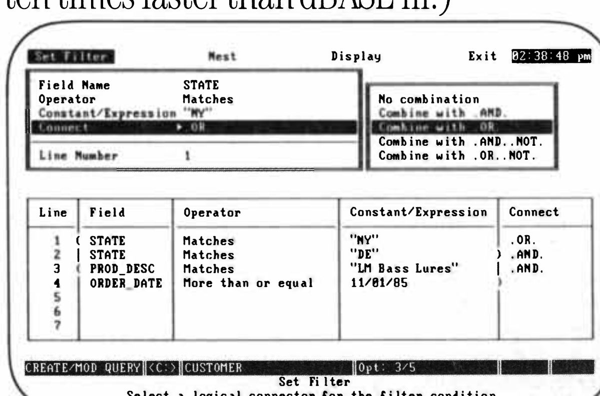
Or using View, access related information in several databases at one time. Without programming.

With Advanced Query System, another new non-programming feature, any user can build complex queries just by selecting from the dBASE III PLUS pull-down menus.

For rapidly creating entire programs, there's even a new Applications Generator.

And for all those who wish to learn to program, the Assistant can be of further assistance. By teaching you programming commands as you go along. Without disrupting your work flow.

These are only a few of the dBASE III PLUS features that can help new users quickly get up to speed. And experienced users quickly increase their speed. (Sorting, for example, is up to two times faster and indexing up to ten times faster than dBASE III.)



Advanced Query System lets you set up and answer complex queries without programming.

The Blood-Brain Barrier

Brain capillaries are unlike those of other organs. Their special properties enable them to serve as stringent gatekeepers between blood and brain. Recent work shows how that feat is accomplished

by Gary W. Goldstein and A. Lorris Betz

In human beings and other complex organisms, life itself depends on homeostasis, the maintenance of a constant internal milieu. Nowhere is this dependence more pronounced than in the brain. Elsewhere in the body the extracellular concentrations of hormones, amino acids and ions such as potassium undergo frequent small fluctuations, particularly after meals or bouts of exercise. If the brain were exposed to such fluctuations, the result might be uncontrolled nervous activity, because some hormones and amino acids serve as neurotransmitters and potassium ion influences the threshold for the firing of nerve cells. Hence the brain must be kept rigorously isolated from transient changes in the composition of the blood.

How is this feat accomplished? The answer lies in the unique structure of the capillaries that supply blood to the tissues of the brain. The cells of brain capillaries, unlike those of other capillaries, form a continuous wall that prevents many substances from entering the brain. The uninterrupted capillary wall provides the basis of the blood-brain barrier, whose existence was first demonstrated conclusively in the 1960's. By cordoning off the brain from the other tissues, the barrier serves a critical function. Yet if the isolation were complete, the brain would die for lack of nourishment. Fortunately the essential nutrients traverse the blood-brain barrier easily, helped across by transport systems that recognize specific molecules and carry them into the brain.

Much of the recent work on the blood-brain barrier, including our own, has concentrated on the transport systems. It has become clear that there are several different types of transporter, each of which has a specific function. The transport systems not only bring nutrients into the brain but also pump surplus substances out, thereby helping to maintain a constant environment for the neurons. By applying newly developed methods for isolating brain capillaries, we have identified many of the transporters in the capillary wall. Our work, along with that of other investigators, is beginning to show how the blood-brain barrier creates the unchanging environment the brain needs. Because the barrier must be pierced in order to treat diseases that affect the brain, the recent findings have considerable significance for medicine as well as for more fundamental studies of biology.

The Concept of a Barrier

What is now understood about the blood-brain barrier represents the culmination of a long line of research. The concept of the barrier arose in the late 19th century when the German bacteriologist Paul Ehrlich observed that certain dyes administered intravenously to small animals stained all the organs except the brain. Ehrlich's interpretation of his result was that the brain had a lower affinity for the dye than the other tissues. In 1913 Edwin E. Goldmann, who had studied under Ehrlich, proved his teacher wrong. Goldmann injected the dye trypan blue directly into the cerebrospinal fluid of rabbits and dogs. Administered by this route, the dye readily stained the entire brain but did not enter the bloodstream to stain the other internal organs. Thus Goldmann showed that the central nervous system is separated from the blood by a barrier of some kind.

Goldmann's finding helped physiologists to make sense of some interesting but puzzling results that had been obtained a few years before by other investigators. In 1898 A. Biedl and R. Kraus had found that bile acids, which are not neurotoxic when injected into a vein, caused seizures and coma when injected directly into the brain. Two years later M. Lewandowsky obtained similar results using sodium ferrocyanide. In explaining these findings the three investigators displayed considerable foresight, since they attributed their results to special permeability properties of the small blood vessels in the brain.

The hypothesis that brain capillaries provide the anatomical basis of the barrier postulated by Goldmann could not be confirmed immediately. In the early 1900's no method was available for examining the fine ultrastructure of the blood vessels. By the 1950's such a tool had become available in the electron microscope, and it was soon turned to the capillaries of the brain. Electron microscopy revealed that the endothelial cells forming the tube of a capillary are joined by what are known as continuous tight junctions. Ordinarily each animal cell is encapsulated in its own outer membrane, which has two leaflets, or layers. At a tight junction the outer leaflets of two adjoining cells merge, so that the cells are physically joined.

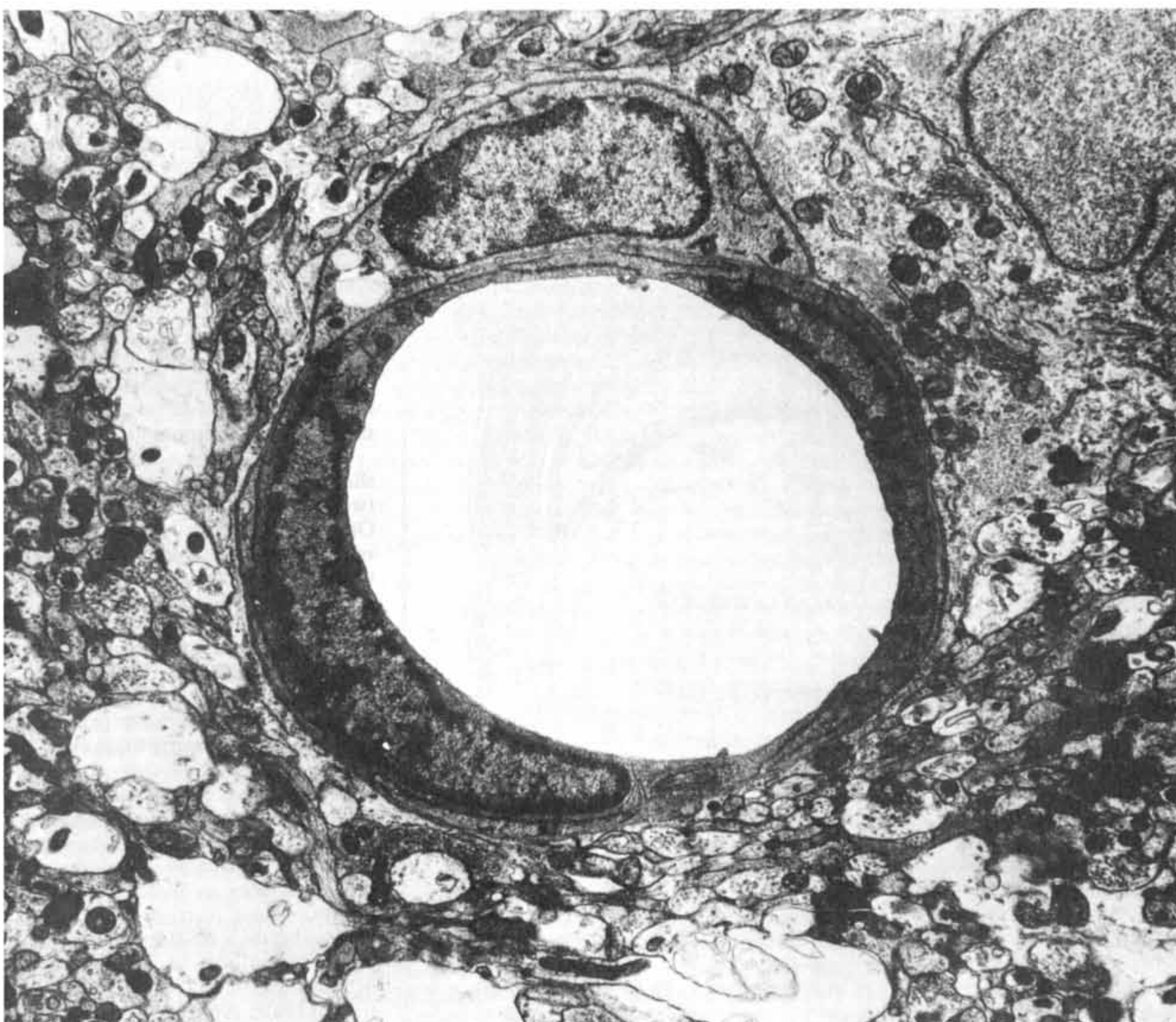
Continuous tight junctions are not the only feature that makes the blood vessels of the brain distinctive. In some organs the cells of the endothelial layer have gaps or channels running all the way through the layer; in the brain such channels are lacking. In addition the brain capillaries are almost completely surrounded by processes (long, slender extensions) of the brain cells known as astrocytes. Astrocytes are members of the largest class of brain cells—the glial cells—which, among other functions, form the myelin that sheathes some neurons. Because the astrocytes hold the capillaries in their grasp, some of the early investigators thought these cells must constitute the blood-brain barrier. That hypothesis has been shown to be false, but the exact function of the astrocytes is still being debated.

As information about the ultrastructure of brain capillaries accumulated it became clear that microvessels with these distinctive features are not present everywhere in the brain. In six or seven small regions gaps or channels extend through the endothelium, enabling substances in the blood to reach the extracellular fluid of the brain or the neurons themselves. Among the areas without a blood-brain barrier are the pituitary gland, the pineal gland and some parts of the hypothalamus. It would seem that the lack of a barrier in these zones opens the way for hormones circulating in the blood to reach secretory neurons in the brain, thereby closing the loop that underlies the feedback regulation of neuroendocrine systems.

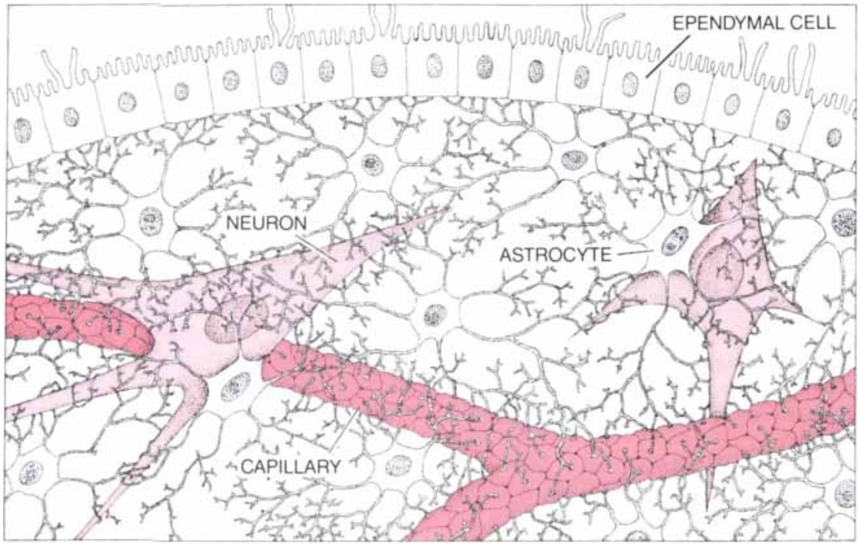
Electron microscopy done in the 1950's provided a much more detailed picture of the structure of brain capillaries than had ever before been available. In the 1960's the development of tracer molecules that could be visualized in the electron microscope made it possible to correlate the newly acquired structural detail with the function of the blood-brain barrier. One such tracer is the enzyme horseradish peroxidase, which is similar in size to proteins normally present in the blood. When the peroxidase is exposed to a specific reagent, electron-dense fragments of the protein form a dark stain that serves to mark the location of the tracer.

Thomas S. Reese and Morris J. Karnovsky of the Harvard Medical School employed horseradish peroxidase in a modern version of the earlier experiments with trypan blue. They injected it into the bloodstream, waited and then examined the brain by electron microscopy. In most organs peroxidase easily penetrates the capillary wall through channels and gaps between endothelial cells or is engulfed in small capsules called pinocytotic vesicles, which are taken up into the endothelial cells themselves. In the brain, on the other hand, the peroxidase was stopped by the tight junctions between endothelial cells and little of the protein was carried into the endothelium in vesicles.

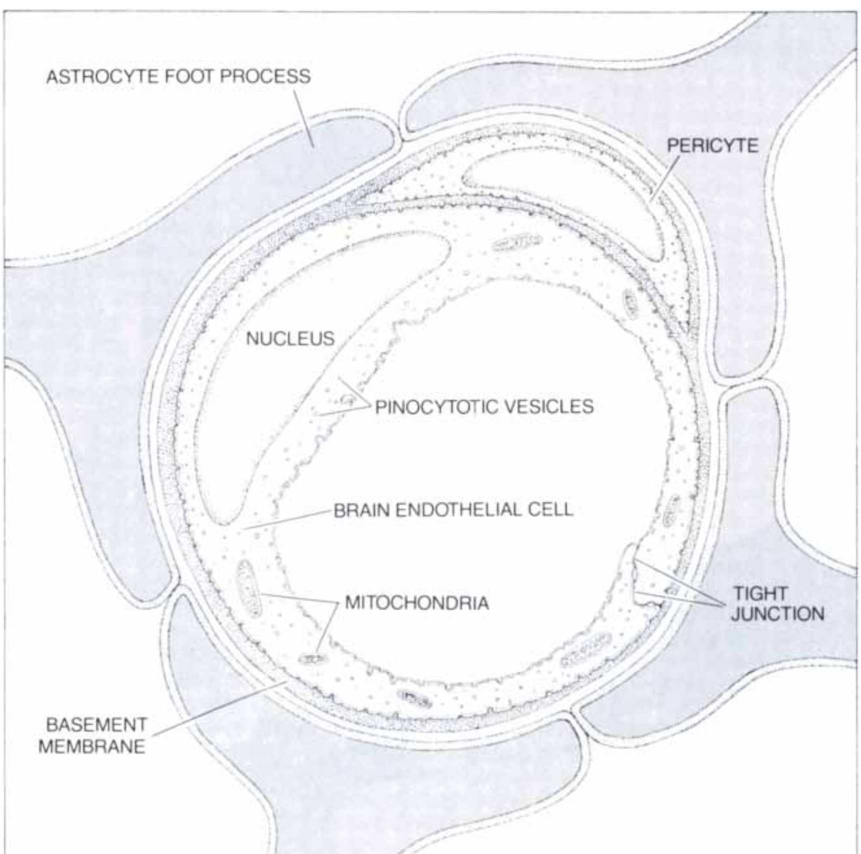
Not long after this dramatic demonstration Reese moved to the National



BRAIN CAPILLARY shown in cross section reveals the anatomic structures that provide the basis for the blood-brain barrier. The two large, crescent-shaped forms that make up the wall of the capillary are endothelial cells. They are joined, at the top and slightly to the right of the bottom, by thin overlapping segments. At these overlaps the outer membranes of the endothelial cells partially fuse to form "tight junctions" that prevent substances in the blood from diffusing freely into the brain. The cell attached to the top of the capillary is a pericyte; its function is not yet completely understood. The capillary, from a rat cerebral cortex, has been enlarged some 10,000 diameters in this micrograph made by P. A. Cancilla of the University of California at Los Angeles School of Medicine.



ASTROCYTES are cells whose long processes (extensions) make contact with several other types of brain cells, as is shown schematically in this drawing. The astrocyte processes touch neurons and the ependymal cells that line the ventricles, which are spaces at the center of the brain. In addition each brain capillary is typically in contact with several astrocytes. Astrocytes are members of the largest class of brain cells: the glial cells. Although their function is not yet fully understood, they may influence capillary permeability.



ASTROCYTE FOOT PROCESSES almost completely surround the brain capillary. Because of this relation it was once thought that the astrocytes form the blood-brain barrier. It is now known that the endothelial cells constitute the barrier. Endothelial cells selectively transport nutrients into the brain, and their many mitochondria probably provide energy for transport. The endothelial cells of the brain have few pinocytotic vesicles. In other organs such vesicles may provide relatively unselective transport across the capillary wall.

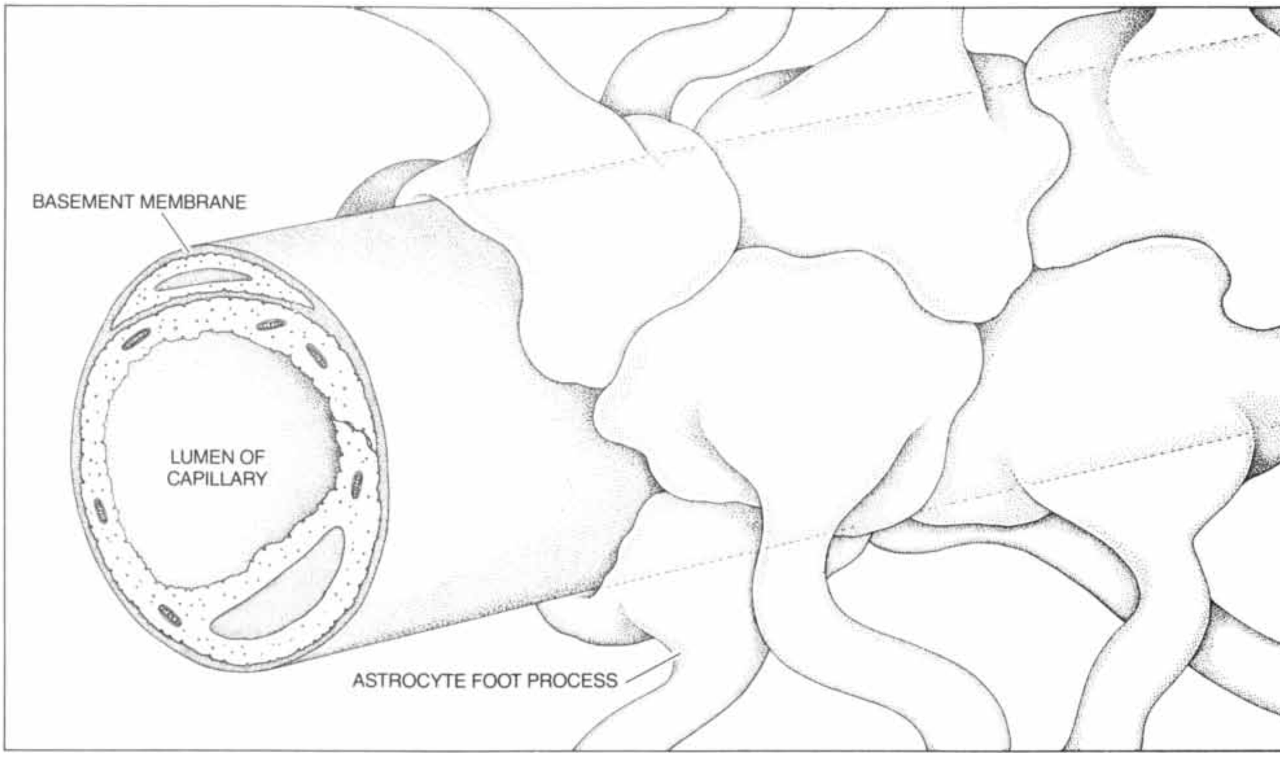
Institute of Neurological and Communicative Diseases and Stroke, where he was joined by Milton W. Brightman. Together they did an experiment that was in a sense the obverse of the one Reese had done at Harvard with Karnovsky. Horseradish peroxidase was injected into one of the cerebral ventricles, the cavities that lie at the center of the brain. The peroxidase permeated the cerebrospinal fluid in the ventricles and flowed into the extracellular spaces of the brain. It was prevented from leaving the brain, however, by the tightly joined endothelial cells. This study showed that the astrocytes did not act as a barrier to the outward movement of the protein. Furthermore, together with the work Reese had done with Karnovsky, it established beyond a doubt that the endothelium was the anatomic site of the blood-brain barrier.

How Do Nutrients Cross?

Having established the existence of the barrier along with its ultrastructural basis, research workers turned to the next phase of the investigation: finding out how the molecules necessary for brain metabolism get across the barrier. The first step was to determine which substances can traverse the barrier and which are blocked by it. To this end physiologists and biochemists turned to new methods of tracing molecules in tissue that rely on radioactively labeled compounds.

A number of the early experiments shared a similar design that required two different radioactive substances. One was a reference compound, chosen because it embodied one of two extremes: it penetrated the brain either almost completely or not at all. The other compound was the test substance. Its rate of penetration was compared with that of the reference molecule by measuring the concentration of the two compounds in the veins leading away from the brain or in the brain itself.

Such tests were done for many substances, and it was found that several chemical properties help to determine how easily a molecule enters the brain. The most decisive factor is lipid solubility, which is roughly equivalent to how easily a substance dissolves in oil. Lipid-soluble molecules readily breach the blood-brain barrier and enter the brain. Among such substances are nicotine, ethanol and heroin, which helps to explain why these compounds are so frequently abused. Compounds that are highly soluble in water, on the other hand, tend not to be taken up into the brain, and that finding applies across a range of molecules extending



CAPILLARY IS ENSHEATHED by the astrocyte foot processes, which form a jigsaw pattern. The basement membrane holds the endothelium together and helps to maintain its overall tubular form. The lumen of the capillary is frequently just wide enough for a red blood cell to pass; as the finest extension of the vascular system, the capillary is the site of exchanges between blood and tissues.

from proteins as large as albumin to ions as small as sodium.

The reason lipid-soluble substances are able to penetrate the blood-brain barrier is that the leaflets of the cell membrane, including those of the capillary endothelium, are composed of lipid molecules. Each lipid molecule has two parts: a small head attached to two long hydrocarbon chains that form a tail. Each leaflet of the cell membrane is made up of many lipid molecules side by side, arranged with their heads facing outward. The complete membrane consists of two such sheets in close apposition. The lipid molecules in each leaflet are not bound closely to one another. Instead they are free to move in relation to one another, thereby forming a structure resembling a liquid in two dimensions. Only substances capable of dissolving in lipids can diffuse through this two-dimensional liquid, and only molecules capable of traversing the membranes of the endothelial cells can enter the brain.

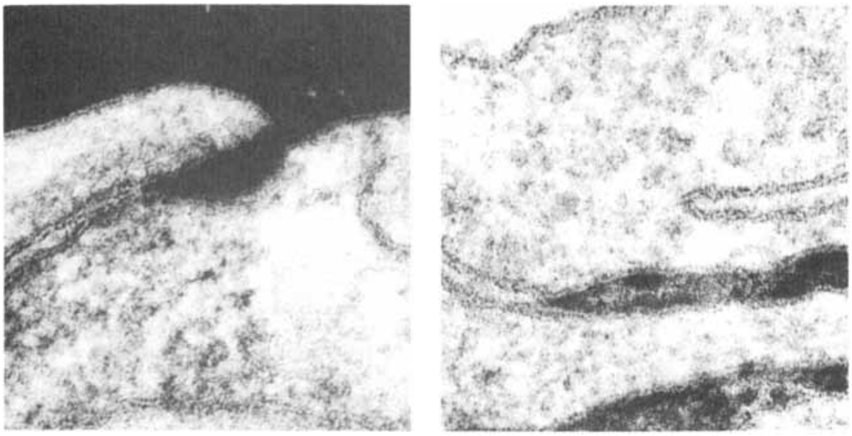
Nutrients Are Transported

Yet in order to function the brain also needs substances that are not lipid-soluble, such as glucose, the main source of energy for brain cells, and certain amino acids that the cells of the brain cannot manufacture for themselves. Because these substances are not soluble in lipids, they cannot cross the barrier by simple diffusion. Each essential nutrient must be recognized and brought across the membrane by a transporter specific to the nutrient molecule. The principle of specificity was first established for the glucose transporter in work that showed how elegantly discriminating the transport systems can be.

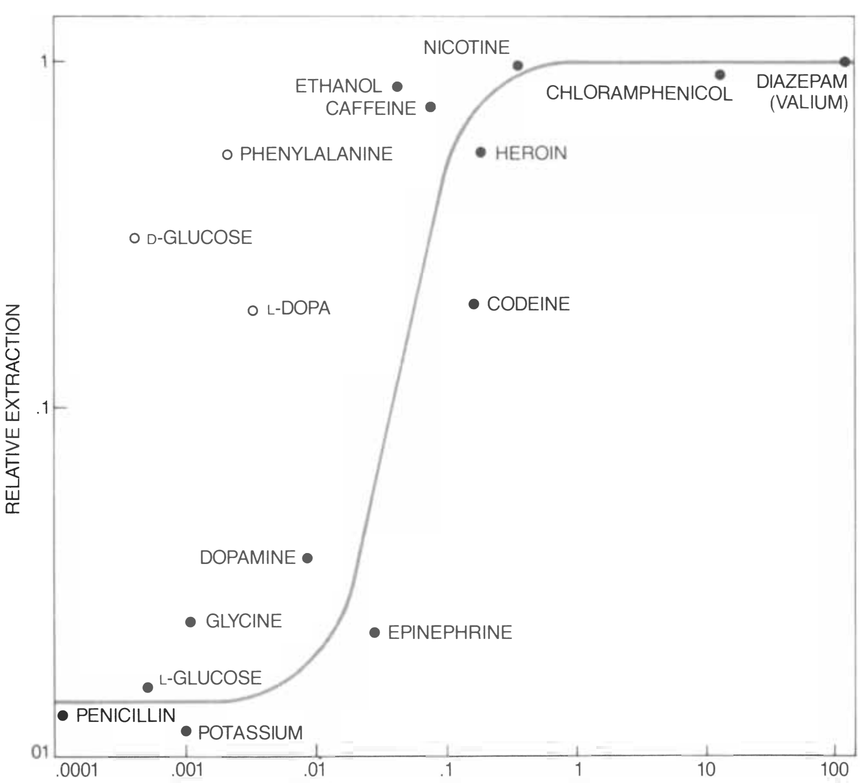
Christian Crone of the University of Copenhagen compared the rate at which two stereoisomers of glucose, called D-glucose and L-glucose, cross the blood-brain barrier. Isomers are molecules in which the same atoms are arranged differently; stereoisomers are mirror-symmetric. Both D-glucose and L-glucose can be tolerated by the body, but only the D molecule is biologically active. Crone found that the endothelium is capable of distinguishing between the stereoisomers. Although a large fraction of D-glucose is extracted from the blood by the brain tissue, no L-glucose is extracted. That result was quite different from the finding in muscle, where the two isomers entered the tissue equally well. Crone concluded that a highly specific transport system carries D-glucose to the brain, and he went on to propose what was then a novel hypothesis: that the transporter resides in the endothelium of the brain capillaries.

We began our collaboration after becoming interested in the blood-brain barrier independently during our research training. The curiosity of one of us (Betz) was aroused during predoctoral training in the laboratory of David D. Gilboe at the University of Wisconsin Medical School at Madison. Employing a modified version of the extraction technique, Betz found that the brain's glucose-transport system shares some properties with the better-understood system that carries glucose into red blood cells. As in blood cells, glucose transport into the brain increases with the concentration of glucose in the blood serum, but only up to a certain level. Beyond that maximum the rate of transport levels off, and so the system is said to be saturable. Like the blood-cell system, the one in the brain is inhibited by the drug cytochalasin *B* and is unaffected by insulin.

Neither the extraction technique nor any other method that entailed intact animals could carry the work much further, however, because all such methods share a crucial limitation: the presence of the brain cells interferes with observations of the capillary endothelium. To examine the operation of the blood-brain barrier without such confounding influences we saw the need to study capillaries in isolation from other tissues. At the time no techniques were available for isolating



BARRIER IN ACTION is captured in classic electron micrographs whose publication in the 1960's proved that the tight junctions between endothelial cells underlie the blood-brain barrier. Each micrograph shows the junction between two endothelial cells in a mouse brain capillary; the cell membranes appear as thin double lines. In both micrographs the lumen is at the top and the brain tissue at the bottom. The dark regions were made by the electron-dense stain horseradish peroxidase. When the stain is injected into the arteries supplying the brain, it reaches the capillary lumen but is prevented from entering the brain by the tight junction (*left*). When injected into the cerebrospinal fluid, it penetrates the cleft between endothelial cells until it is stopped by the tight junction (*right*). The micrograph at the left was made by Thomas S. Reese and Morris J. Karnovsky of the Harvard Medical School, the one at the right by Reese and Milton W. Brightman of the National Institute of Neurological and Communicative Diseases and Stroke after Reese joined Brightman there.



LIPID SOLUBILITY is the chemical property that has the greatest effect on a substance's capacity to pierce the blood-brain barrier. Lipids are fat molecules. The lipid solubility of a compound is measured according to how it distributes itself in a mixture of oil and water (*horizontal axis*). The ability to cross the barrier is measured in terms of how freely a molecule enters the brain compared with molecules known to do so most freely (*vertical axis*). For most substances ease of passage is determined largely by lipid solubility (*solid points*). Certain molecules needed for brain metabolism, however, cross the barrier more readily than their lipid solubility alone would suggest (*open circles*). Such compounds are carried across the barrier by a transport system. Most of the data on the graph are from William H. Oldendorf of the University of California at Los Angeles School of Medicine.

intact microvessels from the surrounding brain tissue, and so the other of us (Goldstein), then a postdoctoral fellow in the laboratory of Ivan Diamond at the University of California at San Francisco Medical Center, set out to develop one. The effort was successful, and not long afterward we joined forces and, first at San Francisco and then at the University of Michigan, took up our work on the blood-brain barrier.

Light on the Transporters

Our first joint project was to extend Betz's finding that the brain-capillary glucose transporter resembles the transport system of the red blood cell. Working with isolated capillaries, we showed that the two transport systems are indeed very similar. In addition, by measuring the amount of glucose transported into the endothelial cell and comparing it with the amount the endothelium metabolizes, we were able to show that only a small fraction of the glucose entering the endothelial cell is consumed there. It follows that most of the sugar passing into the endothelium is ultimately available to fuel the workings of the brain. Those results offered the first direct confirmation of Crone's hypothesis that glucose enters the endothelial cell through the membrane on the "blood" side and leaves it through the membrane on the "brain" side. (Because the interior of the vessel is known as the lumen, the membrane in contact with the blood is called luminal, the one in contact with brain tissue antiluminal.)

As the characteristics of the transport system became clear, attention turned to the transporters themselves. Recently S. I. Harik of Case Western Reserve University employed cytochalasin *B* as a probe to measure the number of glucose transporters in the endothelium. He found that each endothelial cell is richly supplied with transport sites, which helps to explain how the capillary can purge itself of the large surplus of glucose it extracts from the blood. The molecules that carry out this task, however, remain somewhat mysterious. Each transporter is undoubtedly made up of proteins that span the cell membrane, thereby forming a channel through which glucose passes, but the precise composition and configuration of the proteins are still being worked out.

One reason glucose was the first molecule whose passage into the brain was examined in detail is the fact that glucose is the only compound of its class present in significant quantities in the blood. Amino acids present a much more complex picture. The 20 amino acids that are components of proteins can be divided according to their chemical properties, which are largely determined by "side chains" attached to the common amino acid structure. Among the categories into which amino acids fall are the large neutral, the small neutral, the basic and the acidic.

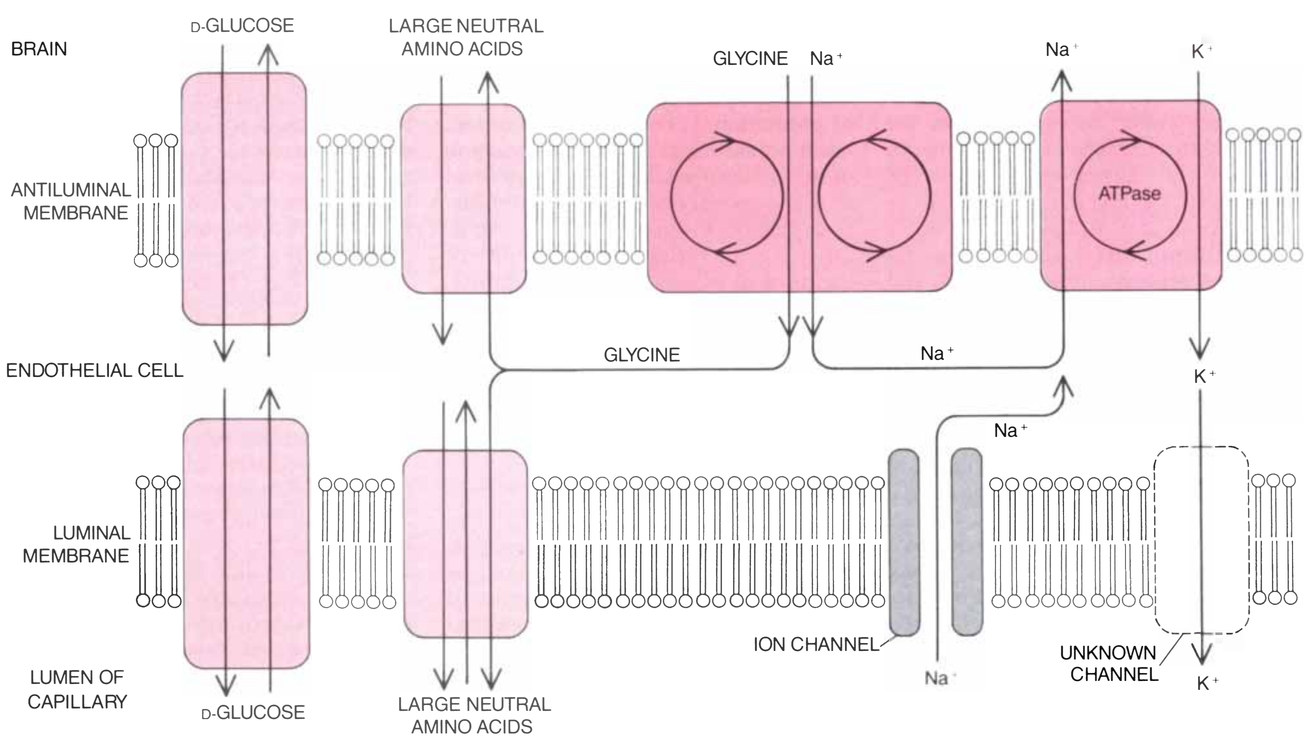
The contrast between the fate of large neutral amino acids and that of the small neutral ones reveals the complexity inherent in the system that brings amino acids to the brain. Large neutral amino acids are particularly important because they are required in the brain for the synthesis of neurotransmitters and proteins. William H. Oldendorf and William M. Pardridge of the University of California at Los Angeles School of Medicine, who have done much of the work in this area, found that there is a specific transporter for large neutral amino acids. That system is one of two specific amino acid carriers described by Oldendorf and Pardridge, the others being for basic and acidic molecules. In the case of the large neutral amino acids at least 10 of them compete for entry into the endothelial cell by means of the same transport mechanism.

Like the carriers for glucose, those for large neutral amino acids are present in both the luminal and the antiluminal membranes and therefore are capable of moving these "essential" amino acids across the endothelium and into the brain. Small neutral amino acids, on the other hand, are not transported from blood to brain. This is perhaps not surprising, since the small molecules can be synthesized by the brain cells and are therefore not "essential." Moreover, at least one member of the group—glycine—is a potent inhibitory neurotransmitter whose concentration must be held at a much lower level in the fluid around the nerve cells than it is in the blood. The transport properties of the capillary endothelium have a central role in maintaining this balance.

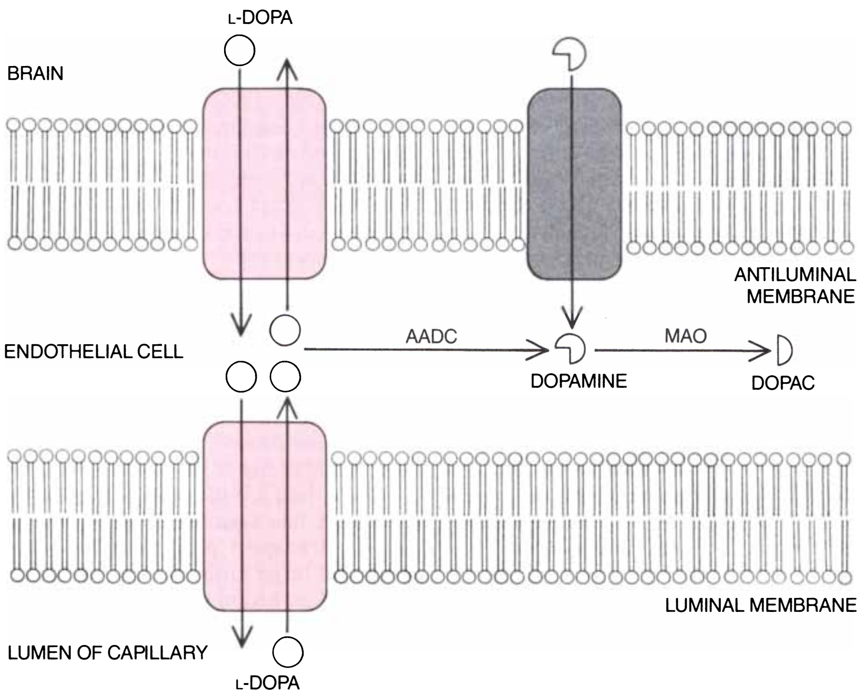
Asymmetric Transport Discovered

Elucidating that role depended in part on being able to study isolated brain capillaries. Several investigators applying brain-uptake methods in whole animals found that the passage of glycine from blood to brain was severely limited, and they concluded that the endothelium lacks glycine transporters. Working with isolated capillaries, however, we found a transport system capable of concentrating glycine within the endothelial cells. We reconciled these apparently contradictory findings by proposing an asymmetric distribution of transporters: the carrier for the small neutral amino acids is present in the antiluminal membrane but not in the luminal one. As a result the small neutral amino acids can be transported out of the brain but cannot be transported into it.

The asymmetric distribution helps to keep the concentration of glycine very low in the fluid surrounding the brain cells. Indeed, the concentration of glycine in the brain fluid is generally lower than it is in the endothelium, a fact that has significant implications for the transport process. When glucose and large amino acids are transported from blood to brain, they move from a region of high concentration to a region where their concentration is lower. In such "downhill" movement the energy of transport is provided by the concentration gradient itself, and the transporters merely facilitate the diffusion that would take place, albeit more slowly, without them. If, as in the case of glycine, a compound moves



COMPLEX SYSTEM OF TRANSPORTERS enables the brain capillary to control movement into and out of the tissues. Some of the transporters merely facilitate osmotic diffusion (*light color*); others are "active" mechanisms requiring a source of energy (*dark color*). D-glucose and large neutral amino acids such as phenylalanine reach the brain by means of transporters found in both membranes of the endothelial cell. These substances flow into as well as out of the brain. Potassium and small neutral amino acids such as glycine move only from brain to blood. They are pumped out of the brain by active transporters found mainly in the antiluminal membrane. The inward movement of glycine is coupled with that of sodium, which provides the needed energy. The energy for potassium transport comes from the splitting of a molecule of ATP by the enzyme ATPase; sodium is simultaneously pumped into the brain. Glycine may leave the endothelium by way of the carrier for large neutral amino acids. Potassium leaves it by an unknown channel.



"METABOLIC" BLOOD-BRAIN BARRIER consists of enzymatic steps by which compounds are modified in the endothelium and rendered unable to enter the brain. L-dopa, an amino acid that is a precursor of several neurotransmitters, enters and leaves the brain by means of the carrier for large neutral amino acids. Once in the endothelium, however, L-dopa may be converted into dopamine and DOPAC in successive steps by the enzymes AADC and MAO. Although dopamine can leave the brain by means of its own carrier, neither dopamine nor DOPAC can cross the antiluminal membrane into the brain. Hence the enzymatic conversions can serve as a means of controlling how much L-dopa reaches the brain.

"uphill" from a region of low concentration to a region of high concentration, energy must be supplied. For the small neutral amino acids the energy comes from sodium ions. The inward movement of glycine is coupled to the movement of sodium, which flows downhill into the endothelium along its own concentration gradient and pulls glycine with it.

Of course, once the small neutral amino acids are inside the endothelial cell they must have a way out or their concentration there would build up to untenable levels. Exit is probably provided by the same transporters that serve the large neutral amino acids. The small molecules have an affinity for the carriers of the large ones. Although the affinity is not as strong as that of the large neutral amino acids for their own carrier, it is strong enough to enable the small molecules to leave the cell.

Additional Asymmetries

Since the notion of an asymmetric distribution of carriers had proved so fruitful for understanding what happens to the small neutral amino acids, we pursued it further in studies of potassium transport. There were several reasons to think the movement of small neutral amino acids might provide a good model for the movement of potassium. Like glycine, potassium has a potent effect on the transmission of nerve impulses, and it is maintained at a low concentration in the fluid surrounding the brain cells. Another similarity is the relative impermeability of the luminal membrane of the endothelium to potassium ions. An additional resemblance was found by Michael W. B. Bradbury, then working at St. Thomas's Hospital Medical School in London, who discovered that potassium is removed from the brain by a saturable transport system.

Stimulated by Bradbury's observation, we showed in isolated capillaries that endothelial cells can transport potassium into themselves. Like the transport of glycine, the transport of potassium is "active" in the sense that it can overcome a concentration gradient. The mechanisms of the two types of active transport, however, are different, as we learned in collaboration with J. Anthony Firth, then at St. George's Hospital Medical School in London. Our work with Firth showed that the antiluminal membrane contains more of the enzyme called sodium-potassium ATPase than the luminal membrane does. That enzyme forms the basis of a pump that simultaneously transports sodium out of the endothelium into the brain and potassium out of the brain into the endothelium.

Since both functions of the pump must work against the prevailing concentration gradient, this mechanism consumes considerable energy. The requisite fuel is provided by the energy-rich molecule ATP (adenosine triphosphate). By splitting a molecule of ATP, the pump enzyme liberates enough energy to push sodium out of the cell while bringing potassium in. Naturally the accumulated potassium must have a means of egress, and that exit is probably provided by pores in the luminal membrane. This complex system enables the endothelium to suppress the level of potassium in the brain and rid itself of the excess potassium; its key is the radically unequal allotment of the pumping enzyme to the two faces of the endothelial cell.

Metabolism as Barrier

As the examples we have given so far suggest, the blood-brain barrier is based largely on the regulation of transport across the endothelium. Yet that is not the only way a compound can be barred from entering the brain. If a substance enters the endothelial cell, it can be converted by metabolic processes there into a chemical form incapable of traversing the antiluminal membrane to reach the brain. Perhaps the most striking demonstration of this mechanism, which has been called the "metabolic" blood-brain barrier, was provided by A. Bertler, B. Falck, C. Owman and E. Rosengrenn of the University of Lund in Sweden. They found that L-dopa, an amino acid precursor of the neurotransmitters dopamine and norepinephrine, readily enters the endothelium. Once there, however, it is modified by enzymes into a form that cannot reach the brain.

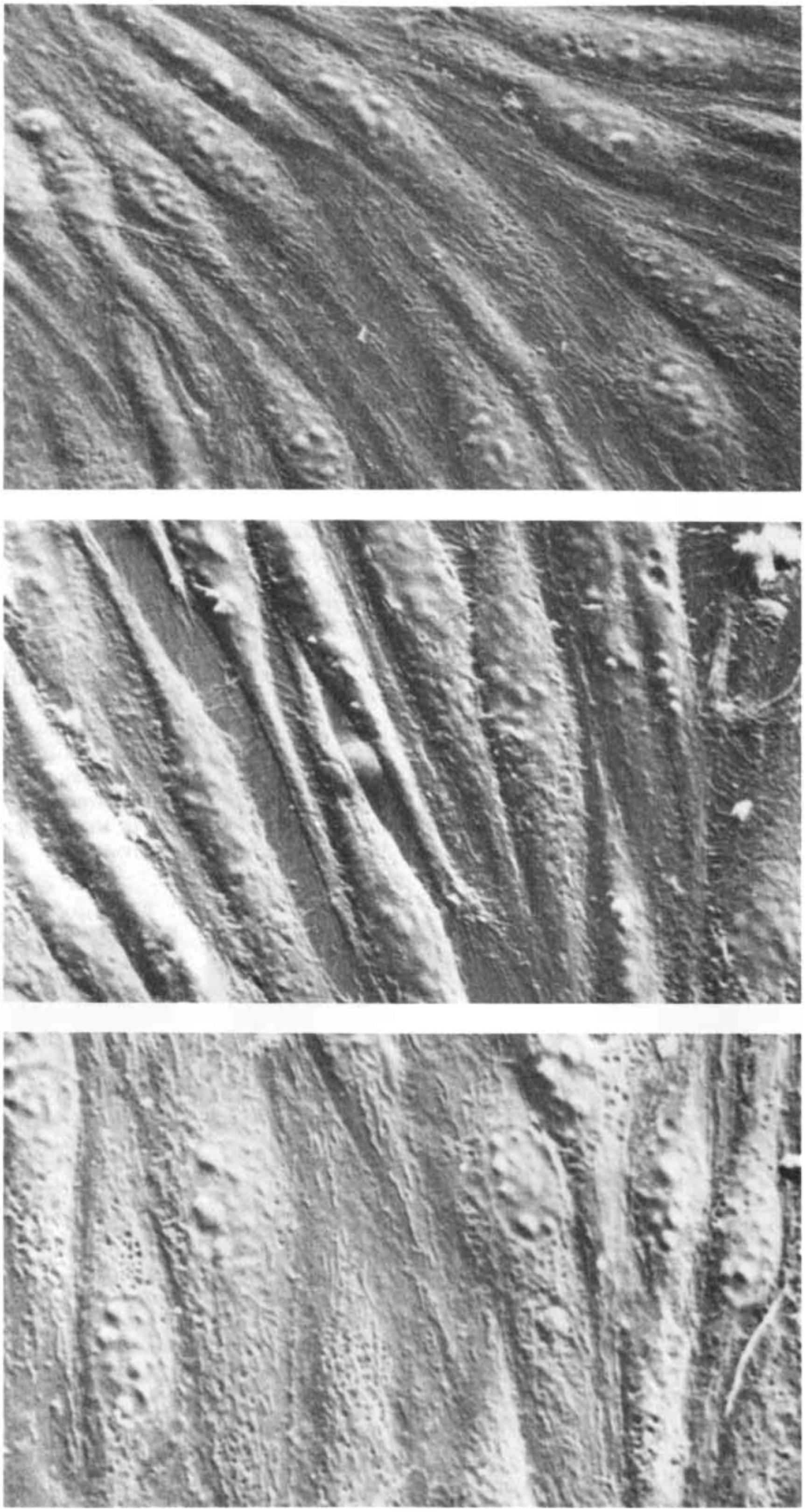
The discovery of the metabolic blood-brain barrier and the elucidation of the complex network of transporters that pepper the plasma membrane have tended to focus the attention of investigators on establishing the properties of the endothelium. Recently, however, some workers have begun to devote their energy to finding out how the endothelium of the brain capillaries acquires its unusual properties. Michael J. Wiley and Patricia A. Stewart of the University of Toronto transplanted brain and muscle tissue in avian embryos. When brain tissue was transplanted to the gut, microvessels from the gut growing into the transplanted brain tissue took on the characteristics of the blood-brain barrier. When muscle tissue was transplanted to the brain, brain microvessels growing into the gut tissue lost their special character. Such findings make it seem likely that the development of the barrier is stimulated by signals from the surrounding brain tissue rather than being genetically programmed in particular endothelial cells.

A likely candidate for the source of the developmental signal is the astrocyte, but it is very difficult to investigate the interaction of astrocytes and endothelial cells in whole animals or even in suspensions of isolated capillaries. Therefore workers in our group and in other groups have devised methods for growing endothelial cells in culture. The cells retain barrier features when first cultured, but after they are passaged (divided and grown in separate culture dishes) they lose the barrier configuration. L. E. DeBault and P. A. Cancilla, then at the University of Iowa, found that if passaged endothelial cells from mouse brain are grown over a layer of cultured glial cells, they regain their barrier properties. Thus we know that glial cells in culture have the transforming effect we suspect them of having in nature.

Culturing the Endothelium

Preparations of endothelial cells in culture are being used to study not only the origins of the blood-brain barrier but also how the barrier functions once it is fully formed. Phillip Bowman of our research group has developed methods for growing brain-capillary endothelial cells in culture, and we are employing his methods to examine changes in the permeability of the barrier. As a result of Bowman's work we can now prepare primary cultures of endothelial cells that display some of the essential characteristics of the blood-brain barrier: they contain few pinocytotic vesicles and they are frequently bound together by tight junctions. When Katerina Dorovini-Zis joined us after completing a fellowship under Brightman, she and Bowman collaborated to carry the work one step further.

Dorovini-Zis had been studying an intriguing phenomenon that was first explored by Stanley I. Rapoport of the National Institute of Aging. Rapoport found that if a hyperosmotic (highly concentrated) solution of sugar is injected into the carotid artery (which supplies the brain) of an experimental animal, the permeability of the blood-brain barrier is rapidly and dramatically increased. The change is temporary, and soon the barrier reverts to its



ENDOTHELIAL CELLS IN CULTURE provide a model for the operation of the blood-brain barrier. The three panels show endothelial cells grown from tissues taken from a cow's brain. When first cultured, the cells form a smooth monolayer (*top*). If they are transferred to a medium lacking calcium, the cells separate and reveal the underlying plastic dish (*middle*). When the cells are replaced in a calcium-containing medium, the monolayer re-forms (*bottom*). Such responses mimic those of the intact blood-brain barrier. Until recently it was not possible to grow brain-capillary endothelial cells in culture, but the authors' colleague Phillip Bowman and others have developed methods for doing so; the cells in the micrographs were grown in the authors' laboratory using Bowman's methods.

original state. To explain the transient enhancement of permeability, Rapoport proposed that the concentrated sugar solution somehow temporarily loosens the tight junctions between endothelial cells.

In her work at the National Institute of Neurological and Communicative Diseases and Stroke, Dorovini-Zis showed that Rapoport's hypothesis was correct. She observed the morphologic changes in the capillaries that follow injection of hyperosmotic sugar solutions into intact animals. Her observations showed that the tight junctions between endothelial cells do indeed separate temporarily after the sugar is injected. In our laboratory she and Bowman performed an analogous set of experiments with endothelial cells in culture. First they demonstrated that the cultured cells provide a functional barrier that keeps horseradish peroxidase from penetrating the tight junctions. Then they showed that after exposure to high concentrations of sugar the tight junctions separate and allow the stain to penetrate; when the sugar solution is removed, the barrier is reestablished.

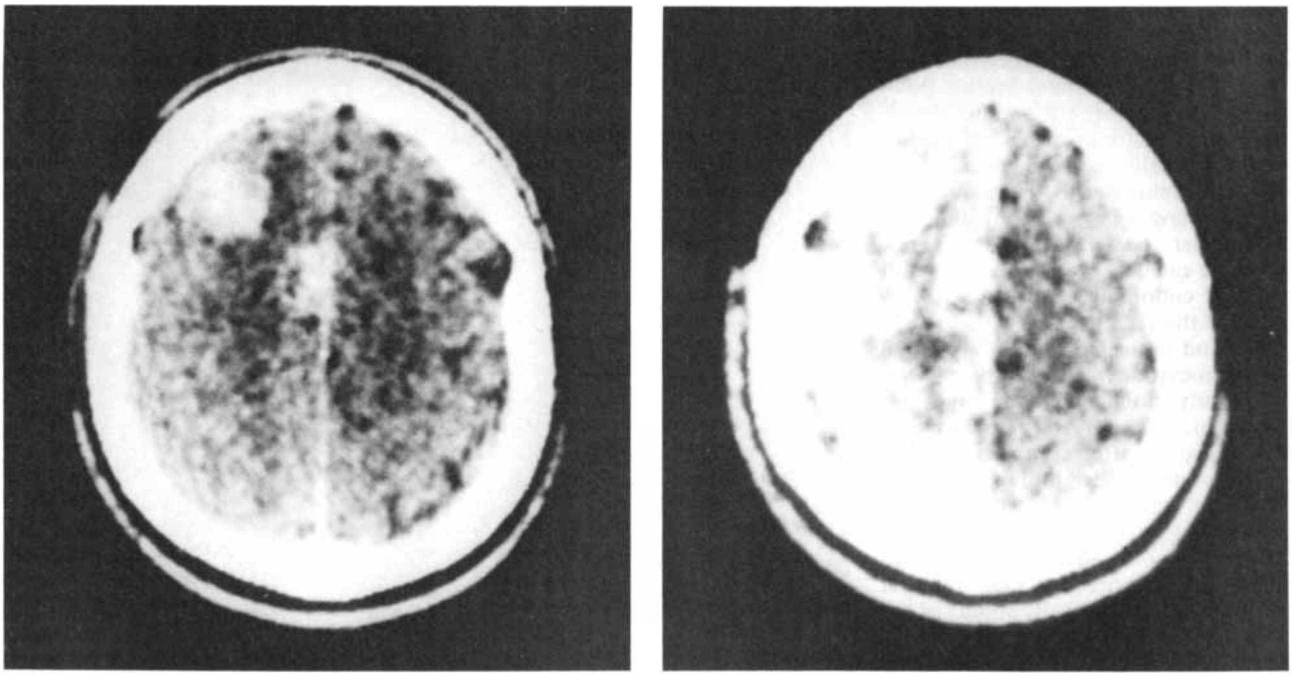
Implications for Medicine

The work of Dorovini-Zis and Bowman has given us a working model of the blood-brain barrier on a culture plate. In the years to come such a model may be valuable not only for basic science but also for its clinical implications. From the point of view of the clinician the fundamental problem presented by the blood-brain barrier is that of getting therapeutic agents through it into the brain to treat diseases that affect brain tissue. Some drugs, such as the antibiotic chloramphenicol, are highly lipid-soluble and therefore enter the brain readily. Many others, however, including such essential ones as penicillin, have a very low solubility in lipids and enter the brain slowly if at all. As it happens, penicillin has so little toxicity that it can be given in extremely large doses, and the high concentration of the antibiotic in the blood helps to compensate for its meager lipid solubility.

Unfortunately few medicines are as relatively harmless as penicillin. Most healing compounds have at least some undesirable side effects, and so they cannot be administered liberally in the hope that a fraction of the dose will reach its target. Finding a more precise mode of administering medicines to the brain generally entails one of two strategies. Either the blood-brain barrier must be temporarily lowered, allowing the drug to pass, or the barrier must somehow be circumvented.

One means of implementing the first strategy has been provided by Rapoport's observation that the permeability of the barrier is greatly enhanced by the injection of a hyperosmotic sugar solution into the carotid artery. Although Rapoport's work was done in animals, the transience of the increase in permeability he observed raised the hope that injecting hyperosmotic solutions into human patients might provide a reversible method of bringing therapeutic agents to the brain. Recently Rapoport's findings have been extended to human beings by Edward A. Neuwelt of the Oregon Health Sciences University. Neuwelt first showed that in patients with brain tumors the barrier can be opened by administering highly concentrated sugar solutions; his recent work suggests that chemotherapeutic drugs given in this way may contribute to regression or disappearance of the tumor.

The second general strategy, circumvention of the barrier, can be carried out by some straightforward and clinically tested techniques or by some more elegant methods that are still experimental. The most direct way to bypass the barrier is to inject the therapeutic substance into the cerebrospinal fluid. Such direct administration has long been common in the treatment of brain tumors. Direct injection



BARRIER IS OPENED by infusing a highly concentrated sugar solution into the carotid artery of a patient with a brain tumor. Both images are computed tomographic (CT) scans of the brain as seen from above; the forehead is at the top. The brain has been injected with a chemical reagent that causes areas where the blood-brain barrier is intact to appear dark and regions of increased permeability to appear light. The image at the left shows the brain before infusion; light areas at the upper left are tumors, where the barrier is somewhat opened. The image at the right, made 30 minutes after the infusion, shows how dramatically the sugar solution increases permeability to the tumor and the surrounding tissue. "Hyperosmotic" opening of the barrier serves as a means of delivering chemotherapeutic agents to the brain. The images were made by Edward A. Neuwelt of the Oregon Health Sciences University.

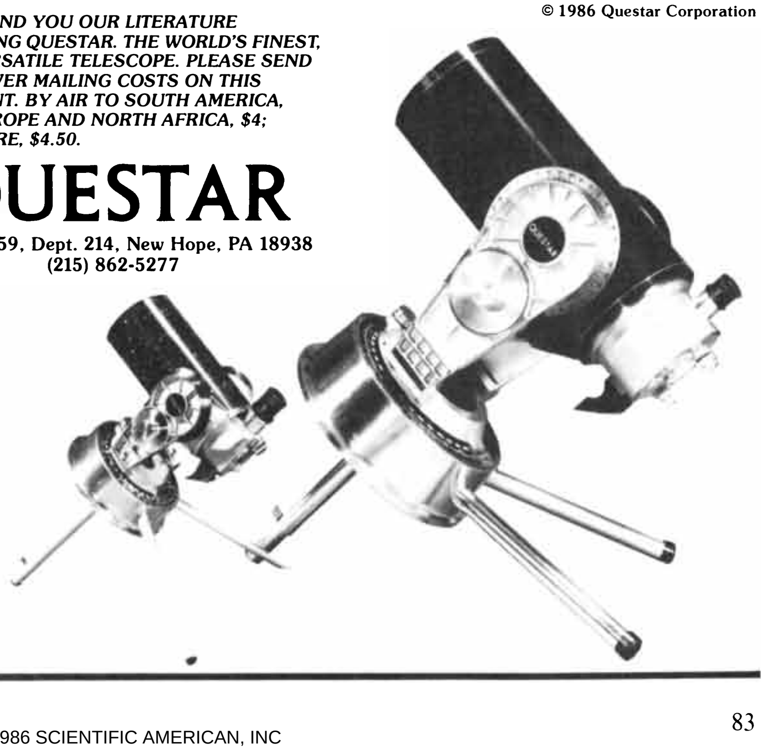
has the virtue of bringing the medicine directly to where it is needed, but in order to reduce the number of injections a large quantity of the drug is often given; this can lead to injurious side effects because of the toxicity of many chemotherapeutic agents.

A far more elegant approach (but one that is as yet thoroughly untested) would rely on the selective permeability of the blood-brain barrier in combination with the metabolic properties of the brain. If a drug could be designed to rapidly enter the brain and there be modified to a form incapable of crossing the barrier, it might be trapped in the brain tissues and have a sustained local effect. Such a principle, which seems promising as an experimental approach, is exemplified in nature by the disparity between heroin and morphine.

Tailoring a Drug

Heroin and morphine are quite similar in chemical structure. The only difference is that the heroin molecule has two acetyl (CH_3CO) groups attached to it at the positions where the morphine molecule has hydroxyl (OH) groups. The chemical properties of the acetyl groups endow heroin with considerable lipid solubility; morphine has very low lipid solubility. As a result of this chemical difference heroin enters the brain much more efficiently than morphine. (That may be one reason heroin is more commonly abused than morphine, in spite of the fact that both substances have potent effects on brain function.) Once heroin has entered the brain, however, enzymes there remove the acetyl groups and convert it into morphine, in essence trapping it in the brain. A therapeutic drug, tailored to take advantage of this pathway, could be delivered in a carefully controlled dosage and yet have a potent effect.

Such principles and others suggest that the recent advances in understanding the blood-brain barrier will have important clinical applications in the years to come. Perhaps the greatest value of the work done since the 1970's, however, lies not in its practical significance but in the satisfaction that comes from understanding a fundamental biological phenomenon. What makes us human can in large part be traced to a single organ: the brain. What has been learned about the blood-brain barrier in the past few years goes a long way toward showing how, within the controlled internal milieu of the body, an even more stringent regulation of physiological conditions makes it possible for the brain to carry out its remarkable work.



The Colors of Things

Color "illusions" devised (for the first time) on the display screen of a computer are evidence that color is not perceived just by sensing the light from individual surfaces in a scene

by Philippe Brou, Thomas R. Sciascia, Lynette Linden and Jerome Y. Lettvin

Color, like beauty, is in the eye of the beholder. It is a private experience and, like any other experience, is not accessible to measure. There is, however, a common belief that color, unlike beauty, is determined directly by physical cause: the spectrum of the light that falls on the eye. In particular the eye is taken to resemble a color television camera, which measures how much energy there is in the long wavelengths (red), middle wavelengths (green) and short wavelengths (blue) of the light at each point in an image. The eye, like the camera, has three types of color sensors, and so it is thought the color at every point is perceived by sensing the redness, greenness and blueness of that point.

This general belief is the impetus for many of the color "illusions" devised by investigators who study visual perception. The underlying notion in such displays is that you ought to see one color but in fact see another. Usually the "illusions" are taken to show the fallibility of the senses: to show the eye does not really work as well as it should. An alternative view is that the conventional notions about color vision are quite wrong if they are so easily and consistently violated. The "illusions" do not reveal defects. They yield insights into the unconventional design of color vision.

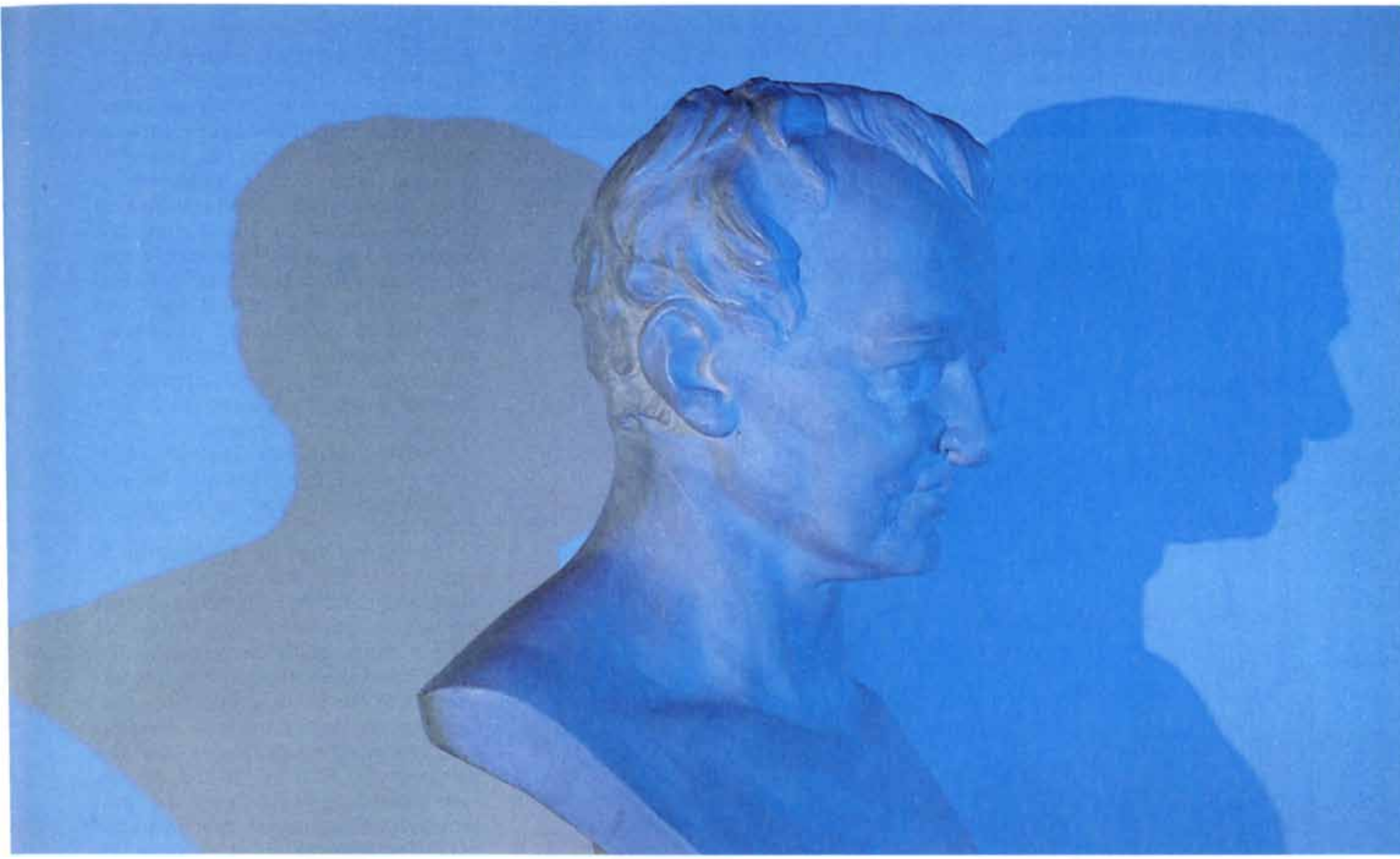
The unconventional notion we shall champion in this article—and test in a sequence of illustrations—is that the perceived colors of things in the visual world do not depend slavishly on the light from each object, sensed independently of all the other things in the world, but on a comparison of the lights from an object and its surroundings. That assertion seems mysterious. After all, the colors of objects are perceived as an intrinsic property of their surfaces. A red rose in daylight seems absolutely red, not red in relation to what is around it. But this perception of intrinsic redness does not signify that the redness is given only by the sense data in the image of the rose alone. We shall be concerned not with accounting for how perception seems to the perceiver but with the process applied to the sense data so that perception is possible.

The key that something is wrong with the conventional notions about the colors of things is in fact a color "illusion"—the most universal "illusion," an "illusion" so much in one's daily experience that it escapes notice. It is the color constancy of objects in daylight. Outdoor illumination is not fixed; it changes not only in brightness, from dawn to high noon to twilight, but also in its spectrum. At dawn the light is rosy. In the afternoon it is distinctly yellow. The north light preferred by artists is the blue of the sky when the sun is excluded. The diffuse light under a leafy arbor looks fairly green. Yet under the various forms of daylight a white sheet of paper is always seen as white. In fact all the surfaces one sees outdoors stay quite constant in color however the daylight varies.

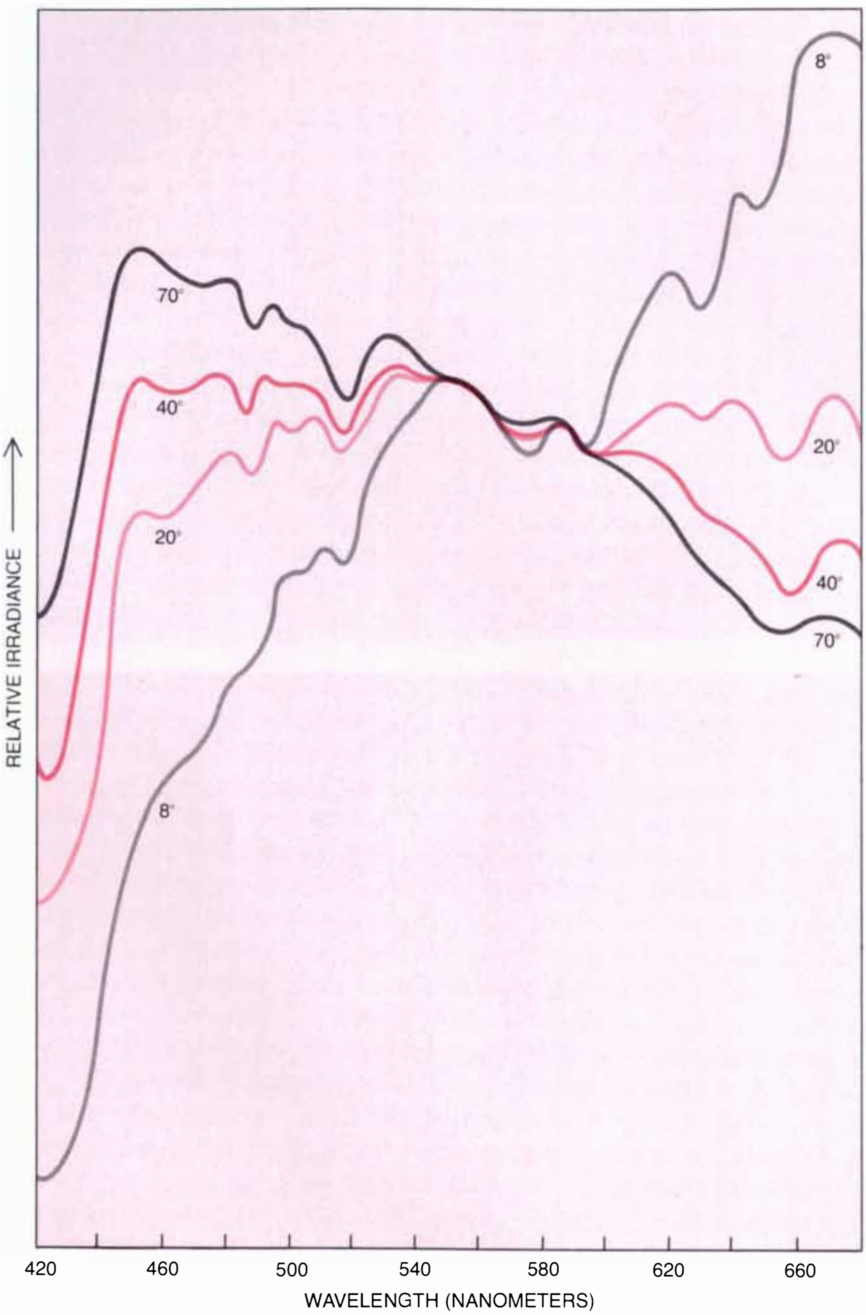
Today color constancy, if it is mentioned at all, is treated by a skeptical footnote in textbooks on vision. The skepticism is not hard to understand. Since many of us virtually live our lives under artificial lighting, we have become inured to the experience that faces and cosmetics, for example, change color between incandescent and fluorescent lighting. Yet the skepticism is wrong. Investigators of a century ago knew that color constancy poses a serious problem for efforts to understand perception. The constancy is so reliable that one routinely compares by memory the face color and lip color of someone who now stands under one form of daylight with the colors seen an hour ago, a day ago, a week ago, and under another form of daylight, and thus detects changes due to blushing or paling or signifying disease such as jaundice (which turns the skin yellow) or anoxia (which turns it blue). Such changes are much smaller than the possible changes in daylight color. It is as if, in the words of the 19th-century physiologist Hermann von Helmholtz, we "discount the illuminant" when we perceive color.

Helmholtz supposed that we know, by memory or intention, what colors we ought to see, and so adjust our perceptions to discount the lighting. But he was uncomfortable with the idea, and rightly so. It calls for a miracle, since one cannot tell, by means of any imaginable sense organ, what light plays on anything except that organ. That is, you cannot tell what light illuminates a surface; you only receive a part of that light—the part the surface reflects to your eye. How then can you "discount the illuminant"?

COLORED SHADOWS constitute an "illusion" that offers insight into the unconventional workings of color perception. In both photographs a bust of the 19th-century explorer and scientist Alexander von Humboldt has been placed in front of a white background. The bust is illuminated from one side by a beam of white light and from the other side by a beam of tinted light. The bust faces the tinted beam; hence the shadow behind the back of the head contains only white light. It does not look colorless, however: it has the color complementary to the opposite, tinted shadow. Thus the shadow behind the back of the head in the top picture looks "warmer" than the corresponding shadow in the bottom picture. Masking the shadows' surroundings will show that the shadows are virtually identical. The "illusion" hints that color perception relies on comparisons across boundaries in an image.



Our attempt to answer that question must take account of a difficulty that bedevils all efforts to talk about color. There are really two distinct types of explanation of color vision. The first type has to do with how information from a visual image is converted into sense data. That is, the explanation concerns itself with how the cones, the color sensors of the eye, react to light. In each cone the arrival of light bleaches a light-sensitive pigment; thus the explanation concerns itself with the rates of the pigment-bleaching events caused by arriving photons, or particles of light.



VARIETIES OF DAYLIGHT differ markedly from one another in their physical properties, notably their spectrum, yet the perceived colors of objects in daylight are remarkably constant in spite of the changing illumination. The phenomenon is called color constancy. The four curves chart the spectrum of the diffuse illumination at ground level at four times of day, when the sun is at various angular distances between the horizon (0 degrees) and the zenith (90 degrees). The curve for the sun at eight degrees was measured about half an hour before sunset. The peak in the spectrum of diffuse daylight then occurs at roughly 660 nanometers, or deep in the red part of the spectrum. Hours earlier, when the sun is high in the sky, at 70 degrees, the peak is at wavelengths some 200 nanometers shorter, or well into the blue part of the spectrum. The four spectra were first published by S. T. Henderson.

The experimental analyses that reveal the laws of this transduction start with Newton. The strategy is to have human subjects act as what are called null measuring devices. The subjects adjust a reference light (they can, for example, adjust the brightnesses of three colored lights that combine to produce a spot of light) until the boundary between the reference and an adjacent test light vanishes and the reference and the test light can no longer be distinguished. The experiments establish that certain lights bleach the visual pigments identically. In other words, the experiments serve only to build up a description of the pigmentary responses. They are independent of anything that may happen to the sense data after the cones produce them.

The second type of explanation is very different. It lies beyond the domain of measurable quantities such as the brightnesses of test lights. Instead it takes as its realm the perception of color. Everything known about the apparatus of the eye and everything inferred about the brain from physiology and psychology shows that the perceiver is directly aware not of raw sense data but only of the consequences of a process applied to the data. The appropriate strategy in devising an explanation of color perception is therefore to take the laws of data acquisition as given (the first type of theory addresses that matter) and search out the rules for dealing with the data.

We think it facilitates such a search to adopt the attitude of the engineer, who is concerned not so much with analyzing the world as with designing a system that fulfills a particular purpose. (The engineer always "builds to specifications.") The art is to conceive of the purpose of color vision and then design a system to realize it.

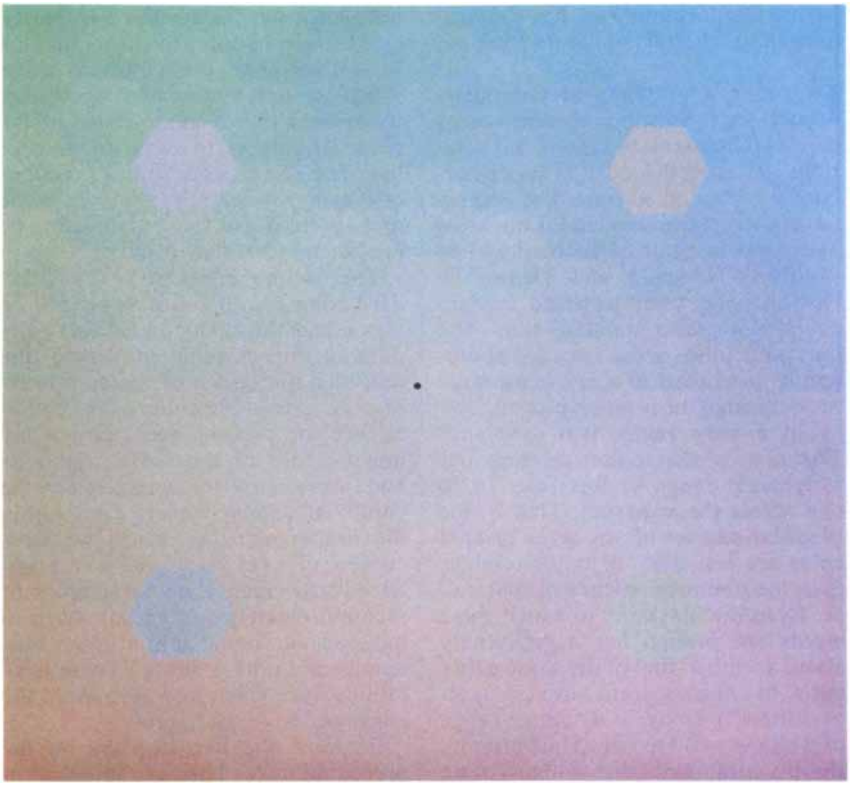
Color vision, as a useful faculty, evolved in a primeval world in which light from the sun—scattered, refracted and reflected—was the chief illuminant. The things it lighted were mostly solid, opaque, nonmetallic objects, which reflected the light according to their material composition and the roughness, irregularities and nonplanarities of their surfaces. The objects were distributed independently over the earth. Any image of such a world is a two-dimensional map of a determined chaos: a bounded region in the image is almost always bounded by many other regions, and the change in light across one of the boundary segments has little predictive value for the change across any other boundary segment. In sum, three conditions—daylight, varied reflectances and a diversity of arrangements—are the constraints under which color vision evolved.

Imagine the purpose of color as an aid to seeing. Color vision improves the ability to tell surfaces apart in a memorable way, so that nourishment, threats and so on can be learned and reliably recognized. Since color vision supports many more distinctions than monochromatic black-and-white vision, there is surely an advantage to seeing in color. Yet unreliable distinctions are useless. Indeed, they are a hindrance. They are busyness without meaning and so are equivalent to noise. This gives value to color constancy. The ability to recognize things would be lessened if their color changed simply because of a change in the illumination.

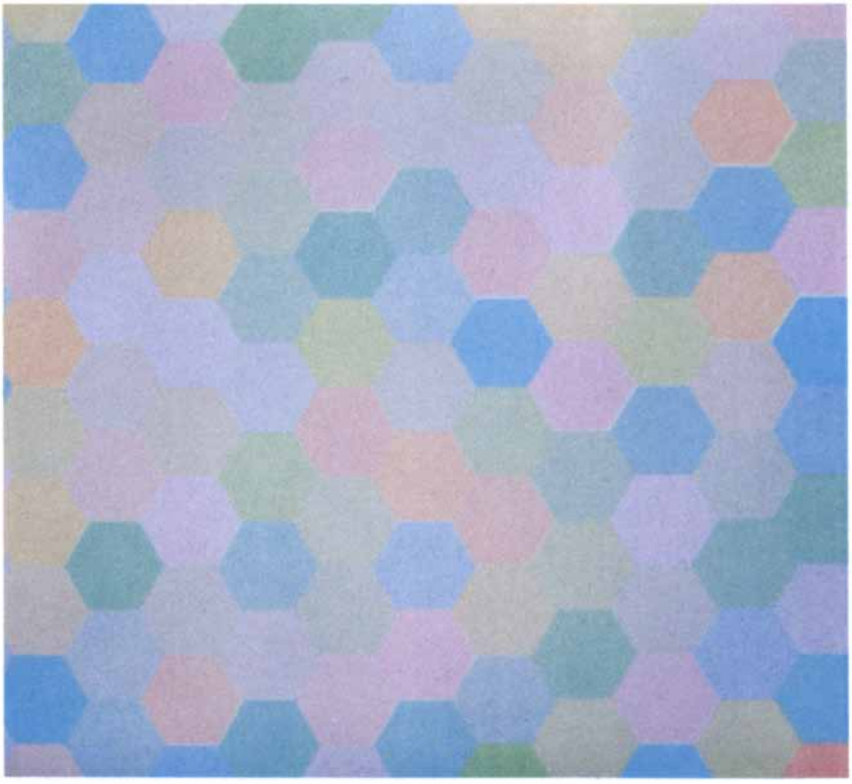
What then, in terms of color, will serve as a memorable quality of a surface, a quality both intrinsic to the surface and independent of the accidental circumstance of what form of sunlight plays on the surface? The only quality that meets these conditions is reflectance: the ability of the surface to reflect light of each given wavelength. As it is ordinarily defined, the reflectance is unknowable. It is the ratio of the light incident on a surface to the light the surface reflects, and (as Helmholtz ruefully knew) the eye has access only to the latter. Hence the notion of employing reflectances as a basis for color vision seems absurd. It is absurd, however, only if one accepts as dogma that the color of a surface depends on the light from that surface alone.

There are other possible tactics. The most straightforward one replaces measures of the light reflected from individual surfaces with comparisons of the lights reflected from pairs of adjacent surfaces. If the surfaces are under the same illumination, the comparison is independent of the incident light. As a result the comparison of the reflected lights becomes equivalent to a comparison of the reflectances (which are unknowable individually). We have arrived, then, at the hypothesis that perceived color is a quality of the reflectance of a surface compared with the reflectances of other surfaces. It is a testable idea; it suggests, for example, that the perception of a stable color at a given point in an image requires that the regions near the point be sufficiently diverse. It is an idea we shall now examine by means of a sequence of color "illusions." They are patterned after displays devised by Edwin H. Land, who founded the Po-

1



2



"ILLUSIONS" IN AN ARRAY of colored hexagons were photographed on the screen of a computer. Each hexagon in the first display also appears in the second. In the first display, however, the hexagons are ordered chromatically; in the second one their positions have been randomized. The randomization makes the range of colors look more vivid. Five hexagons have the same chromaticity (the same spectrum) and retain their position from one display to the other. In the first display they seem to have somewhat different colors; in the second one the color (a fairly neutral gray) is recognizably the same for all five of them.

laroid Corporation and has devoted himself to the study of color vision.

In this era of personal computers with color display screens, color manipulation is easy; indeed, the "illusions" we shall show can be had on ordinary television screens. The display screen was photographed (at the Massachusetts Institute of Technology) by Bradford Howland and Denise D. Denton, who used standard camera equipment and a standard film. The only difficulties arose because photographs published in a magazine have to go through the printing process.

There were really two problems. The first is that color printing has a dynamic range of less than 10 to one across the spectrum. That is, the brightest patches of any given printed color are less than 10 times brighter than the dimmest patches of that color. Even the black ink in which these words are printed has a reflectivity about a tenth of that of the white paper itself. In contrast, transparencies (color "slides") enjoy a dynamic range of well over 100 to one. Thus printing sharply compresses the vividness of an image in a way that a color transparency does not. In order to ensure that the images we photographed would not be much corrupted by such compression, we limited the dynamic range of the color brightness to three to one, or less. The pictures are printed versions of displays that could safely be made more vivid if they were intended to be projected rather than printed.

The second problem is that standard color printing uses three colored inks (magenta, cyan and yellow) quite different in chromaticity from the red, blue and green of the computer-display-screen phosphors. In consequence the printing press cannot lay down greens of the same brightness and saturation as the ones given by the computer display screen, even within the limited dynamic range we have chosen. (To get vivid greens in printed advertisements, manufacturers of menthol cigarettes often pay extra to have an additional ink, a green one, employed in the printing.) These limitations were taken into account in the construction of the figures.

Examine the two displays on the preceding page. They are identical in their components. Each display is an array of hexagonal patches of color, and every hexagon that appears in one display appears somewhere in the other. Every hexagon has a "chromatic signature" defined by three numbers, one giving the intensity of the red pixels (the individual points of brightness on the computer display screen), the second that of the green pixels and the third that of the blue pixels. The only thing that differs from one display to the other is the positions of the hexagons with respect to one another: their addresses on the screen.

In the first display the hexagons are ordered according to their three components. Red increases in intensity from top to bottom of the display, green from bottom to top and blue from left to right. The steps in intensity from each hexagon to the next have been arranged to be somewhat stronger than just visible. The second display is a very different image. The range of colors seems wider and more vivid. Yet the only difference is that the colored hexagons have been shuffled—given random addresses rather than ordered ones.

It is instructive to try to tell the change of a given colored hexagon's address from one display to the other. For a hexagon at the boundary of the first display the new, randomly set position in the second display is relatively easy to find. But we doubt that your eyes will be of much help to you in discovering most of the new positions unless individual hexagons are masked off from the others.

Five of the hexagons in each display are special in two respects. First, their addresses are constant. In both the first display and the second they are the central hexagon and four outlying hexagons, one near each corner of the display. Second, the five hexagons are identical in chromatic signature. The rest all differ in their characteristic triple of numbers.

Try to compare the five. In the first display they should seem to be of differing colors, in spite of the narrow dynamic range of the illustration. (The effect would be stronger if we could increase the dynamic range.) But if, at ordinary reading distance (about 16 inches), you fix your gaze on any one of them, it should turn indeterminate in color.

It is unlikely that you can make out the central hexagon. It is marked with a small dot in its center. If you fix your gaze on the dot (closing one eye), the colors of the four outlying hexagons should seem strong at first, but within a few seconds the hexagons

3

REARRANGING THE RING of colored hexagons that border each gray hexagon does much to abolish the wandering colors of the gray hexagons and make them look reliably gray. In other words, it goes far toward establishing color constancy. The rearranged rings represent the only change from the first display; the rest of the colored hexagons retain their chromatic order. The display therefore suggests that color perception relies in particular on local comparisons across the boundaries between patches of color in a visual image.

should vanish. Indeed, the entire picture should seem to become quite featureless except at its perimeter: its boundary with the white page.

Again the second display is quite different. The five hexagons should look uniformly gray, although they occupy the same positions and have the same chromatic signature they had in the first display. If you fix your gaze on any one of the five, you should find that its color remains reliably gray. Furthermore, fixing your gaze on the central dot should not bring on a general deterioration of the image, as it did for the first display.

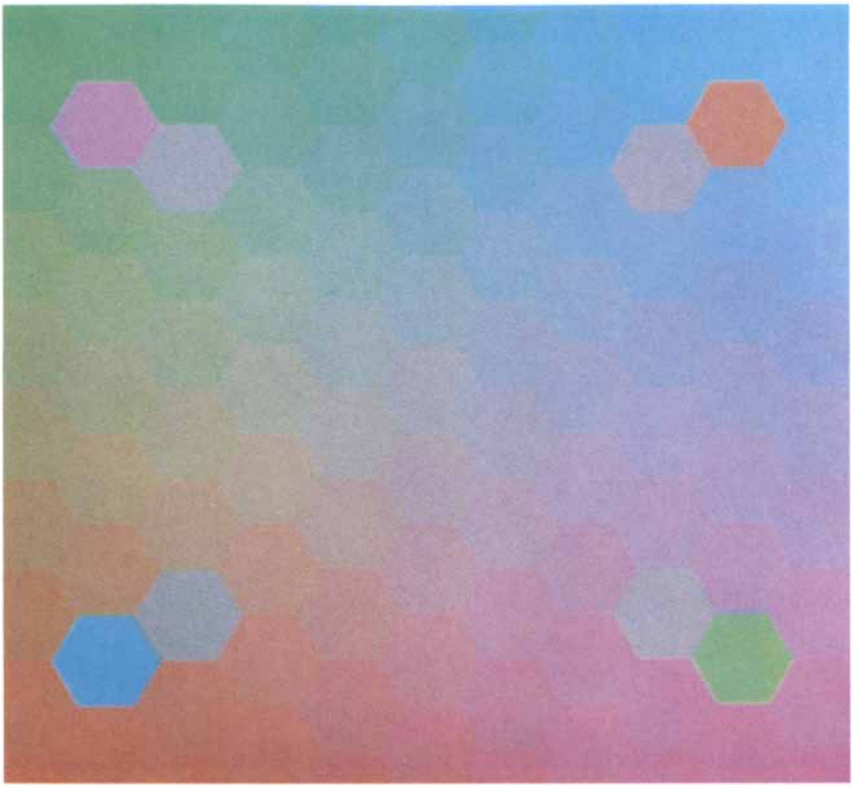
The results of these explorations are readily summarized. The randomizing of the colored hexagons surrounding the five gray hexagons somehow leads to more vivid color perception and at the same time stabilizes the gray hexagons. Conversely, the ordering of the colored hexagons lessens the vividness of their colors and makes the gray hexagons wander in color: it saps their color constancy.

In the third display, shown on the opposite page, we have modified the scheme of the first display only in the immediate vicinity of the four outlying gray hexagons. In particular, the ring of hexagons around each gray hexagon has been rearranged: the hexagons that bordered a gray hexagon in the first display still border it in the third, but their annular order has been changed. The averages of the intensities of the red, the green and the blue pixels immediately around a gray hexagon therefore remain as they were. Yet now the gray hexagons do not differ much in their color.

The fourth display, on this page, is even more faithful to the scheme of the first. Four outer hexagons bounding the outlying four gray hexagons have been exchanged diagonally. Otherwise nothing is altered. Yet again the gray hexagons no longer differ much in their color.

In the fifth and sixth displays, shown on the next page, the phenomena are much as they were in the first and second displays. The difference is that the computer program that generates the images has been altered so that the chromatic signatures are specified in terms of magenta, cyan and yellow, the pigments used in color printing. (The computer display screen still glows in its usual red, blue and green, but the proportions of these glows are adjusted to give combinations that mimic the printing inks.) The color "illusions" still hold, although, as one would expect, the seen colors differ from the ones in the first display.

4



DIAGONAL EXCHANGE of colored hexagons also goes far toward making the gray hexagons look reliably gray. Here a single colored hexagon bordering each gray hexagon has been switched diagonally with its counterpart across the illustration; in all other respects the scheme of the first display is maintained. The exchange shows that the presence of a small area of marked spectral dissimilarity is sufficient impetus for color constancy.

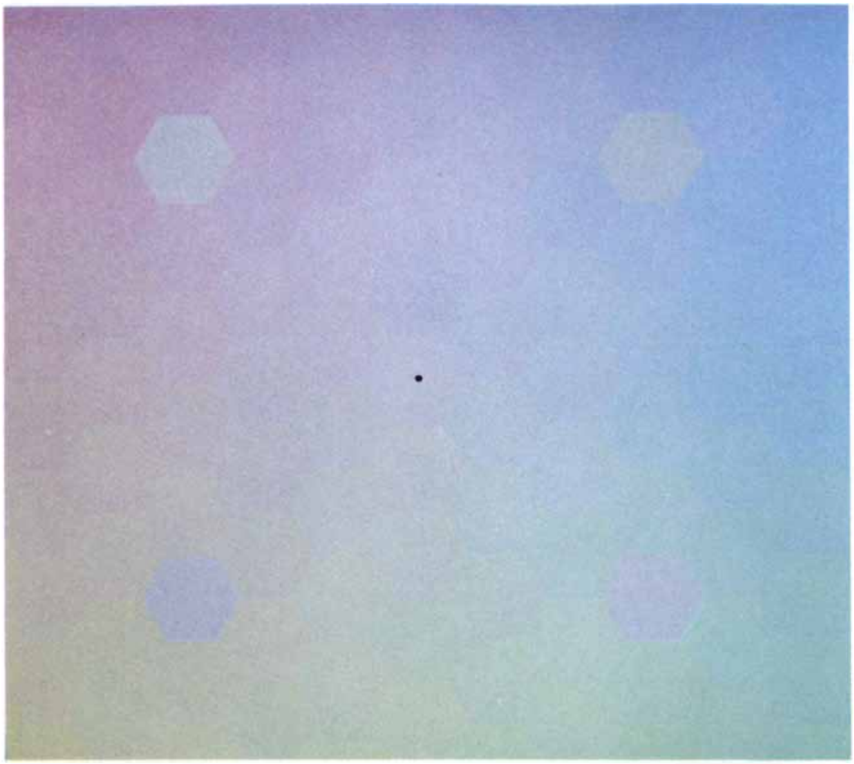
We have one more "illusion" to show. The obverse of color constancy is the phenomenon called colored shadows. Its most familiar form can be seen on page 85, where we display two photographs of a real scene: a bust of the 19th-century scientist Alexander von Humboldt placed in front of a white background. In each photograph the bust is illuminated from one side by a tinted light and from the other side by a white light. It is hard to tell which light is which. One of the shadows in each photograph has the color of the tinted beam. (The bust has blocked off the white beam.) Yet the other shadow (where the tinted beam is blocked off) does not look colorless; it has the color complementary to the first shadow. The effect is particularly striking in patches of shadow on the bust itself. (In the illustration the tinted lights are to the right. The beams were adjusted to limit the dynamic range. A more florid effect can be had with projectors and transparencies.)

Why are colored shadows the obverse of color constancy? In both phenomena the color attributed to a light is different from what it "really is"—different from what the physical properties of the light would lead one to predict. In colored shadows the identical spectral distribution (that is, the same physical stimulus) has different colors. In color constancy different distributions have the same color.

In essence, then, the perceived colors of the five gray hexagons in the first display constitute a demonstration of colored shadows. Imagine that you could adjust the reflectance of each of the four outlying gray hexagons in the display so that they would all appear to be the same color as the central hexagon (the one with the dot). The four adjustments would be different. You would have turned colored shadows (the same stimulus looking different) into a semblance of color constancy (different stimuli looking the same). If the remaining hexagons were then shuffled, as they are in the second display, the five hexagons would have distinctly different colors among themselves.

Plainly the processing of color information from sense data does not give a slavish one-to-one correspon-

5



6

THE SAME "ILLUSIONS" as the ones in the first two displays are shown here for colored hexagons whose chromaticities have been adjusted to resemble combinations of the inks employed in color printing. That is, the hexagons vary in steps of brightness for cyan (a greenish blue), magenta (a purplish red) and yellow. Again the five gray hexagons in the display seem to have different colors when they are surrounded by ordered chromaticities, and they are seen to be the same when their surroundings are randomized. As one would expect, the perceived colors of the gray hexagons differ from the ones in the first display.

dence of perceived color to the spectrum of light. There is no such correspondence in one's daily outdoor experience, which is of color constancies and colored shadows everywhere. And so the rules for color processing must be different in kind from the laws that govern the responses of the visual pigments to light.

Some hints at the nature of the rules are suggested by the displays. The first hint is that colors are determined at boundaries and vertexes (where boundaries meet). Those are the only places in the pictures where the chromatic data change, so that ratios of reflectances can be taken. In this regard the first display is particularly instructive. Again, fix your gaze (with one eye closed) on the central dot in the picture. Observe how the four outlying hexagons vanish after a few seconds, then reappear, then vanish. With a little attention to detail, you should discover that the reappearance always accompanies a movement of the eye, so that after a little practice you can actually control the vanishing and reappearance.

This is a crude example of the phenomenon of stabilized images on the retina. It has been known since the 19th century that if an image is held steady on the retina, it vanishes, to be replaced by a peculiar feeling of blindness called the "empty field." The vanishing is due mainly to the fact that each receptor (each cone) in the eye adapts, or adjusts itself, to the light on it. Receptors are designed to signal to the nervous system only change, not a steady state. Vision is good as long as the eye moves about. When the eye rests, vision deteriorates. This confirms in no uncertain way that the boundary and vertex information is crucial for seeing, just as we inferred from the phenomenon of color constancy. The eye must move so that individual receptors can experience boundary crossings.

A second point is that the color of an area depends not only on the surroundings of the area but also on the running history of the region of the retina on which the area's image falls. That is, a set of receptors exposed to many different lights in succession will give a signal for any specific presentation of light that is different from the signal they will give if their history of exposures is an impoverished one. This conclusion too derives from the first display. Recall that if you fix your gaze on any of the four outlying hexagons, it rapidly becomes indeterminate in color. On the other hand, if you observe the picture cursively, letting your gaze wander about it, the impression of four differently colored hexagons remains.

The phenomenon of stabilized images suggests the explanation. Since a stabilized image vanishes, the strategy of looking fixedly at any surface in order to determine its color is unavailing. But since the receptors do not adapt instantaneously (modern apparatus shows that an image, however bright and contrasty, vanishes in about two seconds), the ordinary jiggle of the gaze gives abundant and varied information to the individual cones.

The second display and the ones that follow it teach the same lessons. They show that as the surroundings of an area become more varied, the color of the area becomes more definite and also more stable over time. Moreover, the color becomes more reliably a correlate of the area's reflectance. In the real world, diversity is the rule. There is rarely an orderly display of the kind shown in the first image—except in the patterns that give some animals protective coloration. This natural diversity, and the continual change of the direction of the gaze, ensures that the spatial surroundings of an area and the temporal history of the exposure of cones to light are as variegated as the engineer of a color-vision system could wish.

We turn, then, to some of the concerns of an engineer who undertakes to design a color-vision system. One of the most useful operations in any sensing system, natural or artificial, is a running normalization. In psychology it is called adaptation; in engineering it is automatic gain control. The idea is to adjust the sensitivity of the system to the average level of input so that all changes are made to lie in the same limited dynamic range. This is done by taking a running average of the input and using it to set the gain, or amplification factor. Imagine, for instance, a camera lens that darkens in bright light, grows clear in dim light, and does it so well that photographs taken with the same exposure time at bright noon and at twilight have the same quality.

The strategy of automatic gain control, applied to cone sensitivity, might take the following form. Call the intensity of the light currently falling on a cone L. The running history of arriving light is designated A. The latter is an average of the intensities of lights that recently fell on the cone and is adjusted so that the effect of a light diminishes with its distance into the past. Helmholtz called it "dark light."

Examine the difference between the light and the dark light, divided by the sum of the two (that is, the ratio of $L - A$ to $L + A$). It achieves automatic gain control. Under steady light, such that A (the dark light) is the same as L (the light), the ratio is zero. If L suddenly increases, the ratio becomes positive; if the increase is maintained, A builds up to L and the ratio lapses back to zero. If L suddenly decreases, the ratio becomes negative and then, as A reduces to L, returns to zero. If L increases or decreases by a fixed fraction of its value, the response is a fixed change in the ratio. Finally, a sudden increase of L to an immense multiple of A can only drive the ratio to a value approaching 1; a sudden decrease of L to a minute fraction of A can only drive the ratio to a value approaching -1. Indeed, the ratio can never go beyond ± 1 (the so-called compression limits).

You must imagine now a sheet of cones, each of which incorporates this strategy. If they were all independent of one another, any image whatever, stabilized on that sheet, would give zero signal everywhere. Conceive of the cones, however, as interconnected laterally, so as to communicate the A's to one another with a strength inversely related to distance. A stabilized image still gives zero signal everywhere. Yet a change in L at any cone is now read not only with respect to the cone's own adaptation level, A, but also with respect to all the A's of the cones around it. That is, the individual cones provide a temporal average of the dark light and their interconnections provide a spatial average. Three interconnected sheets of such cones, one for each of the three cone pigments and interdigitated with one another, provide a trichromatic reference for a change in L in any part of the image. The output signals tell nothing of the normalizations that underlie them; they have automatically adjusted themselves to compensate for changing circumstances of illumination. The system is well on its way to achieving color constancy.

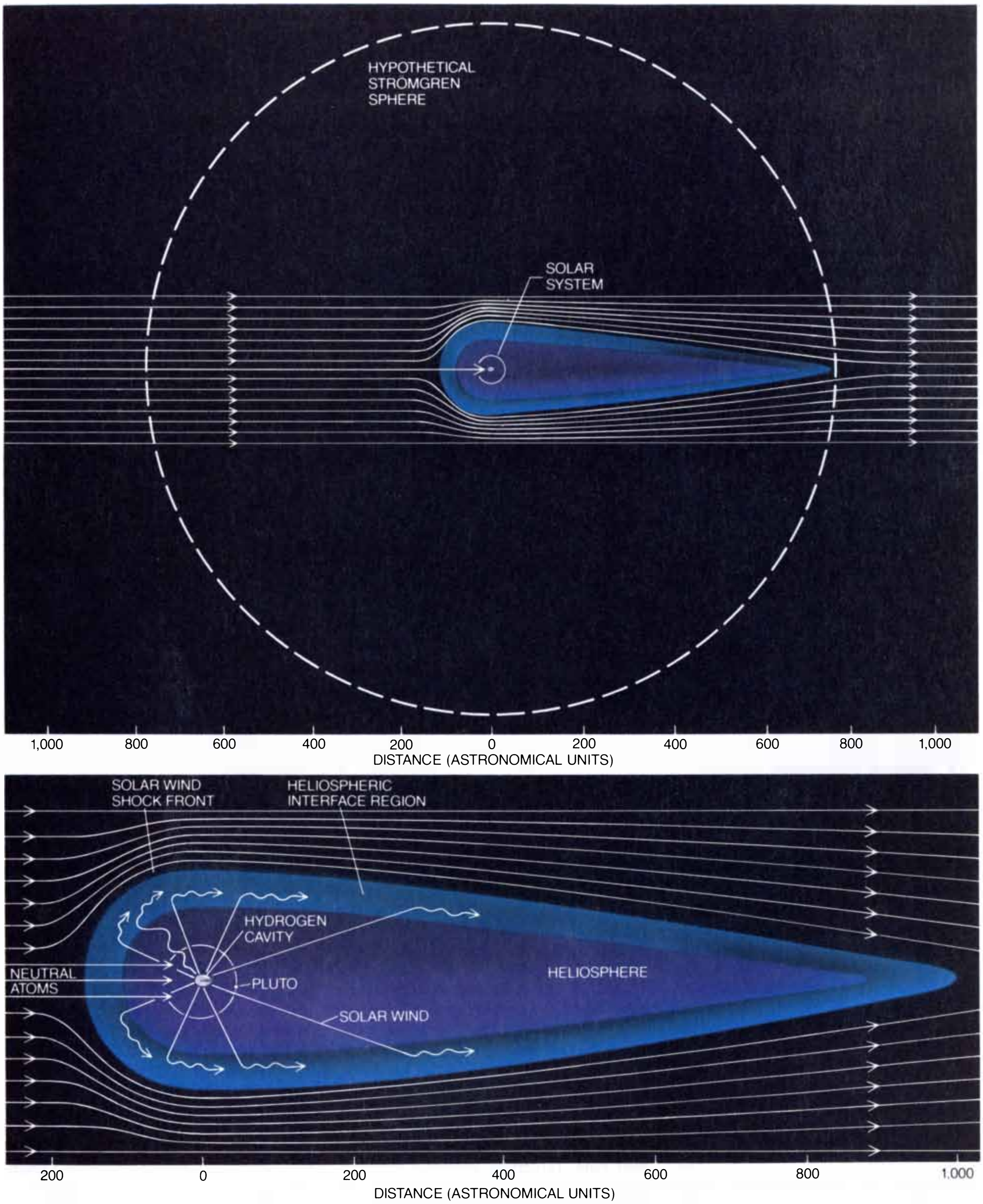
At this point diversity, as a feature of the real world, enters in an important role. Diversity provides that the history of all local retinal areas, for an eye that moves about, is on the whole much the same. It also provides that the average of the various lights falling on any local retinal area has a good chance of approaching the average for the wider visual field around it. Further, it provides that the spatial averages resemble the temporal averages. Hence the diversity of the world ensures that local comparisons across boundaries have a good chance of yielding reliable color constancies for an entire scene.

Diversity applies not only to the distribution of reflected lights but also to the distribution of reflectance ratios across boundaries and around vertexes. These ratios are largely independent of the relative sizes of the areas that are bounded, so that these ratios are the basic information in vision, an imbalance in size between neighboring areas becomes unimportant over a fairly wide range of imbalance. Indeed, in the fourth display a small area of marked spectral dissimilarity from the ordered background is enough to ensure the color constancy of the adjacent gray hexagon.

To what extent does this color theory apply to the human visual system? It is known that the visual system employs a succession of processing stations. The cones pass information to neural circuitry in the retina, which recruits further stations by means of its output channel, the optic nerve. The nerve engages the brain structure called the lateral geniculate body, which in turn engages the part of the cerebral cortex called the primary visual cortex. The cones, then, are the very first elements in a complex sensory system. It may seem surprising that we assign them (and their interconnections) so much responsibility in the color process.

Physiological evidence has established, however, that the detection of boundaries and vertexes begins in the retina. And since each boundary signifies a change in quality (a difference in color or brightness, for instance), the detection of a boundary implies the prior existence of that quality, so that a difference can be detected. (Distinguishing a boundary between red and green implies the prior distinction of redness and greenness.) It would be extremely uneconomical, and also unreliable, for the retina to defer the color process to later stages of the visual system by exporting information about qualities such as color by channels that are separate from the channels that signal boundaries. Moreover, it would take up an inordinate amount of channel space.

Simple economy leads, therefore, to the position that the processing of color determinants by the cones and their intimately associated retinal apparatus offers a reliable system—one that is preferable, we think, to any deferred processing by more central tissue, such as the visual cortex. It remains to be seen whether this is the strategy nature has chosen.



INTERACTION OF THE SUN with the surrounding interstellar medium, consisting primarily of hydrogen and helium gas, is illustrated schematically. In the early 1960's astrophysicists thought radiation from the sun could ionize the interstellar medium, carving a huge zone of plasma that would extend far beyond the entire solar system (*top*). Such a zone is called a Strömgren sphere, or H II region. The discovery of large amounts of neutral (un-ionized) hydrogen well inside the solar system rendered that model invalid. Interstellar material does engulf the solar system, but the material is moving with respect to the sun and is able to penetrate well within the solar system before it is ionized. As the gas streams through the solar system its particles are swept into a long tail downwind of the sun (*teardrop region at top and enlarged view at bottom*). A boundary known as a collisionless shock front forms as a consequence of interactions between charged particles in the inflowing interstellar material and the outflowing solar wind, the stream of charged particles emitted by the sun. Immediately inside the shock front lies a turbulent "shell" of plasma called the heliospheric interface region. The outer surface of the interface forms the heliopause; the inner surface of the interface bounds a region known as the heliosphere.

The Sun and the Interstellar Medium

A cloud of interstellar gas is now streaming through the solar system. Past encounters with denser clouds may have substantially affected the earth's climate

by Francesco Paresce and Stuart Bowyer

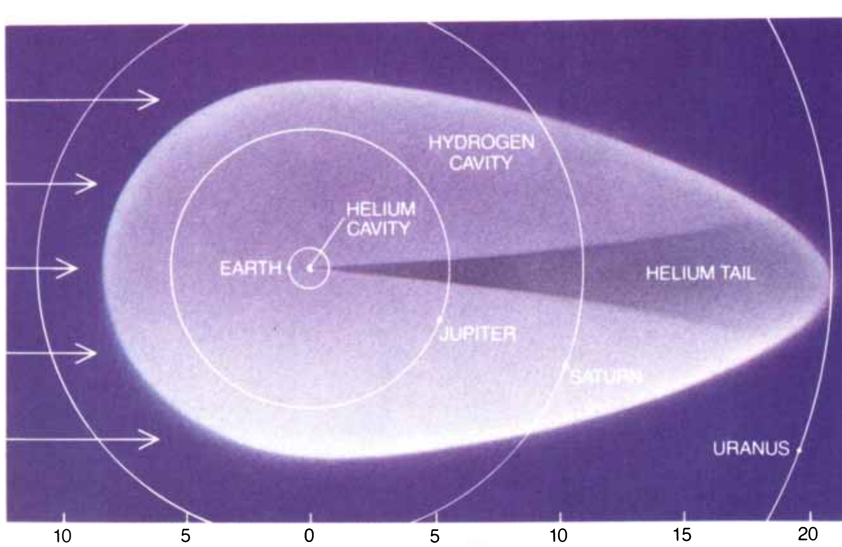
Every star is born within a dense cloud of gas and dust, and during the early stages of its life it resides within the remnants of the cloud. The primordial cloud eventually dissipates, leaving the star to shine in solitary brilliance, but this does not necessarily mark the final meeting of the star with interstellar matter. Later one of many immense clouds of gas drifting randomly through the galaxy may engulf the star in the course of a chance encounter.

The sun is now immersed in a cloud of gas that is relatively tenuous by astronomical standards and enormously tenuous by human standards, containing only about .1 atom per cubic centimeter of space. This local interstellar medium is composed primarily of hydrogen and helium and is streaming through the solar system at the modest astronomical speed of about 20 kilometers per second, or nearly 50,000 miles per hour. The material appears to be coming roughly from the direction of the constellation Centaurus and to be heading in the direction of the constellation Cassiopeia.

It is surprising that a thorough understanding of the interaction of the sun with the interstellar medium has begun to emerge only in the past 15 years. Before then most astrophysicists believed a simple model that successfully accounts for the interaction of hot, massive stars with interstellar clouds could also be applied to the sun, which is relatively cool and small. The subsequent work of many investigators including ourselves has culminated in a new model, rich in detail, that describes the interaction of the sun with the interstellar medium. The model has even led some workers to conjecture that the earth's ice ages may have resulted from past encounters of the solar system with dense molecular clouds.

In discussing how the sun interacts with the interstellar medium it is convenient to begin with the simple model that explains the interaction in the case of hot, massive stars known as spectral types O and B. They are 20 times more massive than the sun and have surface temperatures that exceed 30,000 degrees Celsius (which is five times more than the surface temperature of the sun). Type O and B stars emit prodigious amounts of radiation that has wavelengths of less than 912 angstrom units, or 91.2 billionths of a meter—the critical value at which hydrogen ionizes. Ionization is a process in which an atom is stripped of its valence, or outermost, electrons. In the case of hydrogen, for instance, ionization yields a single negatively charged free electron and a single positively charged free proton.

It follows that a type O or B star can carve a huge zone of highly ionized plasma into virtually any interstellar cloud that might surround it. The zone is called a Strömgren sphere, or an H II region, and may extend 100 to 200 light-years from the star. A particularly striking example of such a process occurs in the Orion nebula, which is a relatively nearby complex of hot young stars and dense interstellar material. In this case the material is sufficiently inhomogeneous so that no



IONIZED CAVITIES of both hydrogen and helium gas surround the sun, but they are several orders of magnitude smaller than the calculated size of the hypothetical static solar Strömgren sphere. Both cavities are well within the bounds of the heliosphere. The outer surfaces are drawn at the point where half of each of the neutral species has been ionized.

actual spherical zones are formed, but the physical process is nonetheless the same. The effects are spectacular enough for the Orion nebula to be a prime subject for astronomical posters and similar illustrations.

The tremendous success of the Strömgren-sphere model in describing the interaction of massive stars with the interstellar medium led astrophysicists to believe the model would also hold for the sun. They thought even the relatively small amount of electromagnetic radiation emanating from the sun would be enough to create a solar Strömgren sphere. Although the calculated .02-light-year-diameter region would be small by galactic standards, it would be large compared with the size of the solar system, extending 25 times beyond the orbit of Pluto, the outermost planet. In this picture the only neutral atoms that would exist within the spherical cavity would be found in the atmospheres of some of the planets. The rest of the region would be filled with plasma.

A startling discovery made in the early 1960's by Herbert Friedman and his colleagues at the Naval Research Laboratory would ultimately lead to the overthrow of the concept of the solar Strömgren sphere. The N.R.L. workers launched several rockets equipped with electromagnetic radiation detectors high above the atmosphere of the earth. The instruments recorded an intense and diffuse glow of ultraviolet light that had a wavelength of 1,216 angstroms. Light having this wavelength is one of the signatures of neutral, or un-ionized, hydrogen. Hydrogen emits 1,216-angstrom light when its electron drops from an energy state called the first excited level to the ground level, or lowest energy state. The intensity of the ultraviolet light indicated that the neutral hydrogen must exist in considerable amounts in interplanetary space.

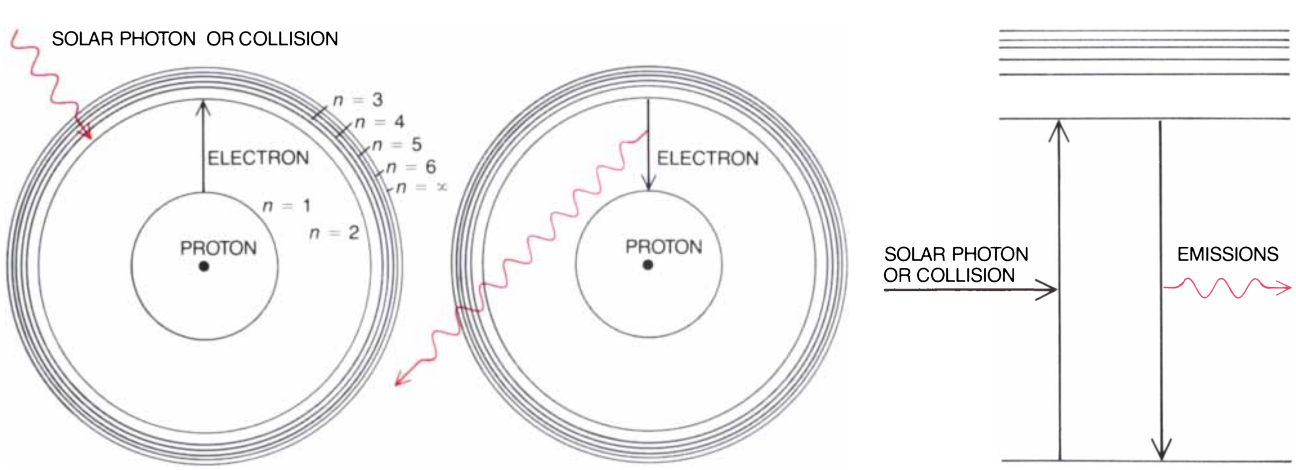
Investigators were puzzled by Friedman's results because according to the solar Strömgren-sphere model, any neutral hydrogen in the immediate vicinity of the solar system (except in the atmosphere of planets) should be ionized quickly by the sun. Some theorists initially tried to explain the anomaly by postulating that gas from distant parts of our galaxy was the source of the 1,216-angstrom ultraviolet light. Further work ruled out that hypothesis. In 1970 W. H. Chambers and his colleagues of the Los Alamos National Laboratory reported that the ultraviolet light exhibited an anisotropy that could not be accounted for if gas in distant parts of the galaxy was the source. Moreover, Phillip W. Mange and Robert R. Meier of the N.R.L. found that the intensity of the ultraviolet light varies directly with the activity of the sun.

At this point and with only these meager clues Hans J. Fahr and Peter W. Blum of the University of Bonn made a bold proposal. They suggested that the source of the ultraviolet light is indeed neutral interstellar gas within the solar system. They argued that the gas remains neutral because it is moving with respect to the sun and can therefore penetrate well within the solar system before being ionized. The effect would be particularly pronounced in the case of the sun, whose radiation output is relatively weak.

Although in retrospect the proposal of Fahr and Blum may seem obvious, at the time it represented an extraordinarily innovative idea. Indeed, the hypothesis was initially greeted with skepticism. Although the present era is one of open scientific inquiry, research is nonetheless carried out by people who share the normal range of human emotions. In particular, one American scientist was so affronted by Fahr and Blum's proposal that he attempted to suppress the publication of their work. Fortunately he did not succeed.

An enormous amount of evidence corroborates Fahr and Blum's theory that an interstellar cloud composed predominantly of neutral gas is drifting through the solar system. In the early 1970's a team of investigators led by J. L. Bertaux and J. E. Blamont of the Aeronomy Department of the French National Center for Scientific Research (CNRS) and Gary E. Thomas and R. F. Krassa of the University of Colorado at Boulder mapped the emission of the 1,216-angstrom ultraviolet light. They based their map on data transmitted to the earth from instruments flown on the satellite *Orbiting Geophysical Observatory*. They determined that most of the light is emanating from neutral hydrogen just inside 10 astronomical units of the sun, which is roughly a fourth of the distance between the sun and Pluto. (An astronomical unit, or AU, is the average distance between the earth and the sun: 1.5×10^8 kilometers.)

At virtually the same time we made a set of measurements that also confirmed the suggestion of Fahr and Blum. Working at the University of California at Berkeley, we developed an instrument with which to detect the 584-angstrom ultraviolet light emitted



SIGNATURE OF NEUTRAL HYDROGEN is ultraviolet light that has a wavelength of 1,216 angstrom units, or 121.6 billionths of a meter. Neutral hydrogen gives off light of this wavelength when its electron drops from an energy state called the first excited level ($n = 2$) to the ground level ($n = 1$), or lowest energy state. Either collisions with other atoms or absorption of 1,216-angstrom radiation can excite the electron to the higher level (*left*). After about 10^{-8} second the hydrogen atom returns to the ground level by emitting 1,216-angstrom radiation in an arbitrary direction (*middle*). The conventional approach to depicting the process is an energy-level diagram (*right*).

by neutral helium when it drops from the first excited level to the ground level. In 1970 we flew the instrument on a rocket and discovered a faint but clearly distinguishable glow that indicated the presence of neutral helium. The discovery was confirmed in 1972 by other investigators who mapped the helium emission extensively with a number of satellites and interplanetary probes. Most notable in this regard were the measurements of Meier and Charles S. Weller, Jr., of the N.R.L., who in 1974 reported data obtained with detectors on a Department of Defense satellite. In 1978 A. Lyle Broadfoot, Shailendra Kumar and Joseph M. Ajello constructed a more detailed map based on data from the *Mariner 10* probe.

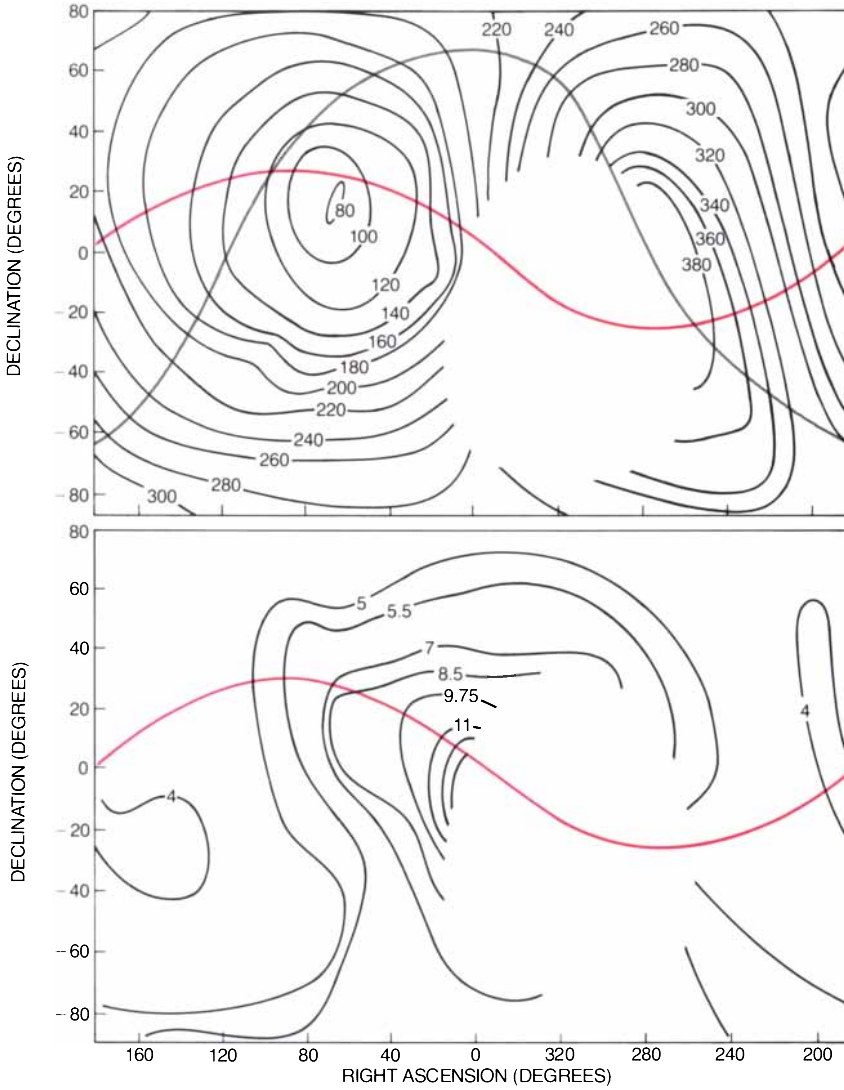
The results confirm the fact that an interstellar cloud consisting primarily of neutral hydrogen and helium atoms engulfs the solar system. The hydrogen atoms absorb 1,216-angstrom ultraviolet light from the sun and reemit radiation of a similar wavelength in another arbitrary direction. An identical process takes place for the neutral helium, except the absorbed light has a wavelength of 584 angstroms. In other words, the neutral hydrogen and helium atoms act as diffusers of solar radiation. An observer who is not looking directly at the sun will see an essentially monochromatic and diffuse radiation pattern. The effect is somewhat like the scattering of sunlight by fog.

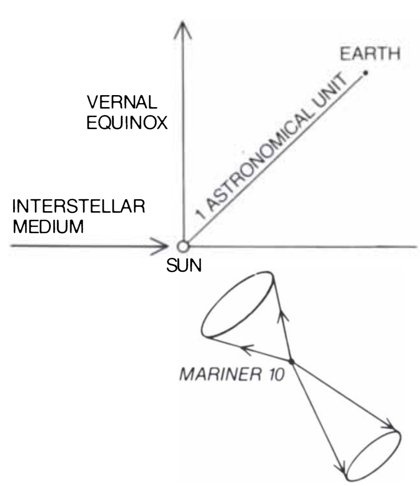
When the material passes through the solar system, the neutral atoms "feel" two opposing forces: the attractive gravitational force of the sun and its repelling radiation pressure. Depending on the stage of the 11-year solar cycle, the radiation pressure can diminish, neutralize or even overpower the gravitational pull of the sun on neutral hydrogen atoms. At times of low solar activity gravity dominates the radiation pressure and draws neutral hydrogen atoms toward the sun. At more active phases in the solar cycle the radiation pressure may balance the gravitational attraction. The atoms will then stream through the solar system in a straight line parallel to their primary flow direction.

At times of high solar activity the radiation pressure forces the neutral hydrogen atoms away from the sun. The radiation pressure of the sun is important for the motion of hydrogen only; neutral helium, which is four times as massive as hydrogen, is affected less. Helium is always focused into a region behind the sun.

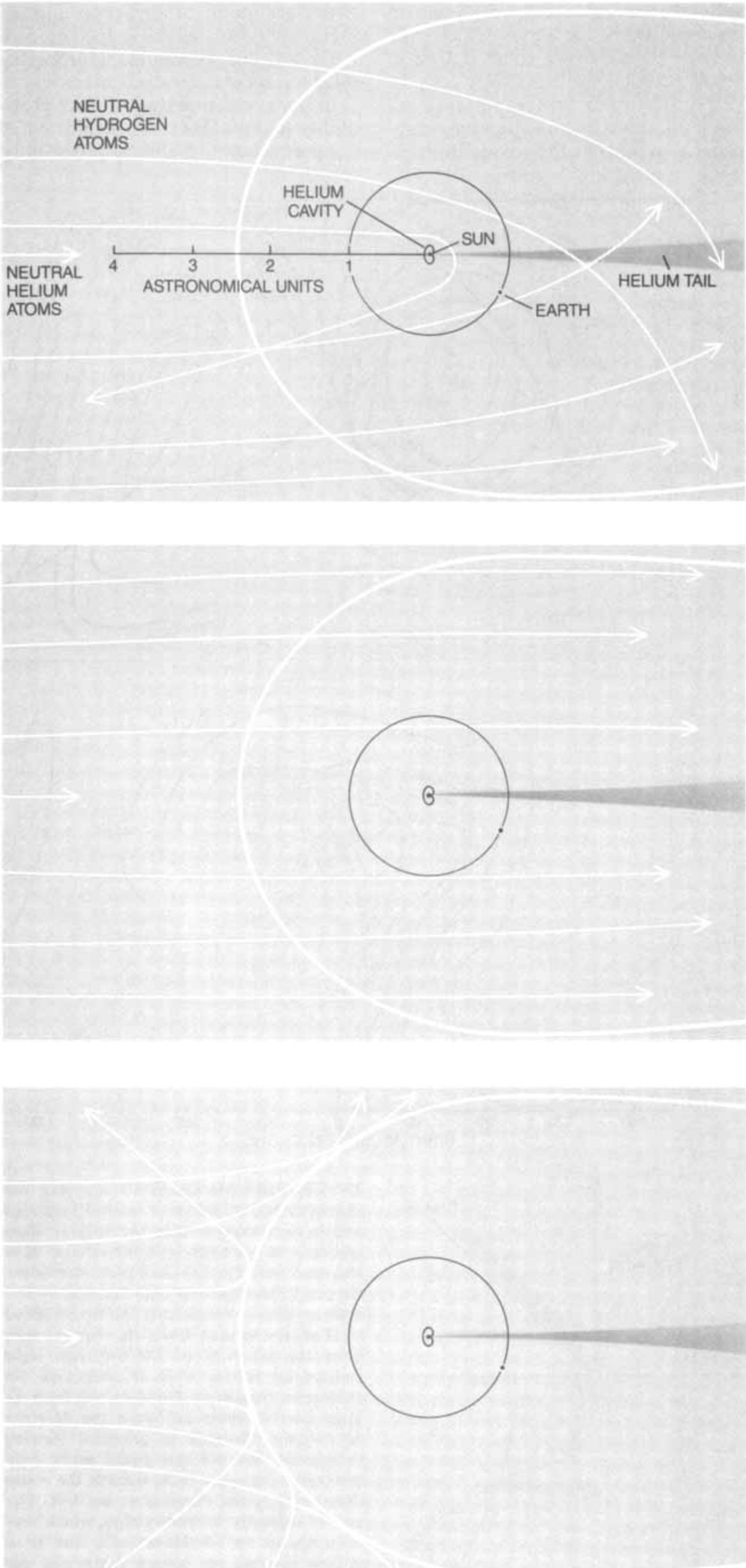
A neutral hydrogen atom traveling at a speed of 10 kilometers per second will take approximately 50 years to traverse 100 AU's, during which time it is under a strong solar influence. Since the solar cycle lasts for only 11 years, the atom will experience all three situations at some point in its approach. In practice the competition between gravitational attraction and radiation pressure will be ragged indeed. An averaging process is therefore applied to model the overall effect.

In any event, there is a distinct probability that an atom will be ionized as it approaches the sun, a probability





INTERSTELLAR CHARTS mapping the characteristic radiation of neutral hydrogen and helium suggest that the solar system contains an abundance of interstellar neutral material. The top map plots the intensity of 1,216-angstrom light emitted by hydrogen when it drops from the first excited level to the ground level; the bottom map plots the intensity of 584-angstrom light emitted by helium when it undergoes the analogous transition. The data are from an ultraviolet detector on board the *Mariner 10* satellite. Because of practical viewing limitations, the detector could probe only the regions of space lying outside the cones illustrated in the diagram at the left. The unit of intensity is the rayleigh, which corresponds to an omnidirectional flux of a million photons per square centimeter per second. (A photon is a quantum of light.)



that increases as the inverse square of the distance. The probability of ionization is also inversely proportional to the radial velocity of the atom with respect to the sun. Different atoms have different rates of ionization; hydrogen, for instance, is more easily ionized than helium. At some point the probability of ionization becomes high enough that essentially all the gas is ionized; the rest of the streamline inside the trajectory is then devoid of neutral species. The locus of such points around the sun defines a region called the solar ionization cavity.

Each interstellar neutral species has its own ionization cavity. In the case of hydrogen the cavity is called an H II region. For helium the cavity is called an He II region. We should point out that the observed H II region is two orders of magnitude smaller than the hypothetical solar Strömgren sphere was predicted to be (.02 light-years, or 1,000 AU's); the observed He II region is an order of magnitude smaller than the H II region. Calculations based on relatively simple models indicate that at about 5 AU's in the upwind direction of the sun the density of neutral hydrogen drops to about half its original value. The density falls to approximately a tenth of its original value at the distance of the earth's orbit.

Even though neutral hydrogen begins to ionize within a cavity approximately 5 to 10 AU's in size, neutral helium remains relatively undisturbed until it is much closer to the sun. Within about .1 AU of the sun helium ionization becomes appreciable and a true ionization cavity exists. Because the cavity is smaller than the earth's orbital distance, the earth is essentially immersed in an undisturbed interstellar gas of neutral helium. At about 10 AU's in the downwind region helium shows an enhancement over its inter-

NEUTRAL ATOMS in the interstellar medium engulfing the solar system are subjected to two opposing forces: the attractive gravitational force of the sun and its repelling radiation pressure. In practice the radiation pressure of the sun is important for the motion of hydrogen only; neutral helium, which is four times as massive as hydrogen, is affected less and is always focused into a region behind the sun. At times of low solar activity gravity dominates radiation pressure and draws neutral hydrogen atoms toward the sun (*top*). During more active phases in the solar cycle the radiation pressure may balance the gravitational attraction. The atoms then stream through the solar system in a straight line parallel to their primary direction of flow (*middle*). At times of high solar activity the radiation pressure sweeps hydrogen atoms out of a paraboloidal zone around the sun (*bottom*).

Take the COMPUTER SCIENCE LIBRARY for only $4.95

when you join The Library of Computer and Information Sciences

You simply agree to buy three more books—at handsome discounts—within the next 12 months.

Save $98.90 on this most important, easy-to-use reference set. It's filled with hundreds of examples, case studies, and tips.

PROGRAMMING LANGUAGES, Second Edition.
By Allen Tucker. An encyclopedia of computer languages. Critically analyzes 11 of the most important programming languages including Pascal, FORTRAN, COBOL, SNOBOL, APL, LISP, PROLOG, C, Ada, and Modula-2. Using the extensive case studies and examples, you can match your application to the right language, and implement it in the most efficient manner. 590 pp.
Publisher's price: $36.95

AN INTRODUCTION TO DATA STRUCTURES WITH APPLICATIONS— 2nd Edition
By Jean-Paul Tremblay and Paul G. Sorenson. A comprehensive resource on all aspects of linear and non-linear data structures, featuring examples and hints on applying them to real world applications. 861 pp.
Publisher's price: $37.95

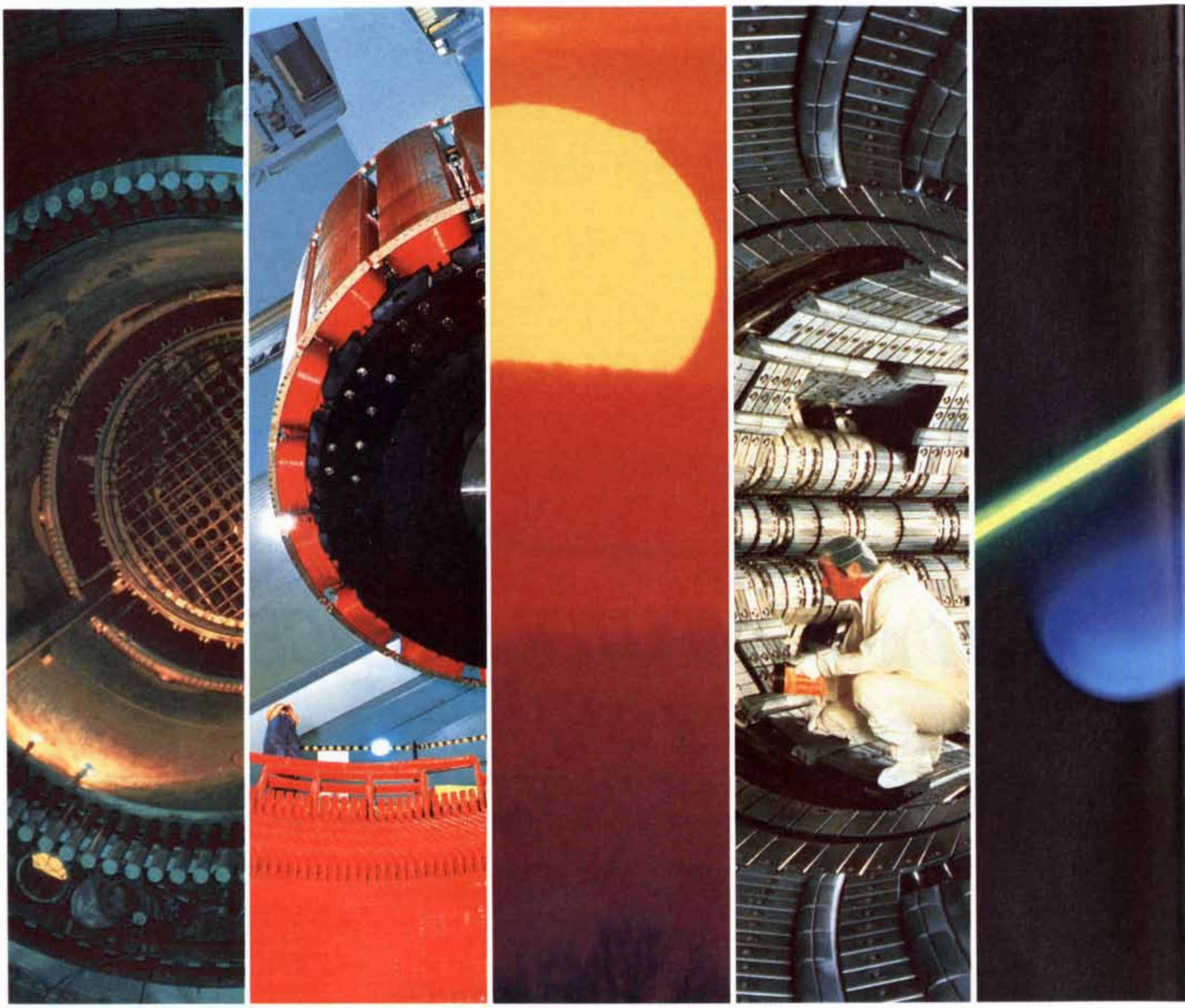
ALGORITHMS
By Robert Sedgewick. A nuts-and-bolts guide to basic problem-solving techniques. Surveys the most important algorithms and data structures including numerical methods, sorting, searching, string processing, geometric methods, graphs, and more. 552 pp.
Publisher's price: $28.95

The Library of Computer and Information Sciences is the oldest, largest book club especially designed for computer professionals. In the incredibly fast-moving world of data processing, where up-to-the-moment knowledge is essential, we make it easy to keep totally informed on all areas of the information sciences. What's more, our selections offer you discounts of up to 30% or more off publisher's prices.

4 Good Reasons to Join

1. The Finest Books. Of the hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.
2. Big Savings. In addition to getting the Computer Science Library for only $4.95 when you join, you keep saving substantially, up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)
3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 70% off the publishers' prices.
4. Convenient Service. At 3–4 week intervals (16 times per year), you will receive the Library of Computer and Information Sciences News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing, and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If reply card has been removed, please write to **The Library of Computer and Information Sciences, Dept. 7-DH5-00728,** Riverside, N.J. 08075 for membership information and an application.
Scientific American 9/86



Hitachi's wide-ranging technologies in energy (from left to right): nuclear power reactor, generator-motor, solar energy development, nuclear fusion plasma testing device, and laser-test of LPG gas combustion.

stellar value by a factor of five or six. The enhancement comes about from gravitational focusing and manifests itself in the form of a "tail" enriched with neutral helium.

A complete description of the interaction of the sun with the nearby interstellar medium must take into account the fact that the instreaming gas is not composed entirely of neutral material. As the charged particles and the accompanying weak magnetic field in the interstellar medium move toward the sun they encounter the outflowing solar wind, the stream of charged particles emanating from the sun. The subsequent interaction forms a collisionless shock front. The location of the shock front depends primarily on the density of the electrons in the cloud and on the strength of the associated magnetic field. Although both parameters are poorly known, the estimate of the location is not extremely sensitive to either of them. Depending on the values chosen for the electron density and the magnetic field strength, the shock front might be anywhere from about 75 to 200 AU's upwind of the sun.

Immediately inside the shock front and having a thickness of perhaps 50 AU's there is a turbulent "shell" of plasma: the heliospheric interface region. The interface is drawn into an elongated teardrop extending perhaps 1,000 AU's downwind of the sun, where the interstellar magnetic field finally pinches it off. The outer surface of the interface is called the heliopause; the inner surface of the interface bounds a region called the heliosphere. Inside the heliospheric interface region the solar wind is deflected into the downwind wake.

As a result of a variety of observations carried out with spaceborne instrumentation, the speed, direction of motion, temperature and density of the inflowing interstellar medium are now reasonably well established. The bulk speed of the cloud is one of the most difficult parameters to determine. The best results have been obtained by measuring the spectral shift of the hydrogen 1,216-angstrom light. Just as the pitch of a train whistle changes from high to low when the train passes a stationary observer, the wavelength of light emitted by a moving atom also changes. Such changes are known as Doppler shifts. In 1977 Thomas F. Adams and Priscilla C. Frisch of the University of Chicago measured the Doppler shift of the hydrogen 1,216-angstrom light from the instreaming gas by means of a spectrometer on board the satellite *Copernicus.* From the data they calculated that the interstellar cloud is moving at a speed of about 22 kilometers per second, or approximately 50,000 miles per hour, in relation to the sun.

In 1984 J. T. Clarke and one of us (Bowyer) at Berkeley, in collaboration with Fahr and G. Lay of the University of Bonn, employed the same approach but with a spectrometer on board the space probe *International Ultraviolet Explorer.* We calculated that the material is moving at a somewhat higher speed of 26 kilometers per second in relation to the sun. The difference between these two results serves as a roughly indicative measure of the uncertainty of the true value of the bulk speed.

Investigators have deduced the direction in which the cloud is moving with respect to the sun by observing, from different locations on the earth's orbit around the sun, where the ultraviolet light is brightest and where it is faintest. Only one flow direction can account for the observed patterns: the material is drifting toward the direction corresponding to the astronomical coordinates right ascension 72 degrees, declination +15 degrees. In terms of astronomical landmarks, as we have mentioned, the cloud appears to be coming from the direction of the constellation Centaurus and drifting toward the direction of the constellation Cassiopeia.

An increasing amount of evidence suggests that there is significance to the direction in which the material is moving. Astronomers have long studied an expanding bubble of gas, commonly referred to as the North Polar Spur, which vaults a large area of the sky encompassing the constellations Ophiuchus, Scorpius and Centaurus. The constellations are about 500 light-years from the earth, roughly in the direction of the galactic center. A variety of observations indicate that the bubble is in fact expanding from a point close to the center of the Scorpius-Centaurus association of stars. The bubble is believed to have originated in an energetic event that took place some one to 10 million years ago in the Scorpius-Centaurus association. Either massive winds from hot stars or the sudden explosion of one or more supernovas would have been sufficient to produce the observed effects. In this context it is quite natural to suggest that the observed bulk-flow direction of the interstellar gas surrounding the solar system is an effect of the Scorpius-Centaurus event; the material may actually be part of the shell of interstellar matter swept along by an expanding shock front emanating from this region.

Other parameters of the local interstellar medium that have been deduced are its temperature and density. The temperature is about 15,000 degrees Kelvin. To measure it, absorption cells filled with hydrogen or helium gas were mounted in front of detectors that are respectively sensitive to the hydrogen 1,216-angstrom light and the helium 584-angstrom light. The cells act as filters that remove a precisely defined segment of the incoming radiation and hence provide detailed information about the emission. In 1977 a team of French and Russian workers led by Bertaux and V. G. Kurt of the Institute of Cosmic Investigation in Moscow flew a hydrogen-absorption cell on a Soviet spacecraft and obtained measurements of the temperature of the inflowing hydrogen.

The analogous approach for helium presents special difficulties because the only materials transparent to helium 584-angstrom radiation (and hence suitable as windows for the absorption cells) are thin metallic foils only a few hundred angstroms thick—less than a thousandth of the thickness of a human hair. Yet the foils must withstand a substantial pressure differential that is typically about a torr (a thousandth of the atmospheric pressure near the surface of the earth) on one side and a vacuum on the other side. In the early 1970's we successfully designed a number of helium-absorption cells with Jay Freeman, who was then a graduate student at Berkeley, and flew the cells on a rocket in 1974 and on the *Apollo Soyuz* test flight in 1975. In 1983 the Bertaux and Kurt group published an elegant analysis of an extensive data set obtained with helium-absorption cells flown on two Soviet spacecraft.

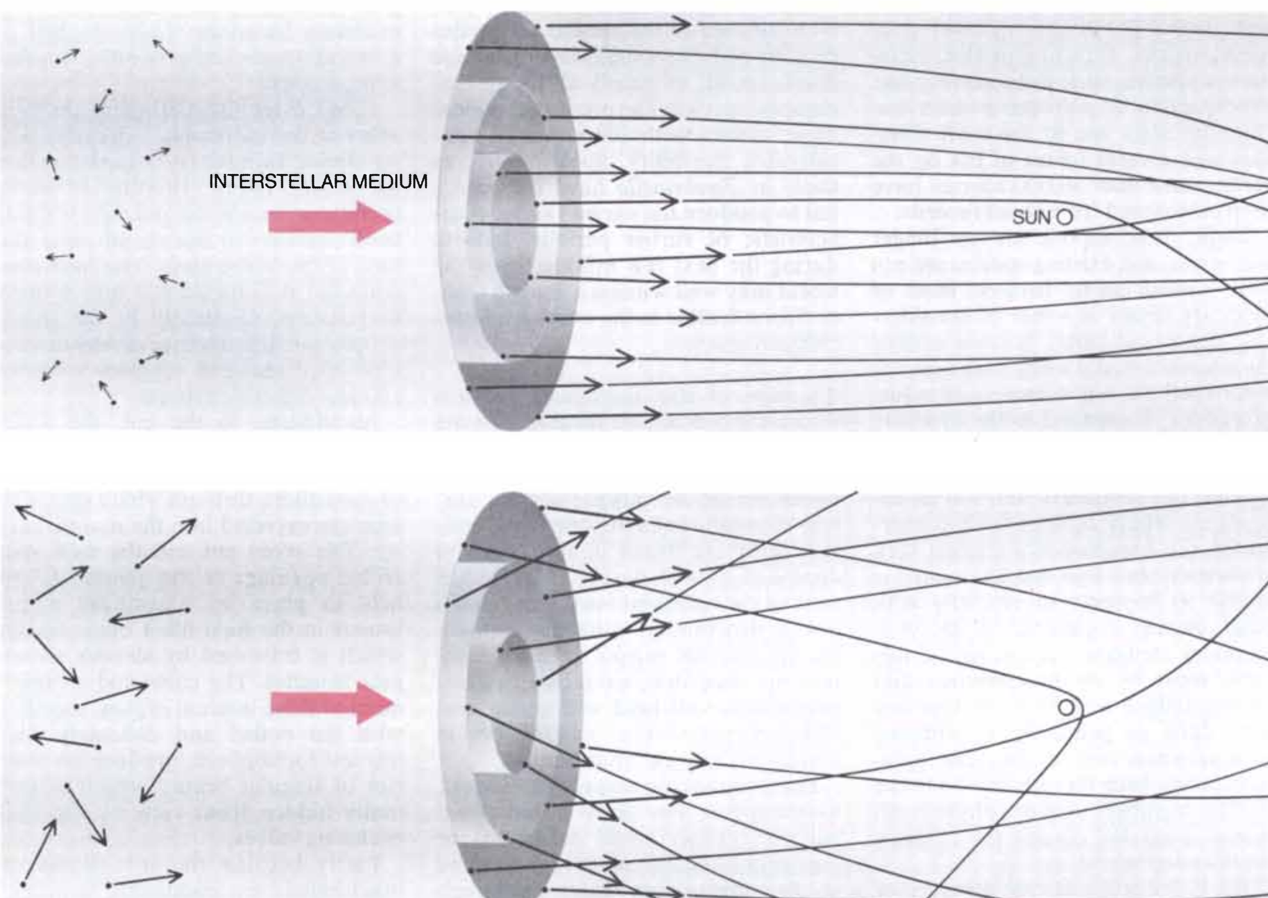
The densities of neutral hydrogen and helium at great distances from the sun can be determined by measuring the absolute intensity of the hydrogen 1,216-angstrom radiation and the helium 584-angstrom radiation. Exploiting such techniques, the density of hydrogen is found to be about .06 atom per cubic centimeter and the density of helium is close to .014 atom per cubic centimeter.

Consequently the ratio of hydrogen to helium contained within the heliosphere is about five, compared with the cosmic, or average, ratio in the universe, which is 10. Astronomers were initially puzzled by the discrepancy between the two ratios. To account for the difference, Fahr and H. W. Ripken of the University of Bonn have suggested that only about half of the original interstellar hydrogen atoms survive the journey through the turbulent heliospheric interface region. With this correction the actual undisturbed interstellar hydrogen density is about .12 atom per cubic centimeter, which brings the ratio of hydrogen to helium to the standard cosmic ratio of 10 and provides an answer to the puzzle.

It is intriguing to speculate what would happen if the sun were to pass through a much denser region of the interstellar medium. After all, clouds with hydrogen densities of 10 to 1,000 atoms per cubic centimeter are found quite commonly throughout intergalactic space. A reasonable estimate is that in its lifetime the sun has encountered at least 100 clouds with densities greater than 100 atoms per cubic centimeter and at least 10 clouds with densities greater than 1,000 atoms per cubic centimeter.

During an encounter with a dense cloud the pressure of the interstellar medium would increase, pushing the solar-wind shock front closer to the sun and compressing the heliosphere. As long as the shock front lies well outside the earth's orbit, the solar wind shields the earth from any direct involvement with the interstellar environment. At some critical density (in some models as low as 150 atoms per cubic centimeter), however, the shock front would be driven within the orbit of the earth. At this point the earth would spend much of its time in the harsh environment of the interstellar cloud. Most hydrogen in dense interstellar clouds is molecular rather than atomic in form. During the 100,000-year traversal time of a typical dense interstellar cloud the atmosphere of the earth would accrete an amount of molecular hydrogen equal to 2,000 times the present amount.

What would the earth do with a large excess of molecular hydrogen? A substantial amount of work has been devoted to this question, and the situation is quite complex. The current thought is that the hydrogen would react with radicals in the upper atmosphere to produce a downward flux of water vapor. The water vapor would condense into clouds. It seems likely that the clouds would reflect a large amount of solar radiation back into space. This in turn would depress the average surface temperature of the earth by several degrees. Although the change may seem negligible, if it were sustained for several thousand years it could trigger an ice age. Even though this hypothesis may be extreme, a substantial climatic effect due to an encounter with a dense molecular cloud seems quite plausible. Future generations will almost certainly have an opportunity to test this hypothesis by direct observation when the solar system encounters a dense interstellar cloud sometime in the next few hundreds of thousands of years.



EFFECT OF TEMPERATURE on the distribution of interstellar neutral atoms around the sun is illustrated for helium. At temperatures near zero degrees Kelvin (−273 degrees Celsius) the typical thermal speeds are negligible with respect to the speed of the cloud, and each particle approaches the sun from the same direction and at the same speed (*top*). This results in an acute clustering of trajectories in the wake. At temperatures of several thousand K. the thermal speeds are comparable to the bulk speed and the flow axis is not as well defined (*bottom*). Consequently there is a blurring of the wake. Here the effect has been exaggerated for the sake of clarity.

Brachiopods

There are two classes of these clamlike creatures. One survives by searching out environments suited to an unchanging form, the other by adapting its form or behavior to the local environment

by Joyce R. Richardson

Paleontologists examining sedimentary formations dating from the Paleozoic era (from 600 to 220 million years ago) often find pairs of fossilized valves, or shells, that superficially look like the remains of bivalve mollusks (such as clams, mussels or scallops). Closer inspection reveals that, unlike mollusks, these fossilized creatures have valves of unequal size and a very different arrangement of internal organs. They are, in fact, members of an altogether different phylum: Brachiopoda. In prehistoric times brachiopods were one of the most abundant and diverse forms of life on the earth: more than 30,000 species have been catalogued from fossil records.

Today brachiopods are no longer numerous, and existing species are not well studied, partly because most of them are found in water of considerable depth and partly because neither the animal's fleshy inner tissue nor its outer shell has any commercial value. Moreover, in contrast to the diversity of the extinct species, the surviving brachiopod lineages (some 300 known species) are singularly uniform in appearance. Many zoologists have interpreted the brachiopod's current lack of diversity as a sign that the animal is unable to compete successfully with other marine organisms in the evolutionary struggle. According to this view, most of the brachiopods that have managed to survive to this day have done so primarily by clinging to underwater cliffs in shoreline habitats, having been literally driven to the wall by members of other phyla (such as mollusks) that occupy the offshore sea-floor sediments.

Yet if longevity is any measure of success, then brachiopods are the most successful organisms extant. The genus *Lingula,* for example, has an unbroken fossil record extending over more than half a billion years to the present day. Not even the major environmental disruption that claimed 90 percent of all marine-animal species toward the end of the Paleozoic seems to have affected it. Recent studies in southern oceans also call for a revision of the conventional view that brachiopods are an evolutionarily stagnant group on the brink of extinction. Species such as those in the genus *Terebratella* can occupy a wide range of habitats. This capability can in large part be attributed to the special properties of their pedicle, a stalklike appendage characteristic of nearly all living brachiopod species. The properties endow these species with life styles of considerable flexibility. Species such as those in *Terebratella* have the potential to produce the varied faunas characteristic of earlier periods. Indeed, during the next few million years the world may well witness a growth rather than a decline in the number of brachiopod species.

In spite of the differences in form among species, all members of the phylum Brachiopoda share a basic physiology that allows them to live in ocean waters of varying depths, current strengths and temperatures. Similar forms are found in soft oozes at depths of 4,000 meters (13,000 feet) as well as on intertidal rock faces. Such a wide distribution in habitat is possible because the organs of respiration, feeding, digestion, excretion and reproduction—all held within the brachiopod's two valves—can function in almost any marine environment.

Brachiopods are suspension feeders, meaning that they ingest minute food particles (either small organisms or particulate organic matter) suspended in the surrounding water. Although this entails processing relatively large quantities of water, the total amount of food needed for survival by individual brachiopods is small, since their metabolic rate is so low. In fact, they can survive more than two years in seawater from which all suspended particles have been removed. (They do not breed under such adverse conditions, however.)

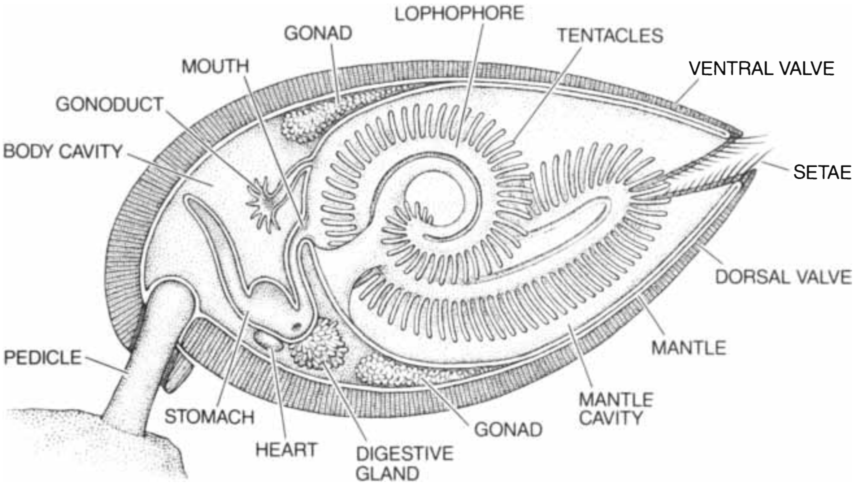
Many of the processes necessary for life take place in one multipurpose organ, the lophophore, which is separated by a membrane (the mantle) from the brachiopod's body cavity, where the internal organs are. The lophophore handles feeding, respiratory and excretory functions; it also doubles as a brood pouch and probably as a fat-storage organ.

Ciliated tentacles fringing the two arms of the lophophore circulate water within the space it occupies, called the mantle cavity. The flow of water facilitates respiration and directs small food particles to the mouth near the base of the lophophore. The food thus collected is consolidated into a mass that rotates constantly in the short, simple gut. The absence of any detectable fecal material suggests a highly efficient digestive process.

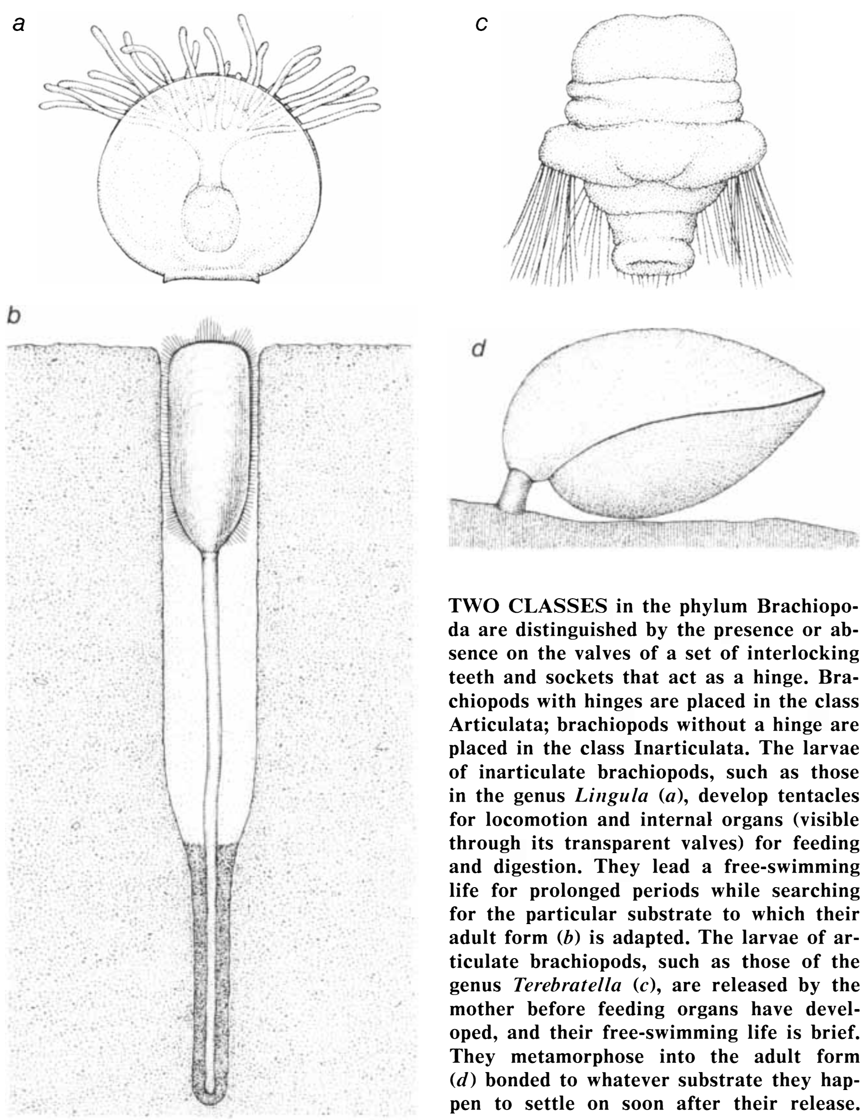
In addition to the gut, the other prominent organs in the brachiopod's body cavity are a tiny heart and a pair of gonoducts, through which sperm or eggs are excreted into the mantle cavity. The green gut and the dark red, frilled openings of the gonoducts are held in place by translucent membranes in the fluid-filled body cavity, which is traversed by slender silver-pink muscles. The color and arrangement of these internal organs, together with the coiled and delicately filamented lophophore, produce an interior of singular beauty, which is normally hidden from view by the two enclosing valves.

Partly because the soft tissues of brachiopods are inacessible to potential predators (except perhaps starfish) when the valves are closed and the pedicle is retracted, the creatures are rarely disturbed by marine animals. Another reason they are not sought as food may be that the little flesh they do have has a particularly unpleasant

BRACHIOPODS were thought to be primarily inhabitants of subtidal rocky cliffs, as exemplified by such species as *Liothyrella uva* (*top*), here shown at a depth of five meters (16 feet) on a rock surface off the coast of Signy Island in the South Orkneys near Antarctica. Closer studies in southern oceans, however, have revealed that they can inhabit a wide variety of environments. *Neothyris lenticularis* (*bottom*), for example, can live among the shell fragments on the muddy sea floor 40 meters (130 feet) under the waters of Paterson Inlet on Stewart Island, New Zealand. A common physiology coupled with adaptations of the pedicle, the stalklike organ of attachment, enables these creatures to live on different substrates in waters of various depths, temperatures and current strengths.



GENERALIZED ANATOMY of a brachiopod is shown in this cross section cut along its plane of symmetry. The most prominent organ, the lophophore, lies in the mantle cavity and is separated from the body cavity, which encloses the other organs. Movements of the lophophore's ciliated tentacles generate water currents inside the two valves, or shells. These internal currents sweep food particles suspended in the water toward the mouth while directing waste products out of the mantle cavity. The lophophore also serves as the brachiopod's organ of respiration, as a brood pouch for larvae and probably as a fat-storage organ.



TWO CLASSES in the phylum Brachiopoda are distinguished by the presence or absence on the valves of a set of interlocking teeth and sockets that act as a hinge. Brachiopods with hinges are placed in the class Articulata; brachiopods without a hinge are placed in the class Inarticulata. The larvae of inarticulate brachiopods, such as those in the genus *Lingula* (*a*), develop tentacles for locomotion and internal organs (visible through its transparent valves) for feeding and digestion. They lead a free-swimming life for prolonged periods while searching for the particular substrate to which their adult form (*b*) is adapted. The larvae of articulate brachiopods, such as those of the genus *Terebratella* (*c*), are released by the mother before feeding organs have developed, and their free-swimming life is brief. They metamorphose into the adult form (*d*) bonded to whatever substrate they happen to settle on soon after their release.

taste. (It seems to be as unpalatable to fish as it is to my colleagues and me: in both field and laboratory situations fresh brachiopod tissue is rejected by fish, even in the absence of alternative foods.)

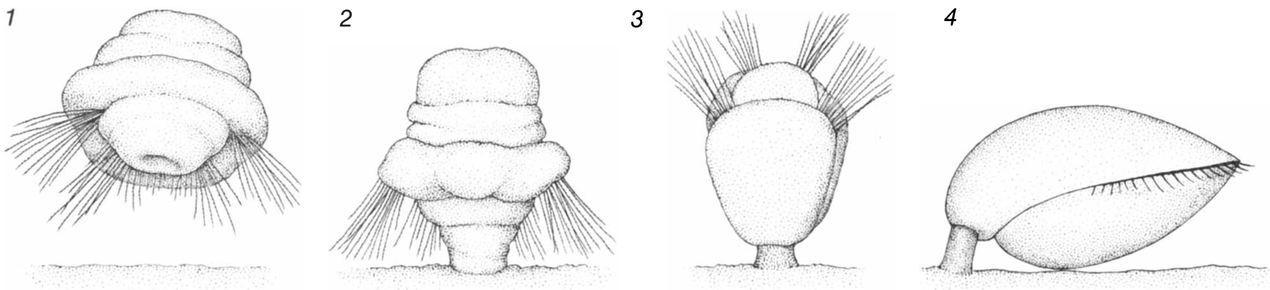
Reproduction is sexual among brachiopods, the sperm being released by the male of the species into the seawater and carried to the female in the feeding currents. The larvae are brooded in the lophophore before they are also released into the surrounding water. (As far as we know, brachiopods do not exhibit any social behavior. Nevertheless, since any tendency to disperse would limit the transmission of sperm to female members of the population, brachiopods tend to cluster in "communities.")

Although brachiopods are fundamentally similar in anatomy and physiology, two classes can be distinguished within the phylum: Articulata and Inarticulata. As the names imply, the Articulata included species (such as those in the genus *Terebratella*) whose valves are hinged together by a set of interlocking teeth and sockets. Members of the class Inarticulata (such as species in the genus *Lingula*) lack such a hinge arrangement; their valves are held in place by muscles. There are other characteristic differences as well, primarily in larval behavior and in the nature of the pedicle.

An inarticulate larva is able to feed and maintain itself for periods of six weeks or longer while it searches for a suitable substrate on which to settle. Once it has found a substrate, the developing pedicle uncoils and emerges from between the larval valves to bond with the underlying surface. The larval life of an articulate brachiopod is brief, on the other hand, and settlement (marked by the bonding of the incipient pedicle of the larva to the substrate) precedes metamorphosis and the development of feeding, digestive and other maintenance systems. In both cases the attachment of the pedicle to a substrate is necessary for the brachiopod to develop into an adult form; if no bond is established, the brachiopod dies while it is still in the larval stage.

The pedicles of articulates vary in both form and composition. In contrast, only two pedicle forms are known for inarticulate brachiopods: a short form, characteristic of species that live on hard substrates (the Acrotretids), and a long form, characteristic of those species that live on unconsolidated sediments (the Lingulids).

The reason for the limited pedicle forms of the inarticulates can be found in the way the organ develops in the



LIFE CYCLE of an articulate brachiopod begins when it is released as a trilobate (three-lobed), free-swimming larva (*1*) from the mother's lophophore. The larva soon settles on or near members of the parental population (*2*). Settlement is accompanied by the bonding of the incipient pedicle (which develops from the hind lobe of the larva) to a substrate. The setae, or bristles, protruding from the larva's underside may aid in this process. Metamorphosis follows bonding, the internal organs and lophophore developing from one lobe and the mantle and valves from another (*3*). In the adult form (*4*) the pedicle and its substrate function jointly as one appendage.

larva. The inarticulate larva consists of two lobes, one developing into the adult body and pedicle and the other developing into the valves. The pedicle develops from a pouchlike outgrowth of the body while the inarticulate is still in the free-swimming larval stage. In contrast, the articulate larva has three lobes. One of them turns into the pedicle on settlement, before the other two can differentiate to become valves, mantle and internal organs. Hence the pedicle and its muscles develop in an articulate brachiopod as a distinct unit separate from the body.

Originally it had been thought that the pedicle was solely an organ for attachment, much like the stem and roots of plants. This misconception arose from the fact that studies of living brachiopods in their natural environments were made on rocks and reefs close to the shore. In such habitats brachiopods do seem to exhibit a sedentary life style. They were therefore generally regarded as being inshore animals, permanently attached to rocky cliffs or outcrops. Those specimens found loose on the sea floor were thought to have been cut loose by misadventure from the rocks on which they had first settled. It was not until marine biologists began to study bottom-dwelling faunas in southern oceans more closely that the pedicle's main purpose became clear: it enables the brachiopod to adjust its position in relation to its surroundings, even on the sea floor.

Unlike more familiar organs of motion such as feet, wings or fins, the pedicle is not a free appendage. From the time of bonding (in the larval stage) onward, the pedicle and the substrate function as one system, regardless of the substrate's size. The brachiopod's motile system therefore consists of the pedicle itself, a cluster of muscles attaching one end of the pedicle to the inner surfaces of the valves, and a substrate bonded to the other end of the pedicle.

Pressure on the valve surfaces (such as that from accumulated sediment) stimulates contraction of the pedicle muscles, which in turn cause movement either of the brachiopod itself or of the pedicle and substrate, depending on the relative masses of the animal and the substrate. For individuals bonded to large substrates, the pedicle functions as a pivot and the shell rotates about it; in the case of brachiopods bonded to small masses, contractions of the pedicle muscles instead move the pedicle and along with it the substrate. If the masses of the brachiopod and the substrate are about equal, then either or both can move depending on the immediate conditions, in particular whether more sediment lies on one body than on the other. Regardless of which mass moves, the result is the same: sediment is shed from the shell surface.

Although there is no difference in the way the pedicle system operates in cliff-hanging and in bottom-dwelling brachiopods, the way the muscle contractions are manifested and the amplitude of motion may differ. In cliff-hangers, where the substrate is typically a firm rock surface, the muscular activity causes the brachiopod to rotate from side to side. On the other hand, bottom dwellers, which are generally attached to bits of gravel or small shell fragments, twist the pedicle-substrate system one way to cause a slight twisting motion of the brachiopod in the opposite direction. These twisting movements may be weak and may seem undirected, but they are as effective in preventing the buildup of sediment on the brachiopod as the stronger rotatory movements of brachiopods fixed to large masses.

In all probability the only reason for these peculiar movements is that they keep the brachiopod from being buried and so maintain its access to water in whatever environment it has the capacity to occupy. Because they are suspension-feeding organisms, access to nonturbid water is vitally important to brachiopods. Observations of many living species confirm that the movements that shed sediment automatically result in a favorable feeding position at the interface between the water and the substrate.

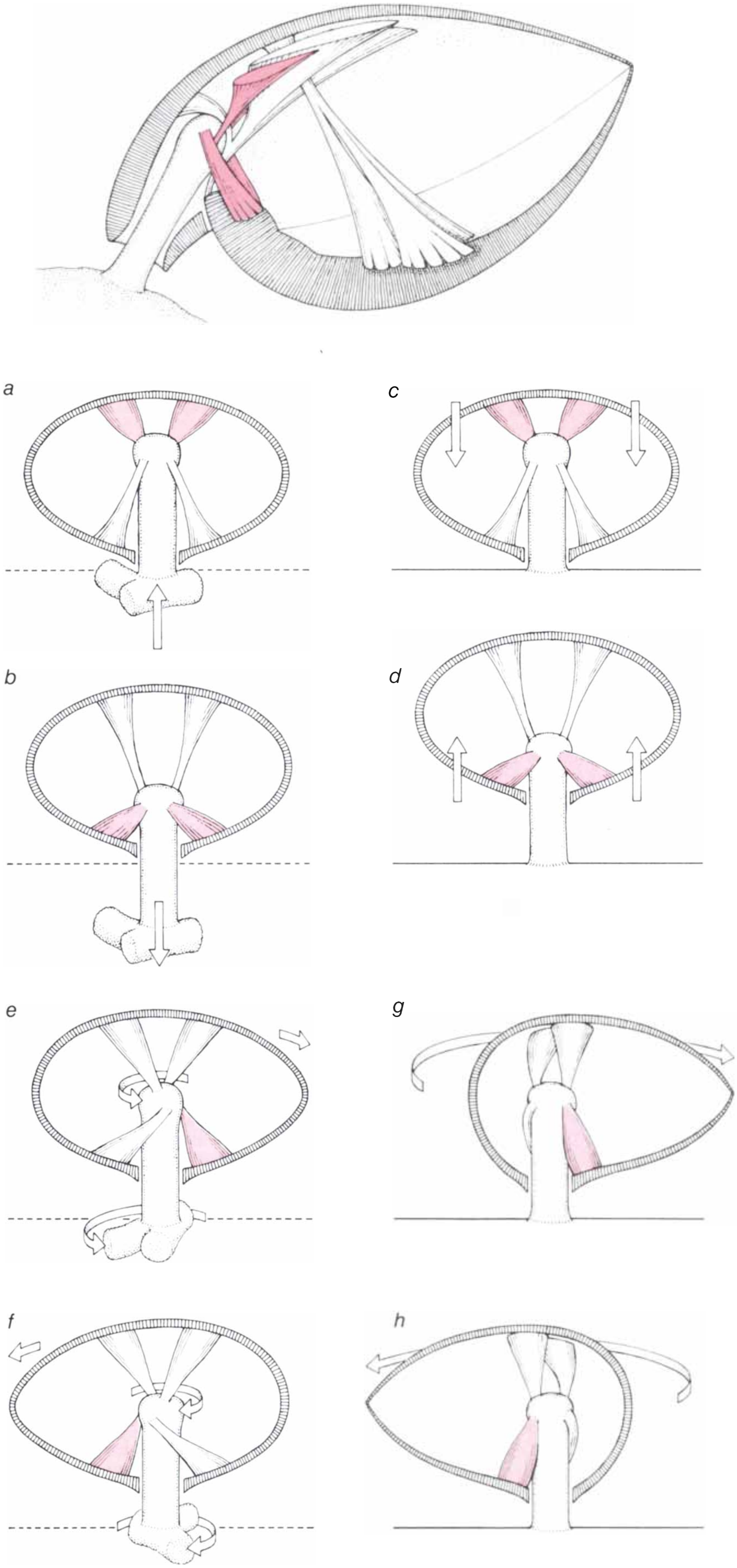
In this context species that can maintain a stable feeding position on substrates of any type are called generalists. Their patterns of behavior may vary as much during the life history of one individual as they do among different members of a population. Other species are labeled specialists because they stick to a uniform behavior that is appropriate for achieving feeding stability only on a limited range of substrates. Specialists can be identified anatomically by various adaptations of the valves, pedicle and muscles that make movement on a particular substrate more efficient but would be incompatible with life on any other type of substrate.

For example, some articulate brachiopods have valves that are weighted at the base like self-righting toys. This differential thickening gives the valves a constant orientation in soft sediments and so facilitates the animal's movement on such substrates. The added weight, however, would restrict free movement on large, solid substrates. By the same token, the pedicles of other specialist species have tended to merge over the course of time with the pedicle muscles into one strongly contractile unit that has an extensive attachment area. Such a design makes movement impossible without a large, hard surface for leverage. Individuals of such species would die before reaching adulthood if they settled on unconsolidated substrates. Generalist species, on the other hand, tend to display less variability in valve shape and pedicle structure. Their more limited morphology nonetheless endows generalists with the capacity to occupy a wide range of environments.

As one looks over the brachiopod fossil record it becomes apparent that most brachiopod lineages have followed a trend toward increased specialization, a trend that culminates in loss of the pedicle. In the absence of a pedicle system the adult organism is immobile, and adjustments of its position in relation to its surroundings is no longer possible. Stability in such a case is achieved by specialized valve shapes and by outgrowths from the valve surfaces.

G. Arthur Cooper and Richard E. Grant of the U.S. National Museum have documented the diversity of brachiopods from earlier eras. The abundant faunas they uncovered from the Glass Mountains of western Texas include numerous articulate forms that lack a pedicle and are therefore highly specialized. Buttressing spines and a cementing secretion are two of the adaptations that endowed some of the extinct nonpediculate species with the capacity to fix themselves firmly onto hard substrates. Other extinct species relied on the weight and flanged shape of their valves as well as on anchoring spines to ensure their stability on soft substrates.

Today, however, all articulate brachiopods (with a few rare exceptions) have pedicles. What happened to the many nonpediculate articulate species of the past? The answer reveals the consequence of specialization: vulnerability to environmental change. The specialization that gradually evolves in response to prolonged colonization



PEDICLE MUSCLES (*dark color*) can be found among other muscles (which open and close the valves) in the posterior section of the brachiopod (*top*). The pairwise contraction (*light color*) of the pedicle muscles (shown here in rear-view cross sections) can produce an up or down motion of either the pedicle and substrate (*a, b*) or of the brachiopod's valves (*c, d*), depending on the size of the substrate. If the pedicle is bonded to a small substrate, contraction of one of the lower pair of pedicle muscles twists the entire pedicle-substrate system, resulting in a weak counterrotating motion of the valves (*e, f*). If the pedicle is attached to a large substrate, by contracting the same muscles a brachiopod can rotate itself strongly (*g, h*). Regardless of whether the brachiopod or the substrate undergoes the stronger motion, such twisting action is effective in shaking off sediment that might bury the animal. By virtue of these simple movements a brachiopod automatically occupies a favorable feeding position at the interface between the water and the substrate.

of a particular substrate depends critically on that substrate. Hence during periods of environmental instability, when the particular substrate to which a specialist species has adapted is no longer available, the species quickly dies out. Generalist species have a greater capacity for riding out environmental change, since they are not dependent on a particular substrate. The rate of extinction of specialist species is related to both the magnitude of environmental change and the degree of specialization.

Environmental disruptions need not be global in extent, like the one at the end of the Paleozoic, to cause the extinction of specialized species. Study of local geographic regions shows that extinctions are not uncommon and that small-scale shifts in geology may be quite devastating. Finn Surlyk of the Geological Survey of Greenland and Marianne Bagge Johansen of the Institute of Historical Geology and Paleontology in Copenhagen have meticulously documented such a local mass extinction in the late Cretaceous. They show that many brachiopod species adapted for chalky substrates found in Danish waters became extinct following a change in sedimentation recorded stratigraphically as a layer of clay overlying the chalk. Of the 35 brachiopod species found in the chalk deposits, only six could be found in the clay—all of them generalists.

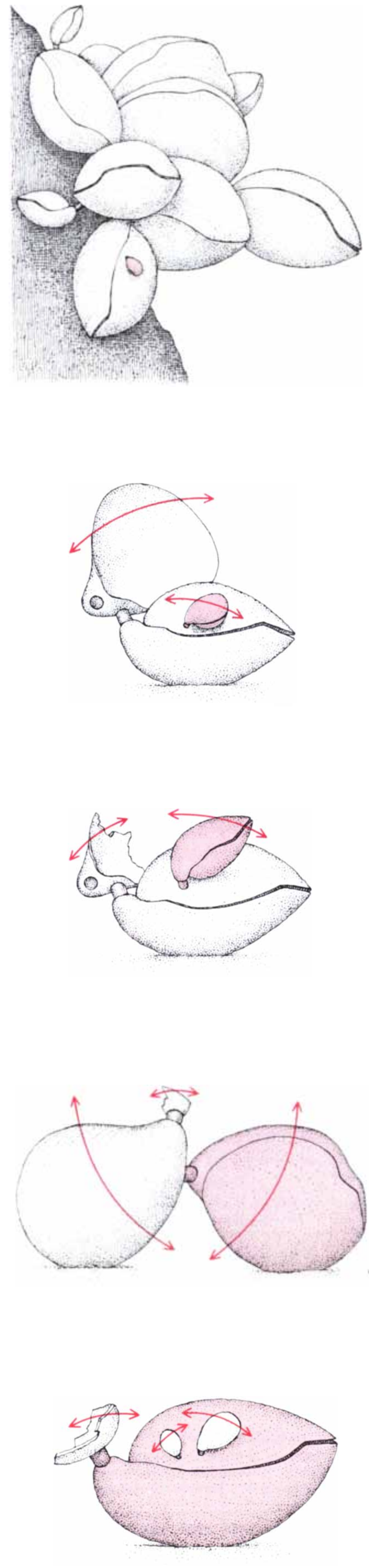
Although periodic extinctions of articulate brachiopods can be explained by the vulnerability of specialized species, a look at the history of inarticulate species would appear to contradict this explanation. Living inarticulate brachiopods are specialized; as I mentioned above, they are adapted either for hard surfaces or for soft muds, but not for both. (In fact, the acrocretid genus *Crania* is characterized by the very acme of specialization: the absence of a pedicle.) Yet they display remarkable longevity and continuity in the fossil record. Whereas the articulate species have fluctuated between periods of specialization (resulting in the formation of new species) and extinction, the inarticulate species have preserved the same specialized forms, and presumably the same life styles, since pre-Paleozoic times. The surviving lineages of both articulates and inarticulates must have achieved invulnerability to environmental change, but, as their strikingly different evolutionary histories attest, that invulnerability has apparently been achieved by completely different methods in the two classes.

The inarticulates, although they are specialists, have survived for eons because their larvae are self-supporting and are therefore not limited to settling in the vicinity of the parental population. If the local environment becomes inhospitable, the inarticulate larvae can move on to other areas in search of the substrate for which their specialized adult forms are adapted.

Among articulates, as I have indicated, it is the generalists that survive: species that can live equally well on any type of substrate. Unlike the larvae of inarticulate species, articulate larvae settle at random directly on or near parental populations and must depend on their adaptability to carry them through periods of environmental change. The continuance of articulate brachiopods therefore depends critically on the maintenance of enough generalist species. This is ensured if varied substrates are present within the larval dispersal range (as is the case in coastal areas), so that long-term colonization of a uniform substrate does not occur and the ensuing specialization is prevented.

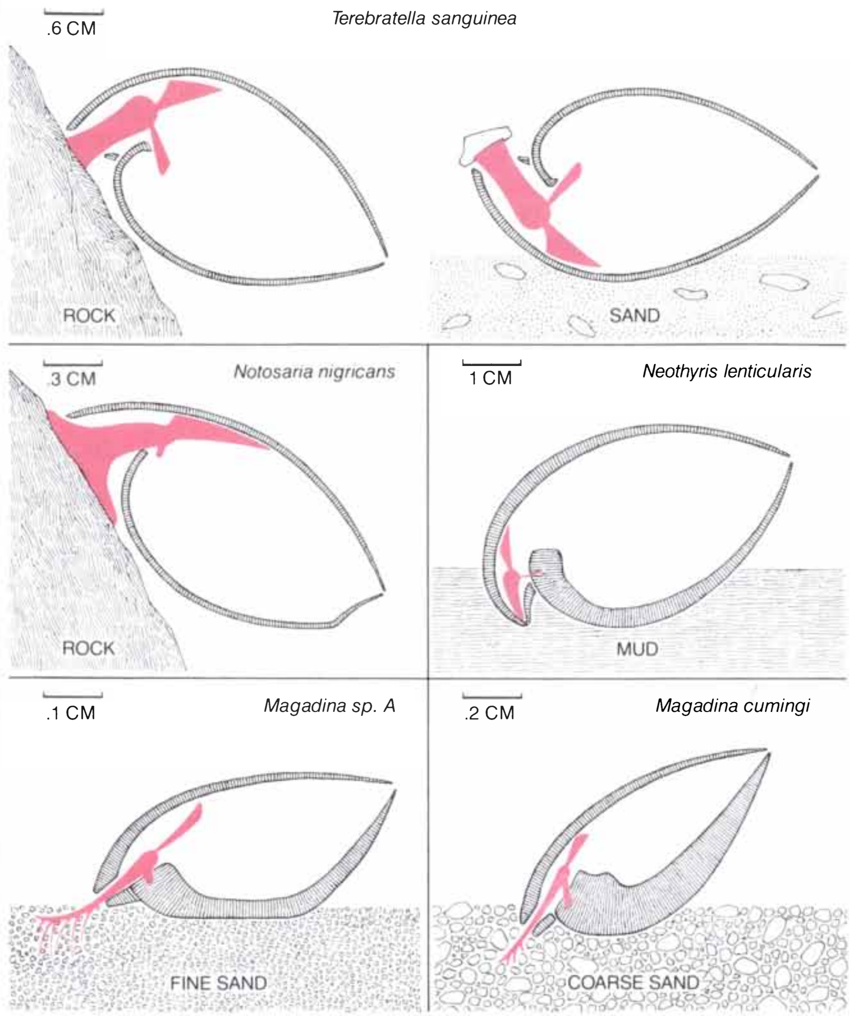
Given this view of the factors that govern the continuance of brachiopod species, it seems unlikely that the phylum is close to being extinguished. A study of Antarctic and subantarctic seas led by Merrill W. Foster of Bradley University showed that articulate brachiopods are common, sometimes even dominant, members of marine communities there. They are also prominent among the marine faunas of South America, Australia and New Zealand.

Some of the brachiopods in those waters are generalists; others have pedicles modified to some degree for



ADAPTABILITY of certain articulate species is illustrated by the changes in behavior that one such individual (*color*) exhibits in the course of a lifetime (*top to bottom*). Although the young brachiopod begins life as a sedentary cliff-hanger, when its substrate (in this case a member of the parental population) breaks off the cliff wall and falls, it can readily adapt to the different life style required of a bottom dweller. During the rest of its life the creature is subject to changes in position resulting from changes in its own size (through growth) and that of its substrate (through disintegration). As the relative sizes of the brachiopod and substrate vary, the developing brachiopod must modify its behavior: first it rotates itself to shake off sediment, but later it must twist the substrate to accomplish the same thing. When the brachiopod and the bonded substrate are approximately equal in mass, then either can move, depending on whether sediment has accumulated more on one than on the other. The lengths of the arrows indicate the relative magnitudes of the movements of the brachiopod and the substrate.

life on either soft or hard substrates. Given sufficient long-term stability of the local area they occupy, species more highly adapted for particular regimes should evolve. Trends toward the loss of the pedicle in muddy sediments and toward its modification as a free appendage in carbonate sands are already evident in two lineages. Nevertheless, most southern brachiopods are little modified for specific regimes. Most of them have characteristics of generalists or of nascent specialists, which are in the process of colonizing new habitats. They do not appear to have the characteristics of a biologically monotonous group that is heading toward extinction.



GENERALIST AND SPECIALIST articulate species are distinguished by anatomical features such as the shape of the valves and the structure of the pedicle system (*color*). Generalist species such as *Terebratella sanguinea* can live either as cliff-hangers or bottom dwellers. In either situation the motion generated by the pedicle system allows members of the species to shed sediment, regardless of orientation. In specialist species the pedicle system and valves have adapted to suit particular substrates. In *Notosaria nigricans,* for example, the pedicle muscles have fused with the pedicle to form one stout contractile unit with a large attachment area. Such an adaptation allows *N. nigricans* to live on large rock surfaces, but the pedicle is ineffective on small, unstable surfaces. The pedicle of *Neothyris lenticularis,* in contrast to that of *Notosaria nigricans,* has atrophied over millenniums, making it useless for attachment to a large substrate. Yet the size, shape and thickness of the lower valve alone enable the species to maintain a stable feeding position in muddy sediments. Two other specialist species, *Magadina sp. A* and *M. cumingi,* have also evolved peculiar valve shapes and pedicles that enable them to successfully colonize sea floors composed of carbonate sands. These two species are unique in that their pedicle pushes them in a ratchetlike manner through overlying sediment. The larvae of these species bond to grains of sand or small shell fragments, but processes that later develop at the pedicle's tip free the adult from the bonded substrate. In periods of environmental change generalist species are more likely to survive than specialists: they are not limited to a particular substrate type.

“The best presentation yet of ‘The New Biology’”*

MOLECULAR CELL BIOLOGY

James Darnell, ROCKEFELLER UNIVERSITY, and
Harvey Lodish and David Baltimore, both at the
WHITEHEAD INSTITUTE and M.I.T.

“The most advanced and the most wide ranging in topics covered.... Will certainly be used by everyone.”
—Steven Ackerman, UNIVERSITY OF MASSACHUSETTS

This is the first book to unify cell biology with genetics and biochemistry. Written by three world-famous scientists, MOLECULAR CELL BIOLOGY provides the most up-to-date and comprehensive treatment available of cell biology and molecular genetics—particularly of the interface of cell structure and biochemistry.

“Excellent, very thorough and with the type of both experimental and integrated approaches which cannot be found [elsewhere] and which are essential to convey the dynamic aspects of this fast growing discipline.”
—M. O. Krause, UNIVERSITY OF NEW BRUNSWICK

MOLECULAR CELL BIOLOGY offers an extraordinarily coherent view of fundamental principles, as well as the key experiments and techniques that led to their formation—an approach that enables readers to see how biological progress is connected to experimental results.

“Extremely well written and illustrated. Excellent use of historical materials and clear description of important experiments and experimental techniques.”
—Donald J. Roufa, KANSAS STATE UNIVERSITY

The quality of the writing is matched by that of the book’s 1,248 illustrations. Like those in *Scientific American,* the illustrations in MOLECULAR CELL BIOLOGY are closely integrated with the text and are accompanied by clear, explanatory captions. Together, the art and captions serve as a virtual ‘short course’ through the book.

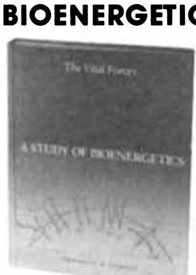
“A clear and excellent presentation....”
—Ralph Greenspan, PRINCETON UNIVERSITY

*Paul R. Elliott, FLORIDA STATE UNIVERSITY

1986, 1,187 pages, 1,248 illustrations
ISBN 0-7167-1448-5, $42.95 A *Scientific American Book*

Also from Scientific American Books/W. H. Freeman and Company...

THE VITAL FORCE: A STUDY OF BIOENERGETICS

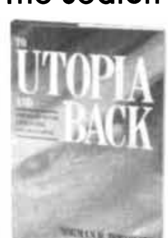


Franklin M. Harold, NATIONAL JEWISH CENTER FOR IMMUNOLOGY AND RESPIRATORY MEDICINE and UNIVERSITY OF COLORADO HEALTH SCIENCES CENTER

The principles of biological energy transduction are now fairly clear, but their significance to cellular and molecular biology is not widely appreciated. THE VITAL FORCE: A STUDY OF BIOENERGETICS makes bioenergetics accessible to a wider audience through an integrated survey of energy transactions in metabolism, membrane transport, movement, and information processing.

1986, 577 pages, 276 illustrations
ISBN 0-7167-1734-4, $37.95

TO UTOPIA AND BACK
The Search for Life in the Solar System



Norman Horowitz

Horowitz, former head of the Jet Propulsion Laboratory’s bioscience section for the Mariner and Viking missions to Mars, takes readers on an illuminating search for life in the solar system. He reports on the microbiological tests for life forms in the Martian environment and what they revealed, extending their lessons to the rest of the solar system. The author also provides a rare inside look at the charged political atmosphere surrounding the six Mariner and Viking flights.

1986, 168 pages, 23 illustrations (8-page color insert)
ISBN 0-7167-1765-4, $17.95

SPECIAL 20% SAVINGS TO READERS OF *SCIENTIFIC AMERICAN*

Please send me:

Quan.	ISBN	Title	List Price	Special Price	Amount
	1448-5	MOLECULAR CELL BIOLOGY	$42.95	$34.59	
	1734-4	THE VITAL FORCE: A STUDY OF BIOENERGETICS	$37.95	$30.59	
	1765-4	TO UTOPIA AND BACK	$17.95	$14.59	
				TOTAL	
				Add $1.50 for shipping and handling, plus $1.00 for each additional book	
				Sales tax (NY, CA, and UT)	
				FINAL TOTAL	

☐ I enclose a check made payable to W. H. Freeman and Company

☐ Charge my:

☐ MasterCard ☐ VISA Account # ____________

Expiration date ____/____ Signature ____________
(Credit card orders must be signed)

Name ____________

Address ____________

City / State / Zip

Mail to: W. H. Freeman and Company
The Book Publishing Arm of *Scientific American* for More Than 20 Years
4419 West 1980 South
Salt Lake City, Utah 84104

Leonardo's Contributions to Theoretical Mechanics

A close look at his visual mode of thinking, particularly in his studies of the crossbow, reveals that his ideas had a bearing on the evolution of four aspects of mechanics

by Vernard Foley and Werner Soedel

It has been argued by some historians of the Renaissance that even though Leonardo da Vinci was talented in many fields, he did not really make worthwhile contributions to such exact sciences as theoretical mechanics. We contend that if one considers recently discovered Leonardo manuscripts and pays close attention to his visual mode of thinking, one can see that his contributions were often substantial. It appears that his work on weaponry, particularly the crossbow, was important in leading his thoughts to these matters. The areas that interested him, to give them their modern names, were the principle of velocity additions, the law of the composition of forces, the concept of the neutral plane and the role played by the center of gravity in a moving body.

Part of the controversy over whether Leonardo worked effectively in the exact sciences stems from the history of his notebooks. After his death they passed into the hands of private collectors, and for a long time access to them was limited. Systematic publication of the notebooks began only at the end of the 19th century. Then for a while Leonardo was seen as the man who had discovered almost everything, centuries before its time.

The fascination with Leonardo also had the effect of focusing interest on his predecessors and contemporaries, however. Soon it was discovered that Leonardo took a great deal from these sources, sometimes even copying their illustrations closely. This finding gave rise to the contention that Leonardo had done nothing original worth mentioning.

The pendulum began to swing the other way with the discovery in Madrid some 20 years ago of two new Leonardo manuscripts, known as the Madrid Codices. Their chief initial student, Ladislao Reti of the University of California at Los Angeles, working with able assistance from the engineer-historian Bern Dibner, pointed out that the manuscripts went far toward restoring Leonardo's reputation as an investigator of natural principles. In order to carry the task further it is important to remember that Leonardo was educated as an apprentice in painting and sculpture and so was chiefly a man of eye and hand. We believe significant additions to the appreciation of him as scientist and engineering theorist can be achieved by putting less emphasis on his words and mathematics and instead giving detailed attention to his sketches.

We begin with an instance that shows Leonardo struggling with weapon-design problems, never fully resolved, that took him into the areas of velocity additions and the composition of forces. Even though gunpowder weapons were evolving rapidly during his lifetime, archers and pikemen were still prominent in the field. Accordingly he devoted much thought to such older weapons as the crossbow. It often happens that a technology reaches its prime only after it comes under pressure from its successor. The process can produce fundamental scientific insights.

Earlier empirical work on crossbows had brought them to a pitch of efficiency even before Leonardo's time. For example, the short bolt (the crossbow's arrow) had achieved about twice the aerodynamic efficiency of traditional handbow arrows. A beginning had also been made in the quest for underlying principles, as one can see from a sketch by an anonymous artist in what is generally called the Hussite Wars manuscript. It shows an attempt to model the crossbow on the form of the circle. The tiller, or stock, of the bow forms a diameter. The nut, or string-release mechanism, falls at the center. The bow sweeps out a third of the circumference. When the string is at rest after a shot, it bisects a radius.

In a sense this picture represents a form of early engineering theory that is worse than having no theory at all. It attempts to subordinate the requirements of functional design to an ideal geometric form. Yet not all the results of this philosophy were bad. Even here there is some glimmering of utility, because a bow will last longer if its curvature is fairly uniform. If that is impossible, the radii of curvature should blend into one another smoothly. Modeling the bow on a circle gets at least the essence of this idea.

In attempting to move beyond the traditional design, Leonardo investigated such ideas as making the arrow shaft fixed and launching only the head. Here he seems to have understood that the reduction of mass in a missile helps to achieve a high exit velocity.

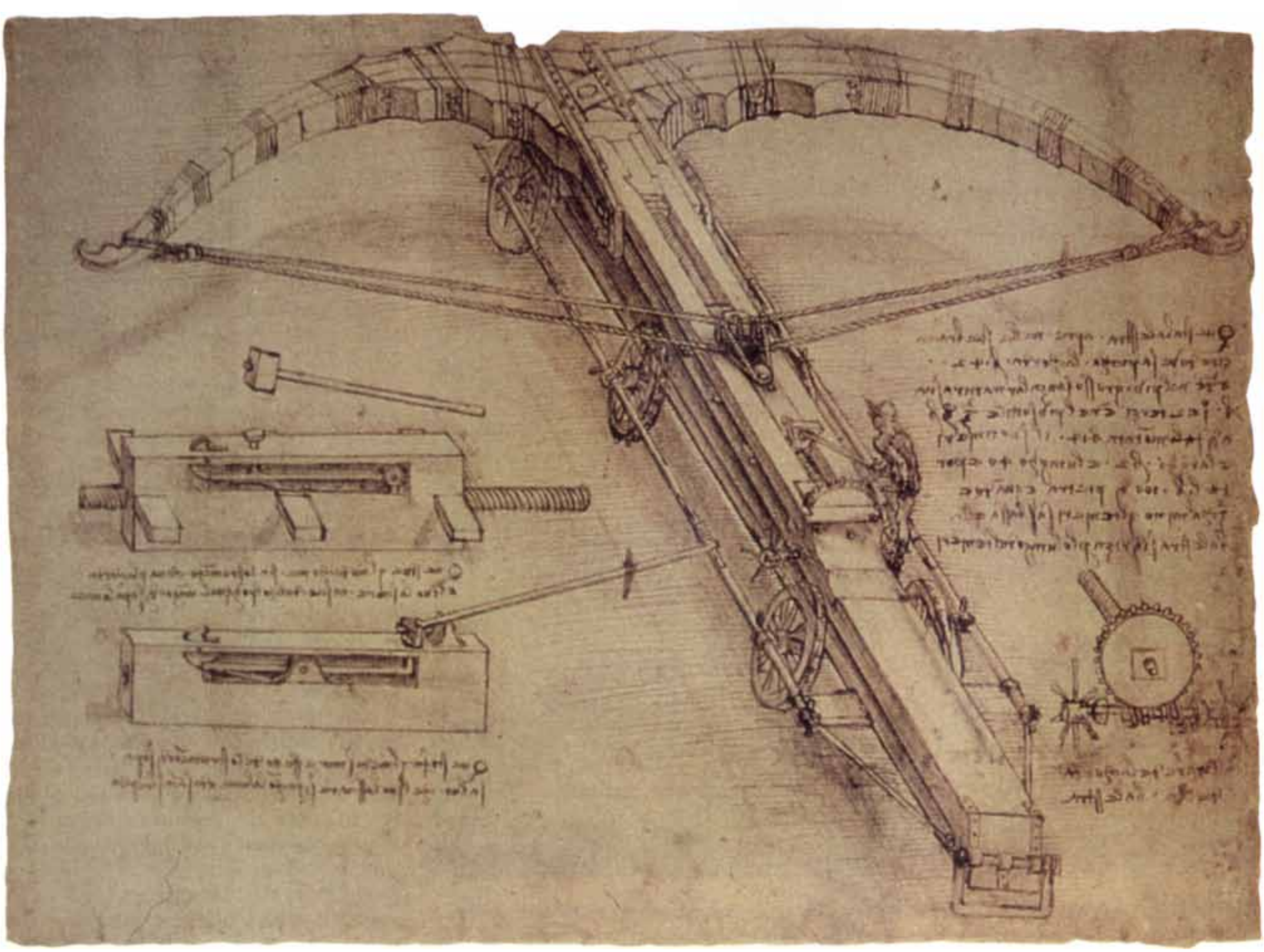
Some of his designs called for multiple bows, working simultaneously in some cases and sequentially in others. In sequential operation a heavy bow would actuate a smaller and lighter one, which would actuate a still smaller one and so on. The last in the series would launch the bolt. Leonardo explicitly thought of this process in terms of additive velocities. He says, for example, that to launch a bolt for maximum range one should shoot forward from a galloping horse and thrust the weapon forward at the moment of loosing the bolt. Actually the combined effect would not greatly increase the velocity of the bolt. Nevertheless, Leonardo's thoughts represent one of the beginnings of the debate over whether one can add velocities indefinitely. Later theorists tended to assume that the process had no limit, until Einstein proposed that the speed of light was an absolute upper bound. At lower speeds the additive principle still holds true.

The law of the composition of forces, also referred to as the notion of the force polygon, was fully established only after Leonardo's time. It concerns the branch of mechanics that asks what happens when two or more forces interact at different angles. In dealing with such problems today one draws rectangles, parallelograms or polygons and solves for the resultants (the diagonals). Leonardo may have had earlier ventures into this matter in the back of his mind when he analyzed crossbows in order to solve a particular design problem. It arises in a process called tillering, which bowmakers employ in preparing the weapon for use.

In making a crossbow it is critical to achieve equal strength in each limb. Inequalities in strength can wrench the bolt sideways out of its launching groove, ruining the accuracy of the shot. Tillering provides the means of checking the strength of the limbs. (Today tillering is done with all bows, but from its name it would seem to have originated with crossbows.) The weapon is mounted on a wall with the string horizontal and the arc of the bow facing upward. Weights of varying size are hung from the center of the string. Each weight holds the bow in a fixed degree of flexure so that the strength of the two limbs can be compared. An easy means of checking is to note whether the center point of the string descends in a vertical line as weight is added.

Leonardo's familiarity with the tillering process may have prompted him to make diagrams (found in Codex Madrid I) that plot the displacement of the bow, with special emphasis on the center point of the string, against the amount of weight applied. He realized that the amount of force needed to draw a bow is low at the outset and increases steadily as the limbs of the bow move. (This anticipates the spring law, formulated much later by Robert Hooke: Force is proportional to displacement.)

Leonardo's term for this relation was "pyramidal" because, like the sides of a pyramid, it began at a point and grew uniformly with displacement of the bow limb. In plotting string position v. draw weight, however, he noted nonlinearities. One is that although force is related to limb displacement in a linear way, it is not so related to string displacement. Applying the consequences of this relation, he seems to have worried that in some bows the string tended to move faster (for a given increment of applied weight) just after release than it did when it was ap-



GIANT STONEBOW was probably sketched by Leonardo in connection with his studies of how to deal with the buildup of tension and compression in a bow. The apparatus was designed to hurl stones. As Leonardo drew it, the front and back of the bow's limbs differ in construction. In particular he shows the back of each limb made with separate blocks, which will touch one another only when the bow is at full draw. The first noticeable failure of an overstrained bow is on the back side, resulting from compression forces.

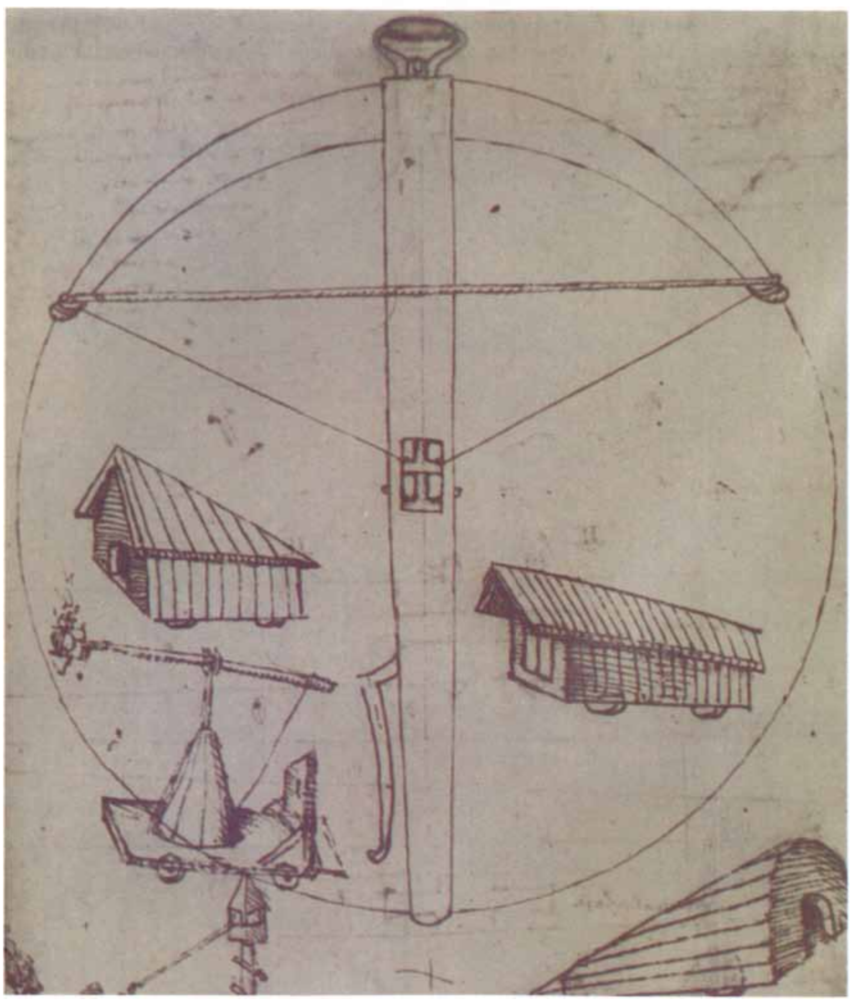
proaching its normal straight position.

This kind of nonlinearity may have been evident in poorly designed bows. Perhaps Leonardo's inference rests on erroneous deduction rather than on measurement, although he sometimes provides numbers. Nevertheless, the problem apparently led him deep into questions of crossbow design. Did the bolt, which accelerated rapidly at the beginning of the shot, actually outrace the lagging string and leave it before the bow had fully relaxed?

Lacking a full understanding of inertia, force and acceleration, Leonardo may not have made up his mind. Definitive statements on both sides can be found in his pages. His concern with the subject, however, led him to a further redesign of the crossbow. It indicates that he was intuitively aware of what would now be called the principle of the composition of forces.

In attempting to grasp and quantify the issue Leonardo flailed about, sometimes wide of the mark. Did a bow that was twice the weight of another shoot twice as far? If the weight of all the bolts needed to reach end to end to the maximum range were determined, would it equal the bow's draw weight? Sometimes he saw deeply indeed, as when he asked whether the vibration of the string after a shot represented a waste of the bow's energy.

Eventually he made a clear and simple statement (found in Codex Madrid I) of the relation between bow force and string displacement: "The power of the mover of the string of the crossbow increases, as much as the angle created at the center of the string decreases." That the remark is not repeated elsewhere may mean Leonardo stuck with it. Certainly he embodied the concept in a fundamental redesign of the bow that he returned to repeatedly: the pulley bow.



THEORETICAL APPROACH to the design of a crossbow appeared in this late-medieval sketch by an anonymous artist in what is usually called the Hussite Wars manuscript. The bow is modeled as an ideal geometric form, the circle. Important parts of the mechanism are at key geometric positions. An example here is the string-release mechanism, which is at the center of the circle. Simple ratios also are evident, one of them being the fact that the bowstring bisects a radius. Such idealistic approaches to engineering design represent one of the starting points for Leonardo's more sophisticated investigations of the subject.

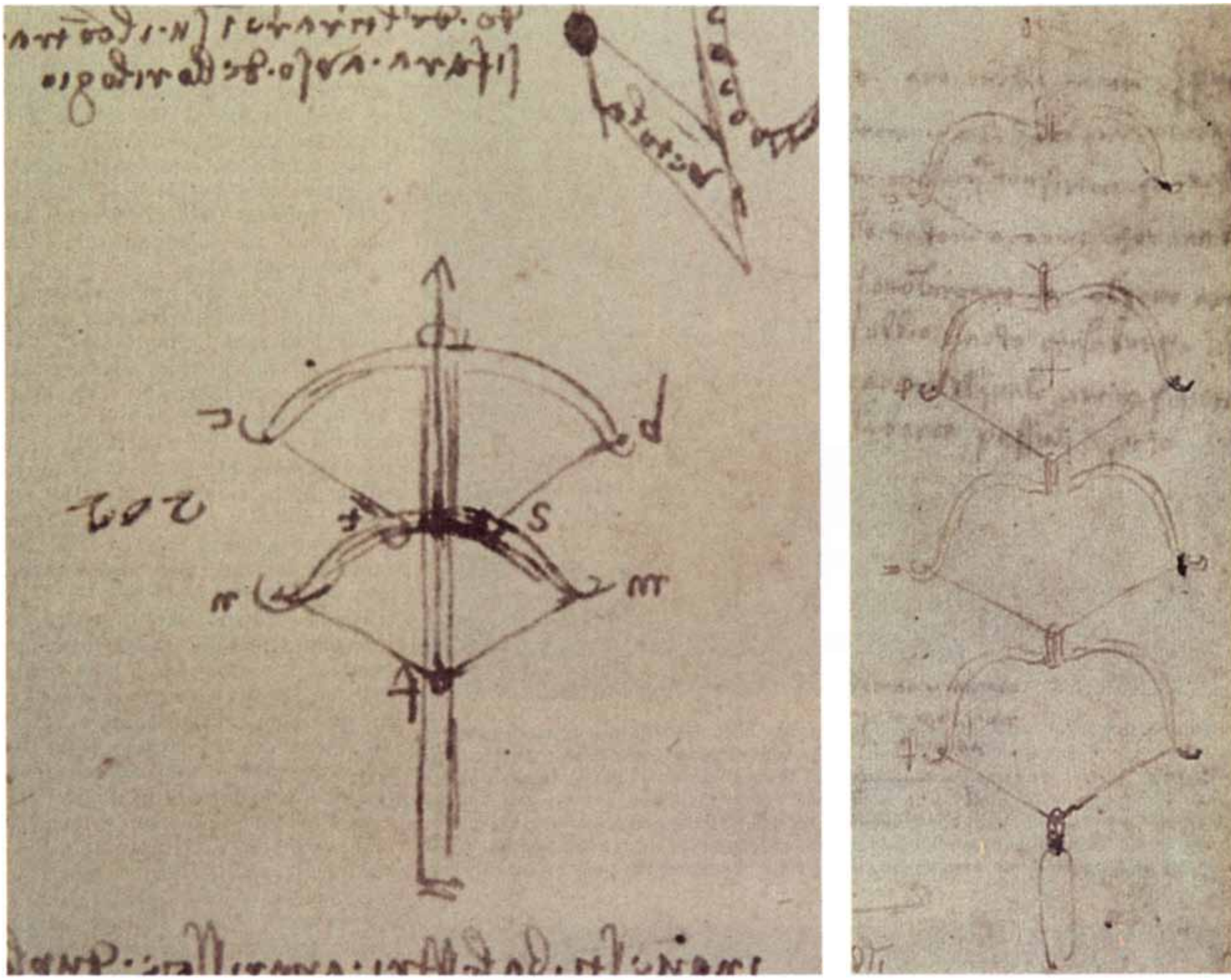
Pulley (or compound) bows, in which the string passes around pulleys, are familiar to modern archers. They are characterized by a high launch velocity and a substantial falloff in draw weight as the bow approaches full draw, and both characteristics are well understood. Leonardo did not reach this degree of understanding, but he did design crossbows whose strings are led around pulleys. His pulleys seem to have always been pivoted solidly on the tiller, rather than moving with the limb tips as the crucial pulleys do in the modern design. Hence his pulley bow would not display the multiplier effect of modern compound bows. In any case, he apparently intended to make a bow that would solve the string-angle problem, increasing the force of launch by decreasing the angle at the center of the string. He also sought to make the bow avoid the waste of energy entailed in a premature launching of the bolt.

In his basic design a highly flexible bow is mounted on a tiller. Some versions show the bow strained almost into a circle. The string runs straight inward from the bow tips, encountering a pair of pulleys mounted on top of the tiller next to the arrow groove. Passing around both pulleys, it loops back to the nut, where it meets the rear of the bolt.

Leonardo seems never to have explained the reason for this design. It appears repeatedly in his drawings, however, together with another odd form of crossbow, again extremely curved, in which the string reaches straight back from the tips to the nut, forming a deep V shape.

The most plausible explanation for both designs is that he wanted to minimize the central angle of the string in order to get better acceleration on launch. He probably also led the string around pulleys to keep the angle between the string and the limbs as close to 90 degrees for as long a time as possible. His intuitive appreciation of the law of force composition allowed him to propose such a radical redesign of a time-tested weapon on the basis of a quantified relation between the energy stored in a bow and the motion of the string. Clearly he had some idea of the mechanical efficiency of this arrangement and sought to improve it.

Leonardo's pulley bow may have been impractical, because the sudden



MULTIPLE BOWS represent Leonardo's interest in what would now be called velocity additions. The idea was to join as many as four bows, each bow launching the next until the last one launched the bolt or arrow. Leonardo's intuition told him correctly that an increase in the energy stored in the bow system would increase the energy and range of the projectile. He did not know the law of the conservation of energy, however, and so he was not fully aware of the losses of energy that would have been entailed in moving several bows and the projectile. It is doubtful that Leonardo's scheme was ever tried, or that it would have yielded much gain in performance.

straightening of the string would have loaded it heavily. Only bows of sophisticated composite construction could have withstood the extreme flexing he envisioned.

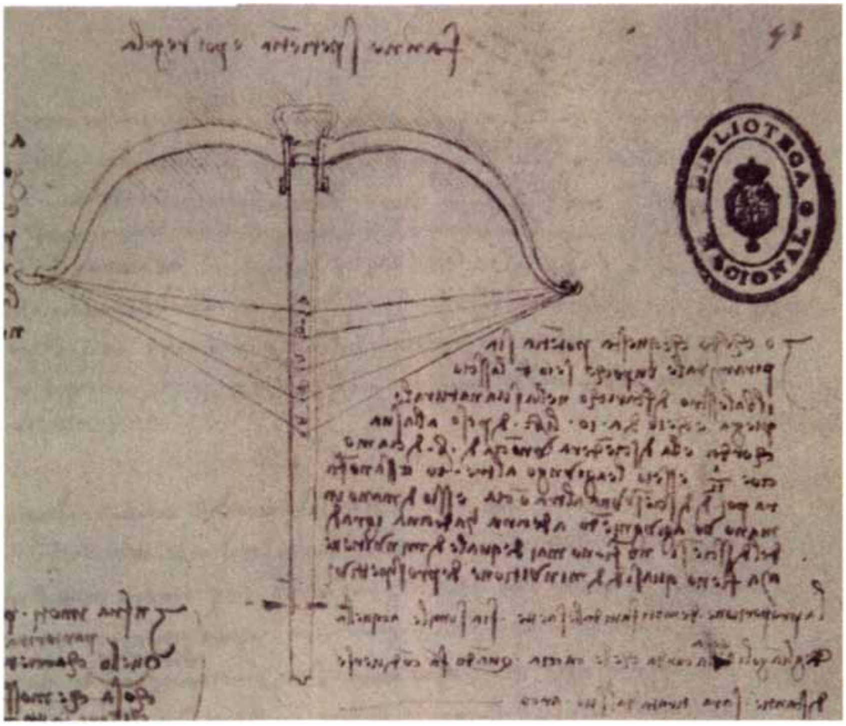
Composite bows of a kind were in existence in his day, however, and they may have led him to his work on what is now known as the neutral plane. The work entailed an even deeper insight into the behavior of matter under stress.

In a typical composite bow of Leonardo's time the back and belly (outer and inner) sides of the limbs were made of different materials, often sinew on the side under tension and horn on the side under compression. Each material is stronger than wood for the particular force to which it is exposed. Between them was a thin layer of wood, by itself hardly enough to give the limb any stiffness. The limbs of such a bow can be bent through more than 180 degrees. Leonardo understood something of how such a bow was made, and the design's use of different materials to absorb tension and compression may have brought him to a fundamental realization of how stresses build up in a structure.

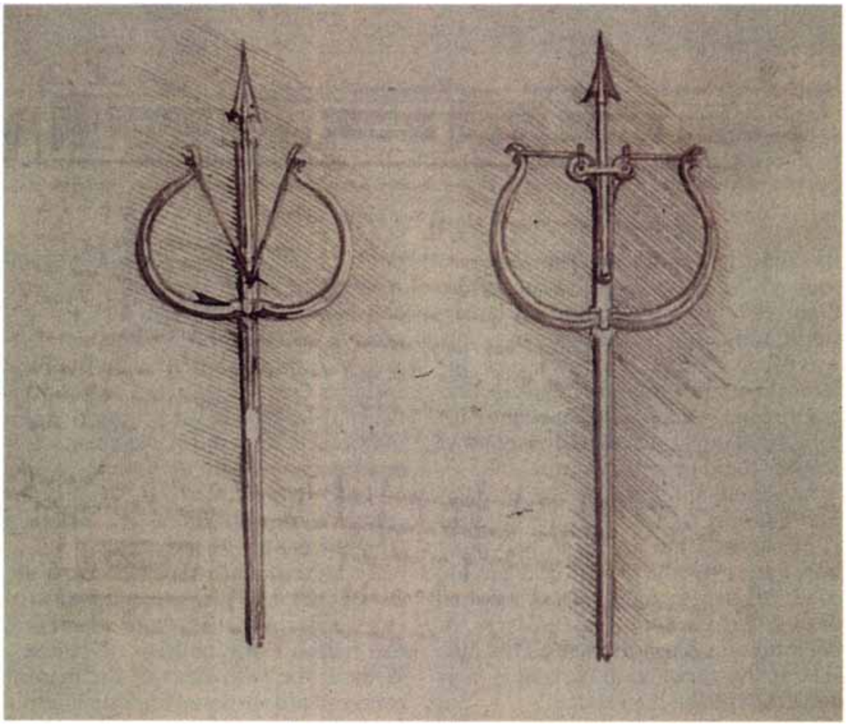
In a small and not very elegant two-part drawing (found in Codex Madrid I) he shows a flat spring in two positions, straight and flexed. Across the edge of the straight spring he has drawn two parallel lines adjacent to the midpoint. They diverge on the outside of the curve as the spring is bent and converge on the inside.

Leonardo's text states that when a spring is bent, the convex part becomes thinner and the concave part thicker. "This modification is pyramidal, and consequently there will never be a change in the middle of the spring." In other words, the distance between the initially parallel lines will grow at the top as much as it diminishes at the bottom. The center part of the spring is thus like a kind of balance between the two sides. It is a zone where the stress is zero: the neutral plane. Leonardo also recognized that the strains of both tension and compression increase in proportion to the distance from the neutral zone.

By identifying the point where the stresses are zero the neutral plane provides a starting place from which one can measure the buildup of stresses. Without the twin ideas of the neutral zone and proportional buildup it is difficult if not impossible (except with a computer) to predict the stress buildup correctly and hence to design, for example, a beam that will be safe from failure. A computer can divide ma-



FORCE RELATIONS caused Leonardo some puzzlement. Here he was examining the relation between the increments of movement of the bowstring and the weight or force applied in cocking the bow. At times he recognized that the relation would not be simply proportional. Because the string at full draw stored more energy than it did at other positions, he thought the arrow might leave the string before receiving all the energy stored in the string. He apparently realized that the central angle formed by the string was a key measure of the weapon's ability to transmit energy, and so he tried to maximize the angle by new designs.



REDESIGN OF CROSSBOW by Leonardo is represented in two of his sketches. He apparently never wrote down his reasons for inverting the bow and making the string angle so acute, in one case (*right*) with the aid of pulleys. Related texts suggest his objective was to develop a weapon that would transmit the bow's energy to the bowstring more efficiently.

chine or building parts into a number of small components and estimate the stresses in each one, but even so the neutral-zone idea is usually employed to keep the programs manageable. Here again, then, Leonardo was dealing with one of the most basic concepts of engineering mechanics. (The Leonardo scholar Carlo Zammatio has also noted the significance of these two small drawings.)

On this topic too a look at Leonardo's sketches helps to trace the genesis of the idea back to the crossbow. An example is his sketch of a giant stonebow, a weapon designed to hurl stones. The bow is drawn by a screw windlass and the stones are launched from a pouch at the center of the double bowstring. Both features were scaled up from manually operated crossbows. Leonardo appears to have realized, however, that enlarging the bow in this way would create problems.

His neutral-zone sketches show his recognition that, for a given degree of bending, the stresses in a bow would increase directly with the thickness. To prevent them from reaching a critical level he redesigned the giant bow. The forepart, which is the tension side, is made of a single piece of timber. The belly (under compression) is made of individual blocks pinned or lashed behind the forepart. They are shaped so that they will touch only when the bow is at or near full draw. This design and others demonstrate Leonardo's belief that tension and compression forces should be handled separately.

Our final example, dealing with the center of gravity, serves like the others to suggest how a serious study of even simple machines can indicate the outlines of deeper fundamental principles. From another small, rough but rather graceful sketch (in Codex Madrid I) one can trace connections between Leonardo's musings on the crossbow and his later theorizing on flight in general.

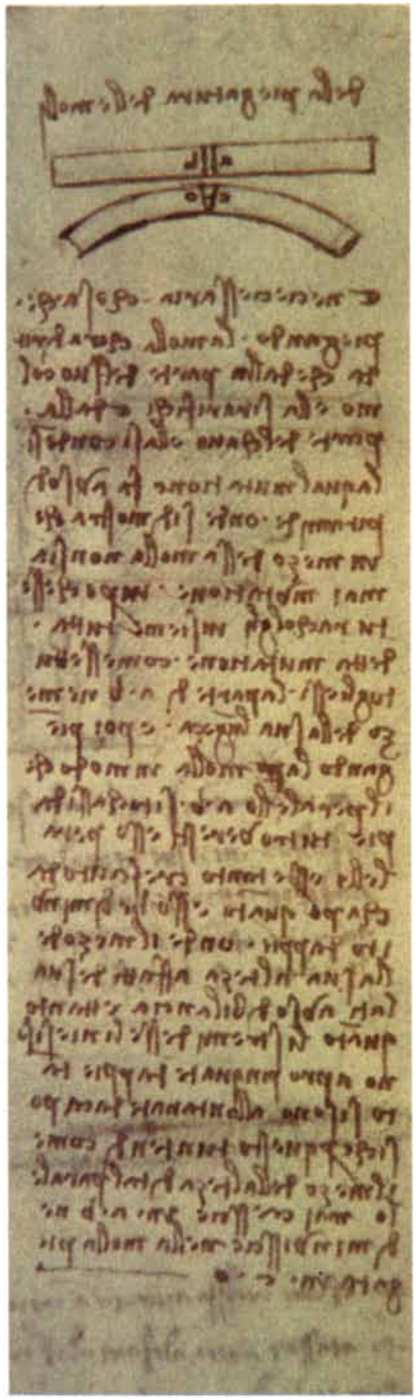
The sketch shows a crossbow being shot vertically. At the outset the bolt is reversed because it had been loaded backward in the bow—something Leonardo may have seen crossbowmen do as a lark in practice. As it flies, it must tumble until the point is forward. On the way down it will reverse this sequence. Leonardo states that the same path will describe the bolt's motion in each case. "The center of natural motion [caused by gravity] and the center of the accidental motion [caused by other forces] are one and the same.... The center of gravity of the arrow will be found at the same center throughout its motion, be it accidental or natural." Comparing these words with the drawing, one sees that the reversing bolt rotates about its center of gravity as it regains its proper flight attitude. All the while the center of gravity is moving along the usual flight path.

In other words, Leonardo has here taken the first steps toward reducing the motion of a bolt or an arrow to that of a mass point moving at a certain velocity along a preordained line. His sketch is therefore important in the history of rotational dynamics, because it simplifies the situation enough to make it treatable mathematically. Even more important, the sketch represents what is to our knowledge the first attempt to apply the idea of the center of gravity in dynamics. (Archimedes had invoked it for static situations.) Marshall Clagett of the Institute for Advanced Study in Princeton, a prominent student of medieval mechanics, has corroborated our opinion on this point.

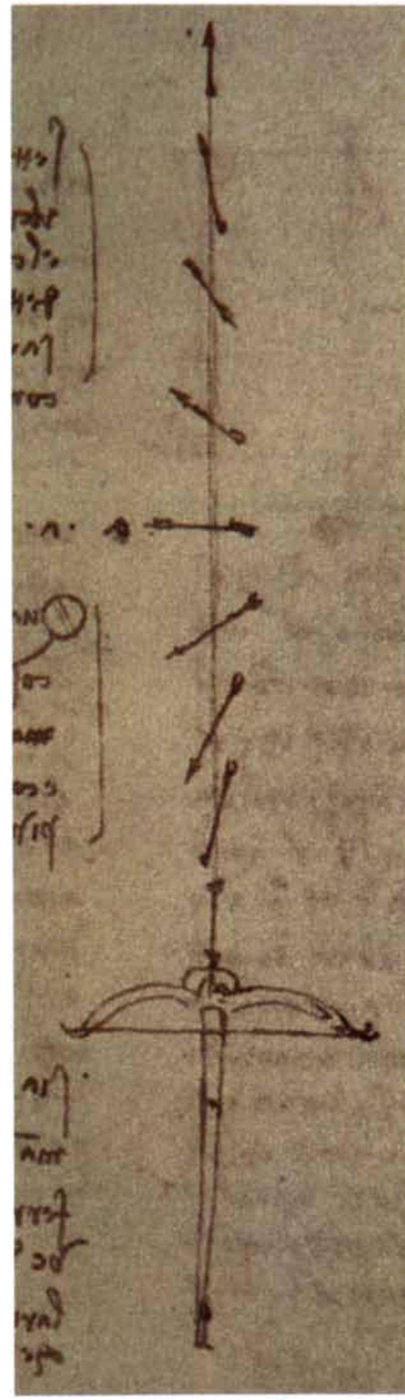
Before the discovery of the Madrid Codices one other sketch by Leonardo illustrating the center of gravity in a moving body was known. Found in a notebook called Codex Forster II, it shows a cube tumbling as it falls. The text states that the center of gravity will remain on the line of descent. The Madrid crossbow-bolt sketch appears to be the earlier of the two. It may well mark the starting point for Leonardo's theoretical investigations of the conditions of flight.

In the manuscript *Sul Volo* and elsewhere Leonardo points out that stable bird flight occurs only when the center of gravity of a bird lies forward of its center of resistance (the point where there is equal pressure front and back). This principle remains as fundamental to aircraft or to rockets as it was to birds when Leonardo applied it.

We found it was not difficult to verify Leonardo's finding experimentally. We selected a weak bow and a heavy arrow (to keep the velocity of the arrow relatively low) and put a band of fluorescent orange tape around the balance point of the arrow for visibility. A special heavy head, incorporating a nock so that the arrow can be shot with the feathers forward, will facilitate shooting and observation by shifting the center of gravity toward the head. It is helpful to shoot against a dark background such as a grove of trees and to face away from the sun. As unskilled archers we found that not every shot showed the mass-point effect cleanly, but in a fair number of cases the arrow dramatically swapped ends around the balance point while the orange marker continued downrange in a stately curve.



NEUTRAL ZONE in a stressed structure is apparently a discovery of Leonardo's. His sketches show a spring or a beam in both straight and flexed positions. The parallel vertical lines he drew at the center of the flat spring diverge on the outside and converge on the inside as the spring is bent. Leonardo's text sets forth his recognition that the beam center is unstressed (the neutral zone) and that the stresses build up from there in proportion to the distance. With this realization of what is now called the concept of the neutral plane it became possible to find the theoretical level of failure in all kinds of structures and machines.



CENTER OF GRAVITY in a body in motion engaged Leonardo's attention in these sketches of an arrow shot vertically. At first it is backward because it was loaded backward in the bow (something Leonardo may have seen done playfully by crossbowmen in practice). The text indicates his recognition that the arrow would rotate about its center of gravity and that the center would follow the same path as if the arrow had been launched normally. The sketch appears to be the first effort to apply the concept of the center of gravity in dynamics. It is a step toward Newton's attempts to consider the planets as moving points of mass.

THE AMATEUR SCIENTIST

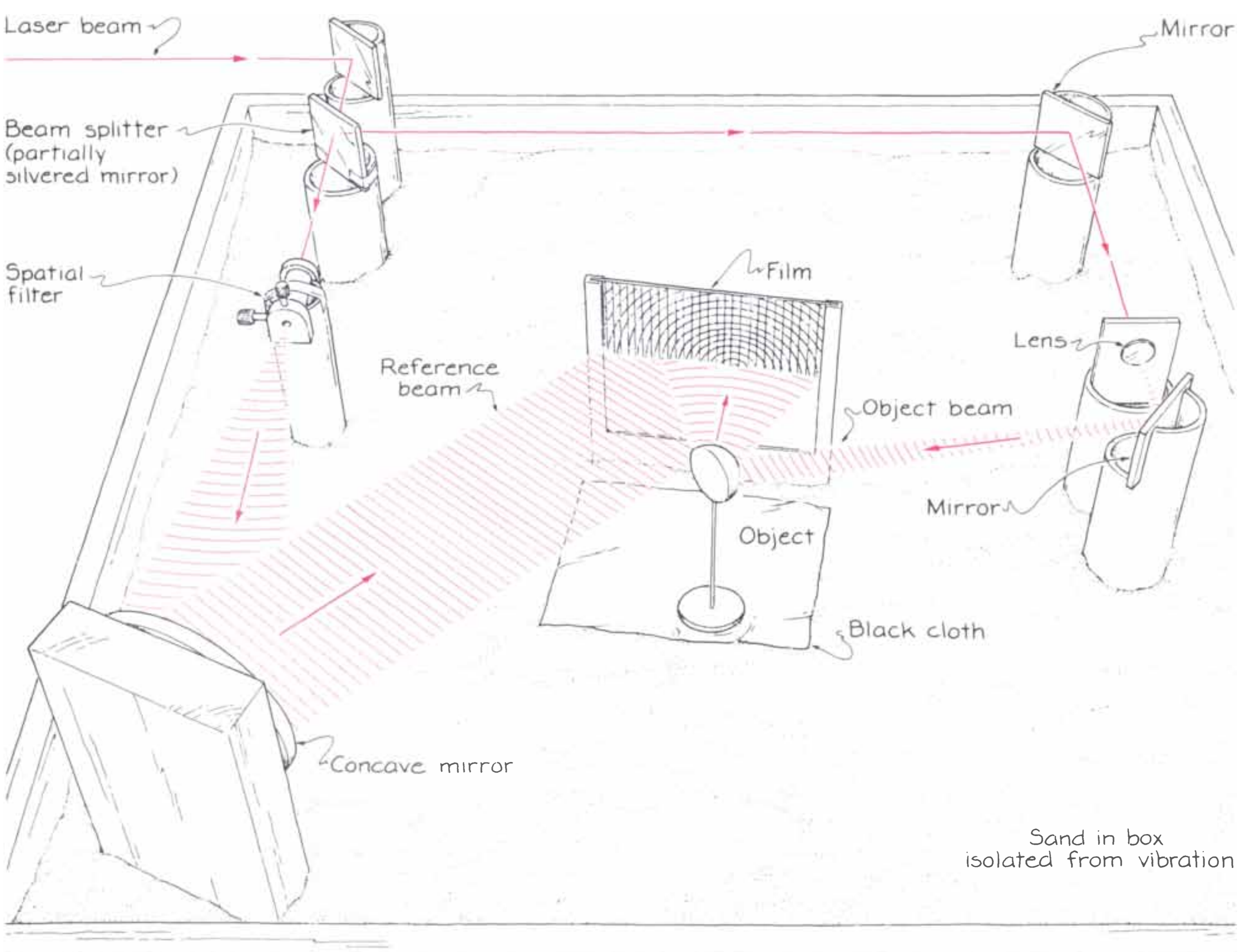
Rainbow holograms, unlike conventional ones, can be observed in ordinary light

by Jearl Walker

When holograms first became available, they presented the user with a difficulty: he needed monochromatic light, such as the light from a laser, to illuminate a hologram so that he could see it. In white light the images were blurry because each wavelength of the visible spectrum formed its own image, slightly displaced from the others. Rainbow holograms, developed in 1969 by Stephen A. Benton of the Polaroid Corporation, are designed to minimize blurring while still producing colorful images in white light. They are employed in works of art and novelty items and as a security measure on certain credit cards.

I have looked into this phenomenon with the help of Douglas S. Hobbs, who studied rainbow holography under the direction of Richard W. Henry of Bucknell University. I have also relied on two excellent books. *Seeing the Light: Optics in Nature, Photography, Color, Vision, and Holography,* by David Falk, Dieter Brill and David Stork, examines the optics of holograms and related devices. *Holography Handbook: Making Holograms the Easy Way,* by Fred Unterseher, Jeannene Hansen and Bob Schlesinger, is rich in detail about how to produce all kinds of hologram on a limited budget.

Because a rainbow hologram is usually made from a standard hologram, one needs to know the essential features of that remarkable artifact. It is not a conventional photograph made when a lens focuses an image of an ob-



An arrangement for making a hologram

ject onto photographic film. Instead it is a photographic record of the interference pattern formed by two beams of light, one of which has been reflected from the object.

In one arrangement the beam from a laser is split with a partially silvered mirror. The part that passes through the mirror goes next through a spatial filter, which consists of a lens focused on a pinhole. The pinhole is at the focal point of a concave mirror. When the light is reflected from the mirror, the resulting beam (the reference beam) consists of plane waves: the wavefronts are straight and the rays are perpendicular to the wavefronts and parallel to one another.

The part of the light that is reflected from the partially silvered mirror is directed by other mirrors and spread by a lens until it is reflected from the object. This beam (the object beam) consists of waves that have curved wavefronts and diverging rays. To make matters simpler I assume the object is a point; the rays diverge from the point and the wavefronts are partial circles centered on it.

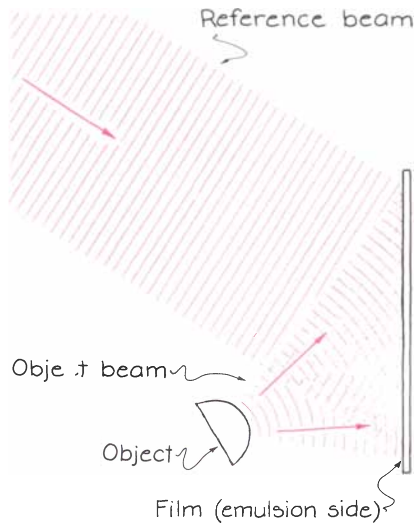
The reference beam and the object beam pass through a photographic film. There they interfere with each other. As a result the film records thin lines where the beams interfere constructively (wave crests coincide with crests and valleys coincide with valleys) to yield bright light that exposes the silver grains. Between the lines are other thin lines arising where the beams interfere destructively (crests coincide with valleys), yielding darkness and leaving the grains unexposed. When the film is developed, the unexposed regions are transparent and the exposed regions are opaque. The film is then a hologram: a permanent record of the interference pattern of the two beams.

The possibility of recording an interference pattern depends on the fact that at each point in the film the waves in the object beam always have a fixed relation of phase with the waves in the reference beam. If they did not, each point in the film would receive destructive interference at one instant and constructive interference at another instant. The entire film would be exposed and no single pattern would be recorded.

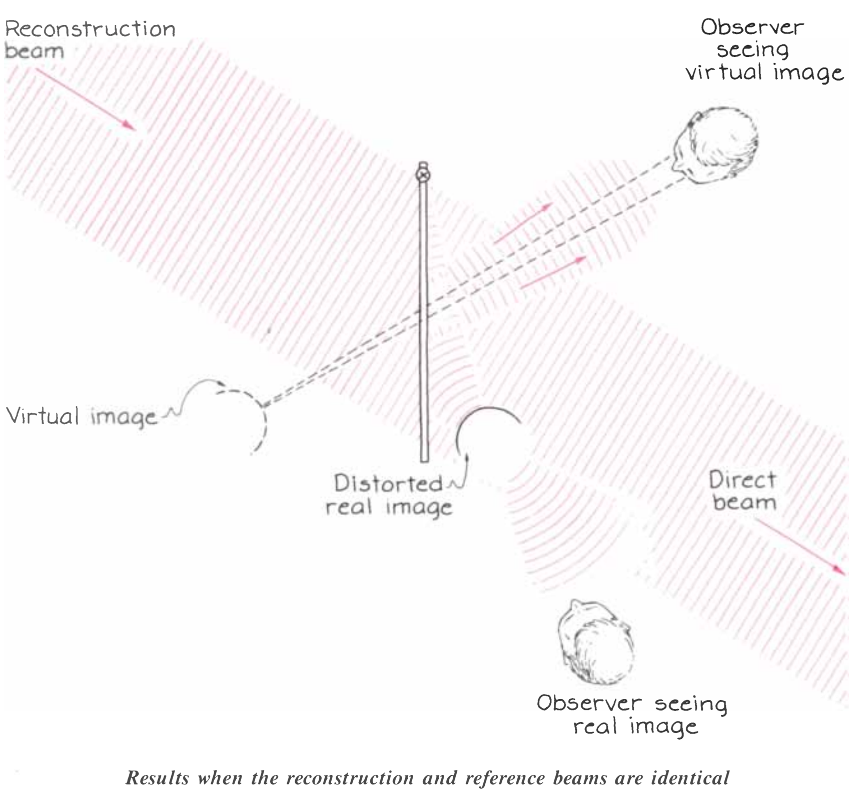
Creating a fixed phase relation between the two beams requires coherent light, such as the light from a laser. One can think of the output from a laser as a succession of long trains of waves. Each train is made up of a continuous wave with a length (the coherence length) roughly the length of the laser tube. When the object beam and the reference beam arrive at the film, they have a fixed phase relation if they are parts of the same train. One achieves this condition by making the distances the two beams travel to reach the film approximately equal. Then each point in the film receives a single type of interference; as a result a single interference pattern is recorded by the film.

The pattern is too fine to be seen. If you examine the hologram in room light, you do not see an image of the object. Nevertheless, the details of the object are recorded there. To display the object the hologram is positioned in a beam of laser light (called the reconstruction beam) identical with the original reference beam. This beam is diffracted as it passes through the pattern of lines within the hologram. Much of the light passes directly through with no change in direction, but some of the light is diffracted to form two new beams, one on each side of the direct beam.

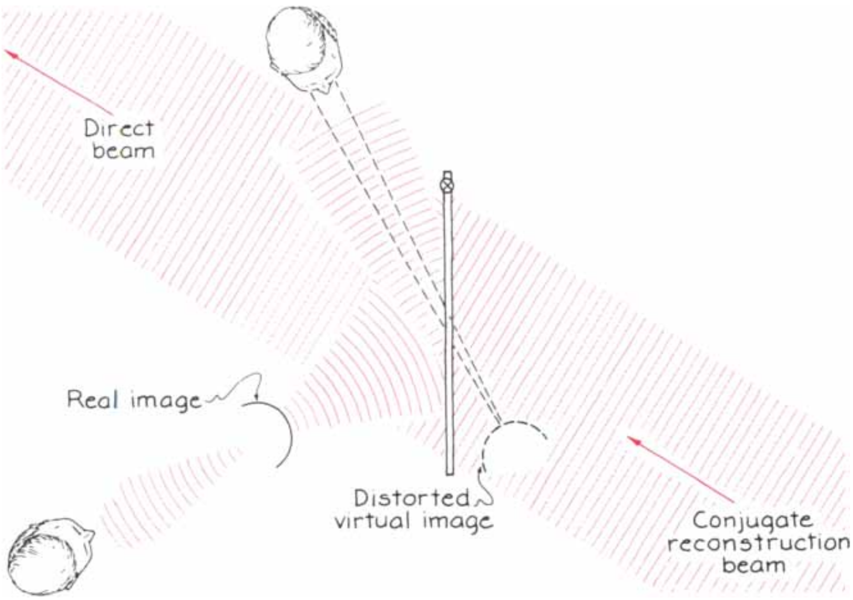
If you look at the hologram when your eyes are intercepting one of these side beams, you see an image of the object. In one of the beams the waves emerge from the hologram traveling in the same directions as the waves in the original object beam did. Intercepted waves are focused onto the retina of each eye. Since your eyes are separated and intercept different parts of the diverging waves, the locations of the images on the retina differ, an effect called parallax. Your visual system interprets the difference in location as evidence that the source of the light lies on the other side of the hologram at the spot originally occupied by the object. In effect your visual system extrapolates the rays of light backward until they cross at the location of the perceived image. For this reason you



How light waves interfere



Results when the reconstruction and reference beams are identical



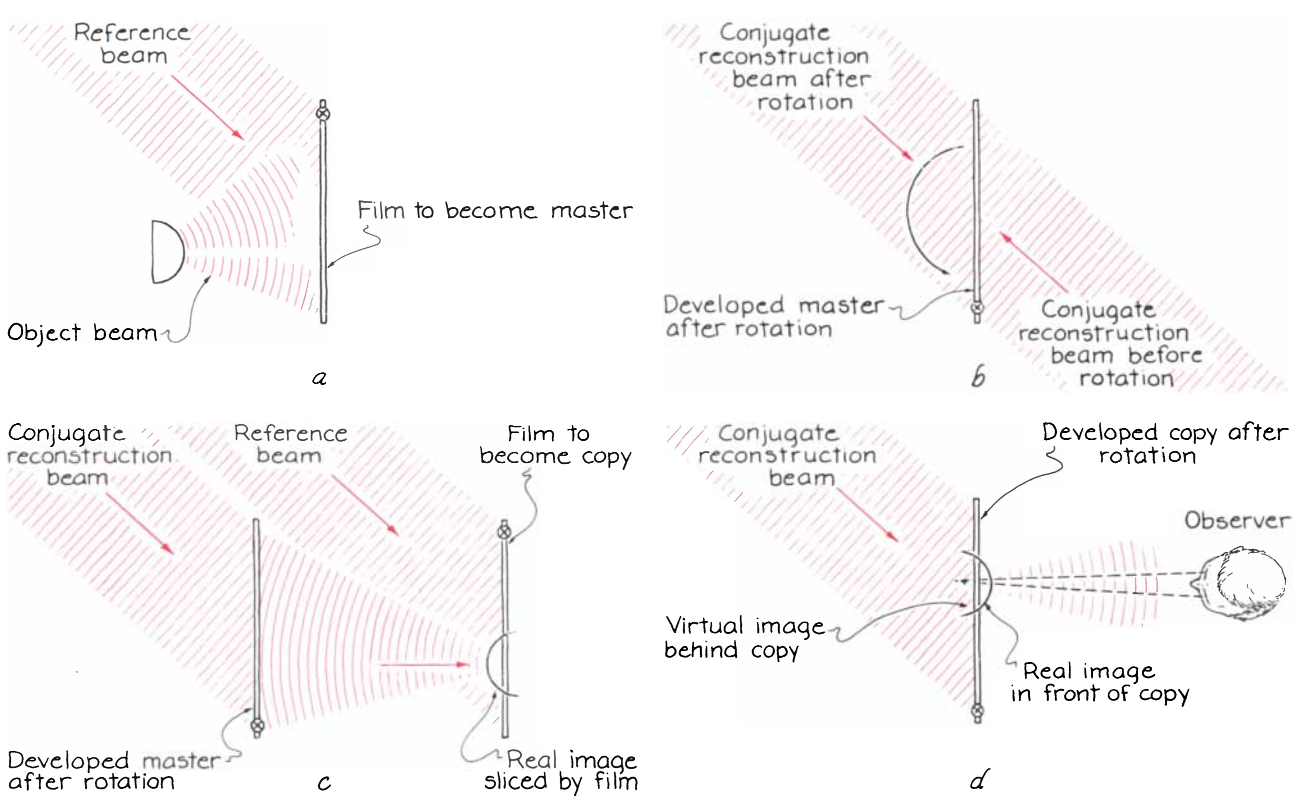
Images made with a conjugate beam

perceive depth in the scene. The image is said to be a virtual one because if you place a blank card at its apparent location, no image forms on the card. The image is only a product of your visual system.

The other beam of diffracted light consists of waves that first converge and then diverge. If you place a blank card at the point of convergence, a real image of the object appears on it. You can also see the image by intercepting with your eyes the light diverging from the image. Again you perceive the image to be at a certain location because of parallax. In this case you see the image in front of the plane of the hologram. The image, however, is usually too distorted to be recognizable; with some arrangements of equipment it may not appear at all.

An undistorted real image can be produced by sending a beam of light through the hologram in a direction opposite to the direction of the original reference beam. This reconstruction beam is said to be conjugate to the reference beam. As before, part of the light diffracts into two side beams, one beam capable of producing a real image and the other a virtual image. This time it is the virtual image that is distorted or missing.

A hologram is actually a diffraction grating. A standard diffraction grating consists of fine, equally spaced parallel lines. The lines in a hologram are not equally spaced because the object beam diverges. At any location in the film the spacing of the lines depends on the angle between the rays in the object and reference beams reaching that point. If the angle is small, the lines are widely spaced. If the angle is large, the lines are narrowly spaced. The variation in the spacing of the lines is the factor that gives rise to the convergence or divergence of the



How to make a copy from a master hologram beams diffracted to the sides when the hologram is illuminated by a reconstruction beam.

Suppose the object is extended instead of being a point. Each point on its surface reflects diverging light to the film, interfering there with the reference beam. The interference pattern recorded on the film is quite complex. Still, when the reconstruction beam is sent through the developed film along the reference-beam path or the conjugate to it, side beams form. Two of them provide undistorted images. If your eyes intercept either beam, the image is three-dimensional because of parallax. If you change your viewpoint within a beam to intercept different parts of the waves, you see a different perspective of the image. This feature enables you to see the front of the object or, by changing your viewpoint, to see along the sides, just as you would if you looked at the original object. (You cannot see an image of the back of the object with this arrangement of equipment because rays from the back never reach the film.)

The undistorted real image that is produced by a conjugate reconstruction beam can be displayed on a card. It is blurry, however, because only parts of it are in focus at any given location of the card. When you see it by intercepting the rays, it has inverted depth. For example, the front of the object seems to be farther away than the sides, giving the impression that the image is inside out. The image is said to be pseudoscopic.

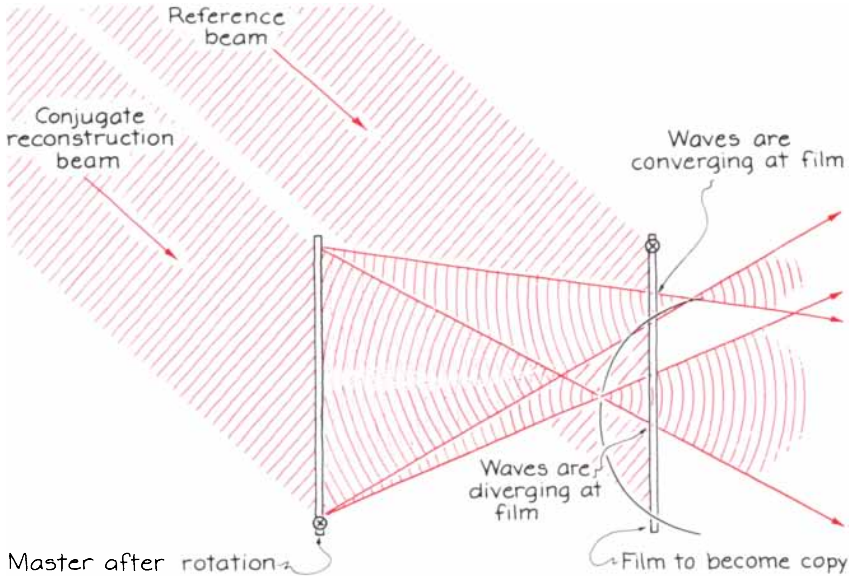
This image from a hologram can be valuable in a technique in which a second hologram is made from the first. Such an arrangement is called focused-image or image-plane holography since the copy is positioned within the real image cast by the master. The real image is produced by sending through the master hologram a reconstruction beam conjugate to the master's reference beam.

In order to avoid rearranging the positions of the laser and various optical devices, one would like to employ the same beam that previously served as the reference beam. This can be accomplished by rotating the master 180 degrees about a vertical axis so that what was the reference beam now becomes the conjugate reconstruction beam. Because of the rotation, the light passes through the master in the direction opposite to the path of the original reference beam.

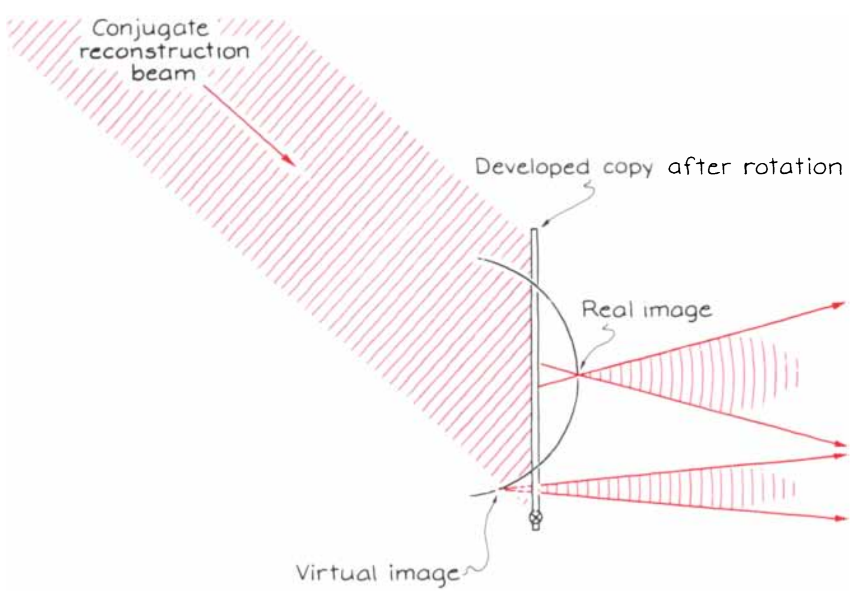
The beam that forms the real image then functions as the object beam for the copy. A reference beam also illuminates the copy. The film records the resulting interference pattern of the two beams. After the copy is developed it is rotated about a vertical axis so that what was a reference beam is now the conjugate reconstruction beam. When your eyes intercept the light diffracted from the beam by the copy, you perceive a three-dimensional image that appears to straddle the plane of the hologram. The part in front of the plane is a real image. The part that seems to be behind the plane is a virtual image.

The image forms as follows. The light traveling from the master to the copy converges to form a real three-dimensional image through which the copy film makes a two-dimensional slice. The rays that form an image of the front of the object converge in front of the film. When they pass through the film, they are diverging. The rays that form an image of the other sides of the object converge behind the film. As those rays pass through the film they are still converging. The interference between these two sets of rays and the reference beam records the convergence and divergence of the rays.

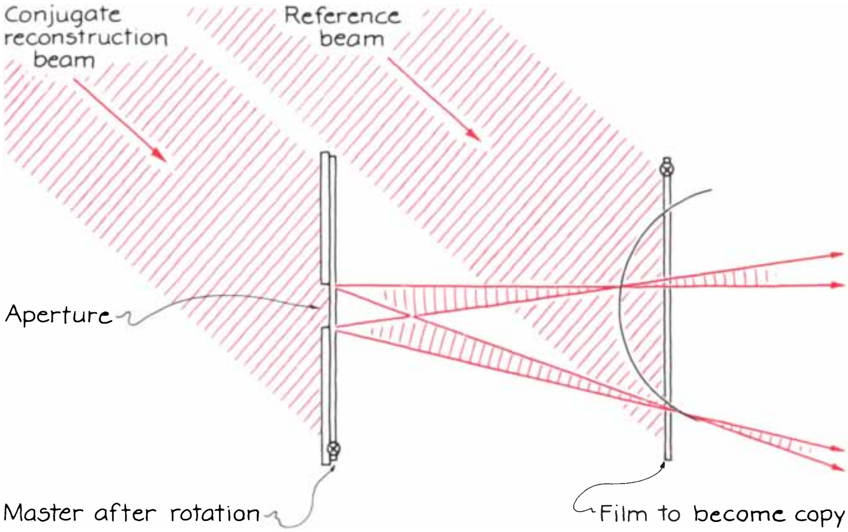
When the copy is developed and a conjugate beam is sent through it, the



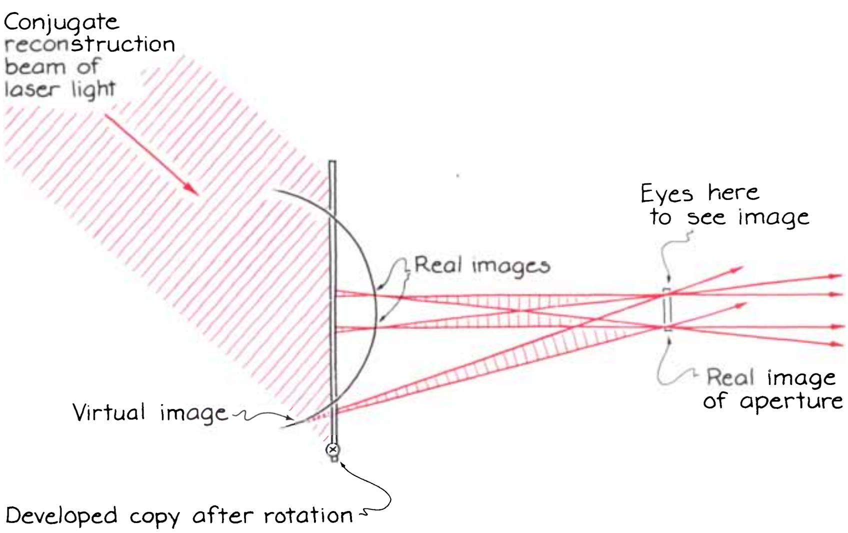
Wave action when a film intercepts a real image



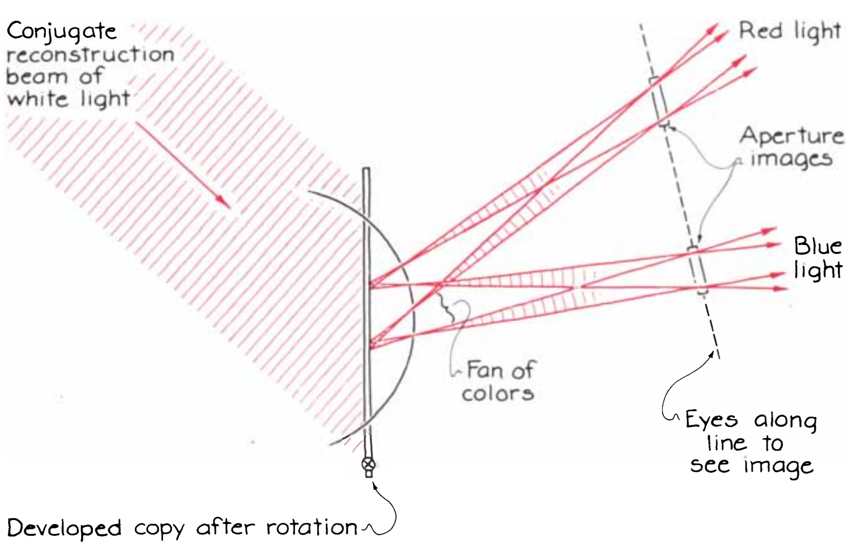
The image seen in a copy hologram



Using an aperture in a master hologram



The aperture image made by a copy hologram



Colored images from a rainbow hologram

diffraction of the light reverses what the rays did during the exposure of the copy. Where rays had been converging they now diverge and where they diverged they now converge. When your eyes intercept the rays, you see a real image of the front of the object lying in front of the copy and a virtual image of the other sides of the object lying behind the copy. Although the master's real image is pseudoscopic, the final image shows normal depth.

If white light is sent through the copy instead of monochromatic light from a laser, each wavelength in the visible spectrum is diffracted by a slightly different amount, creating its own image. The resulting overlap of many colored images yields a blurry composite image. Benton discovered a way for an observer to be able to see clear images when a copy is illuminated by white light. The result is called a rainbow hologram because of the vivid colors of its images.

In making a rainbow hologram only a narrow strip of the master is illuminated by its reconstruction beam. Although information about depth perpendicular to the strip is lost, information about depth parallel to the strip is preserved. Hence a rainbow hologram is a compromise: the possibility of parallax is eliminated in one direction so that the image can be seen in white light without blurring.

In one method the master is masked except for an aperture running across its width. In the top illustration at the left, which shows the setup from overhead, the aperture is vertical. Parallax information about the object is kept in the vertical direction but is lost in the horizontal direction. The aperture's width is also recorded in the copy.

The developed copy is viewed in a conjugate laser beam. (Again, in order for the laser beam that served as a reference beam for the copy to now serve as the conjugate reconstruction beam, the copy is rotated about the vertical axis.) The rays of light emerging from the copy pass through a real image of the aperture that lies in front of it. In the middle illustration at the left the aperture image extends into the page. You can see the image of the object if you position your eyes anywhere in the aperture image. By moving your eyes to a new location within the aperture image you see a different perspective of the object. Information about depth is retained along the direction of the aperture image (vertically in this case). If you move your eyes out of that image, however, the image of the object disappears. Information about depth is lost along the direction at right angles to the aperture image.

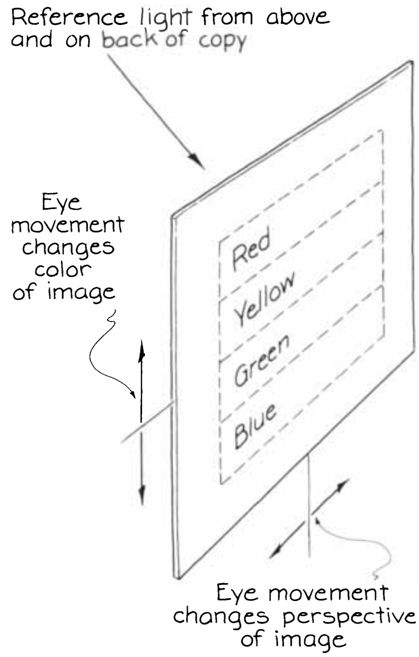
When white light is sent through the developed copy, each wavelength is diffracted by different amounts and creates an aperture image at a different location in front of the hologram. If you position your eyes in the aperture image created by red light, you see a red image of the object. As you move your eyes through the other aperture images, the image smoothly changes color. No blurring results from this procedure if your eyes are at an aperture image. If you move too close to or too far from the hologram, the rays passing through the various aperture images overlap and a blur results.

Hobbs experimented with ways of making and conveniently viewing a rainbow hologram. Assuming it would be held vertically in descending light that strikes the far side, he decided to turn the object onto its side when he made the master. After the copy was developed he rotated it so that the descending light was conjugate to the reference beam the copy had when it was exposed. By this procedure he made the top of the object appear at the top of the hologram.

Hobbs also arranged for viewing along a line approximately perpendicular to the plane of the copy. Since green light is in the middle of the visible spectrum, the rays forming the green aperture image should represent that perpendicular line. By moving your eyes upward and closer to the hologram you bring up the red aperture image; moving downward and farther from the hologram makes the image blue.

Hobbs calculated that if the rays of green light are to leave the copy approximately perpendicular to its plane after reaching the copy at about 45 degrees, the spacing between diffraction lines within the hologram should be about .78 micrometer. (This number and the other numbers that follow are only representative. The complex pattern of lines actually has a range of spacings.) At that spacing light at the red end of the spectrum (wavelength .7 micrometer) is diffracted about 11 degrees to one side of the green rays and light at the blue end (wavelength .4 micrometer) is diffracted about 11 degrees to the other side. Intermediate colors are diffracted at intermediate angles. The full spectrum of aperture images lies in a fan covering an angle of about 22 degrees.

Hobbs also gave thought to the distance between the copy and the observer's eyes. If the green aperture image is 35 centimeters from the copy, the extreme red one is about 28 centimeters away and the extreme blue one is about 48 centimeters away. These distances make viewing quite convenient.



Where the colored images appear

The line separation within the copy depends on the wavelength of the light that made it and on the angle between the reference beam and the object beam (from the aperture on the master) that illuminate it. If the light from a helium-neon laser (wavelength .63 micrometer) is to produce a line separation of about .78 micrometer, the angle between the two beams should be about 54 degrees.

Finally Hobbs considered the light source that would serve as the reconstruction beam. If the source is the sun, the beam consists of approximately plane waves. Since the beam is supposed to be conjugate to the reference beam employed in making the copy, the reference beam must also consist of plane waves. If the light source is a light bulb, the reconstruction beam consists of diverging rays. The conjugate of such a beam is a beam of converging rays, and so the reference beam would have to be made up of converging rays too.

Many suggestions about the making of rainbow holograms can be found in *Holography Handbook*. They include ways to concentrate the light passing through the aperture on the master by means of a cylindrical lens made from a test tube filled with glycerin. With such a lens you may even be able to do without making the master. The lens can focus a narrow, real image of the object directly onto the copy.

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

COMPUTER RECREATIONS

MATHEMATICAL GAMES. Martin Gardner in *Scientific American,* Vol. 224, No. 2, pages 112–117; February, 1971.

COMPUTER RECREATIONS. Brian Hayes in *Scientific American,* Vol. 249, No. 4, pages 22–36; October, 1983.

THE MICROWAVE PROBLEM

BIOLOGICAL EFFECTS AND MEDICAL APPLICATIONS OF ELECTROMAGNETIC ENERGY. *Proceedings of the IEEE,* Vol. 68, No. 1; January, 1980.

THE ORIGINS OF U.S. SAFETY STANDARDS FOR MICROWAVE RADIATION. Nicholas H. Steneck, Harold J. Cook, Arthur J. Vander and Gordon L. Kane in *Science,* Vol. 208, No. 4449, pages 1230–1237; June 13, 1980.

EFFECTS OF LONG-TERM LOW-LEVEL RADIOFREQUENCY RADIATION EXPOSURE ON RATS. A. W. Guy et al. U.S.A.F. School of Aerospace Medicine, 1983–85.

RADIOFREQUENCY RADIATION EXPOSURE LIMITS IN EASTERN EUROPE. Peter Czerski in *Journal of Microwave Power,* Vol. 20, No. 4, pages 233–239; 1985.

BIOLOGICAL EFFECTS AND EXPOSURE CRITERIA FOR RADIOFREQUENCY ELECTROMAGNETIC FIELDS. National Council on Radiation Protection and Measurements, Report No. 86, Bethesda, Md., 1986.

PREDICTING CHEMISTRY FROM TOPOLOGY

MOLECULAR CONNECTIVITY IN CHEMISTRY AND DRUG RESEARCH. Lamont B. Kier and Lowell H. Hall. Academic Press, 1976.

CHEMICAL GRAPH THEORY. Nenad Trinajstić. CRC Press, 1983.

INFORMATION-THEORETIC INDICES FOR CHARACTERIZATION OF CHEMICAL STRUCTURES. Danail Bonchev. Research Studies Press, distributed by John Wiley & Sons, Inc., 1983.

SUPERSTRINGS

AN INTRODUCTION TO THE THEORY OF DUAL MODELS AND STRINGS. J. Scherk in *Reviews of Modern Physics,* Vol. 47, No. 1, pages 123–164; January, 1975.

UNIFICATION OF FORCES AND PARTICLES IN SUPERSTRING THEORIES. Michael B. Green in *Nature,* Vol. 314, No. 6010, pages 409–414; April 4, 1985.

SUPERSTRINGS AND SUPERSYMMETRY. Alvaro de Rújula in *Nature,* Vol. 320, No. 6064, page 678; April 24, 1986.

SUPERSTRING THEORY. M. B. Green, J. H. Schwarz and E. Witten. Cambridge University Press, in press.

THE BLOOD-BRAIN BARRIER

BLOOD-BRAIN BARRIER IN PHYSIOLOGY AND MEDICINE. Stanley I. Rapoport. Raven Press, 1976.

THE CONCEPT OF A BLOOD-BRAIN BARRIER. Michael Bradbury. John Wiley & Sons, Inc., 1979.

POLARITY OF THE BLOOD-BRAIN BARRIER: DISTRIBUTION OF ENZYMES BETWEEN THE LUMINAL AND ANTILUMINAL MEMBRANES OF BRAIN CAPILLARY ENDOTHELIAL CELLS. A. Lorris Betz, J. Anthony Firth and Gary W. Goldstein in *Brain Research,* Vol. 192, No. 1, pages 17–28; June 16, 1980.

BRAIN METABOLISM: A PERSPECTIVE FROM THE BLOOD-BRAIN BARRIER. William M. Pardridge in *Physiological Reviews,* Vol. 63, No. 4, pages 1481–1535; October, 1983.

SPECIALIZED PROPERTIES AND SOLUTE TRANSPORT IN BRAIN CAPILLARIES. A. Lorris Betz and Gary W. Goldstein in *Annual Review of Physiology,* Vol. 48, pages 241–250; 1986.

THE COLORS OF THINGS

JOURNAL DE EUGÈNE DELACROIX. Introduction and notes by André Joubin. Librairie Plon, 1932.

TREATISE ON PHYSIOLOGICAL OPTICS. Herman von Helmholtz. Dover Publications, Inc., 1962.

THE FERRIER LECTURE, 1962: VISUAL ADAPTATION. W. A. H. Rushton in *Proceedings of the Royal Society, Series B,* Vol. 162, No. 986, pages 20–46; March 16, 1965.

DAYLIGHT AND ITS SPECTRUM. S. T. Henderson. John Wiley & Sons, Inc., 1977.

THE RETINEX THEORY OF COLOR VISION. Edwin H. Land in *Scientific American,* Vol. 237, No. 6, pages 108–128; December, 1977.

THE SUN AND THE INTERSTELLAR MEDIUM

THE EXTRATERRESTRIAL UV-BACKGROUND AND THE NEARBY INTERSTELLAR MEDIUM. H. J. Fahr in *Space Sciences Reviews,* Vol. 15, No. 4, pages 483–540; February, 1974.

SYMPOSIUM ON THE STUDY OF THE SUN AND INTERPLANETARY MEDIUM IN THREE DIMENSIONS. Edited by L. A. Fisk and W. I. Axford. Goddard Space Flight Center, 1976.

THE INTERSTELLAR WIND AND ITS INFLUENCE ON THE INTERPLANETARY ENVIRONMENT. Gary E. Thomas in *Annual Review of Earth and Planetary Sciences,* Vol. 6, pages 173–204; 1978.

BRACHIOPODS

LIVING AND FOSSIL BRACHIOPODS. M. J. S. Rudwick. Hutchinson University Library, 1970.

PERMIAN BRACHIOPODS OF WEST TEXAS. G. Arthur Cooper and Richard E. Grant. Smithsonian Institution Press for sale by U.S. Government Printing Office, 1972–77.

BRACHIOPODS AND PEDICLES. Joyce R. Richardson in *Paleobiology,* Vol. 7, No. 1, pages 87–95; 1981.

BRACHIOPODS. Merrill W. Foster in *Synopsis and Classification of Living Organisms,* edited by Sybil P. Parker. McGraw-Hill Book Company, 1982.

END-CRETACEOUS BRACHIOPOD EXTINCTIONS IN THE CHALK OF DENMARK. Finn Surlyk and Marianne Bagge Johansen in *Science,* Vol. 223, No. 4641, pages 1174–1177; March 16, 1984.

LEONARDO'S CONTRIBUTIONS TO THEORETICAL MECHANICS

MEMORIE SULLA GEOMETRIA E LA MECCANICA DI LEONARDO DA VINCI. Roberto Marcolongo. Stabilimento Industrie Editoriali Meridionali, Naples, 1937.

THE MECHANICAL INVESTIGATIONS OF LEONARDO DA VINCI. Ivor B. Hart. University of California Press, 1963.

THE MECHANICS OF LEONARDO DA VINCI. Clifford A. Truesdell in *Essays in the History of Mechanics.* Springer-Verlag, 1968.

LEONARDO DA VINCI. MECHANICS. Marshall Clagett in *The Dictionary of Scientific Biography: Vol. 8,* edited by Charles Coulston Gillispie. Charles Scribner's Sons, 1973.

THE AMATEUR SCIENTIST

HOLOGRAPHY HANDBOOK: MAKING HOLOGRAMS THE EASY WAY. Fred Unterseher, Jeannene Hansen and Bob Schlesinger. Ross Books, 1982.

SEEING THE LIGHT: OPTICS IN NATURE, PHOTOGRAPHY, COLOR, VISION, AND HOLOGRAPHY. David Falk, Dieter Brill and David Stork. Harper & Row, Publishers, 1986.


HI WIPER
LO WIPER
WIPERS OFF
DELAY PUSH
MIN MAX
PUSH TO WASH
DEFROST
R DEFOG
TRIP SET
GAGE
GCC GRAPHIC CONTROL CENTER
SOMEDAY, POOR VISIBILITY WILL BE SOMETHING YOU JUST DON'T SEE ANYMORE.
At GM, our concern for safety goes beyond safety belts and disc brakes. We're working on advancements in electronics to help you actually avoid accidents.
Night vision is an infrared imaging system. It senses the difference in temperature between an object and its background, then displays an enhanced image of what's out there on a screen. GM is working on the technology now, and somewhere down the road, it could be in your car.
So in fog, rain, snow or darkness, you'll be better able to see what's on the road when you can hardly see the road.
THE GM ODYSSEY: SCIENCE NOT FICTION
GM