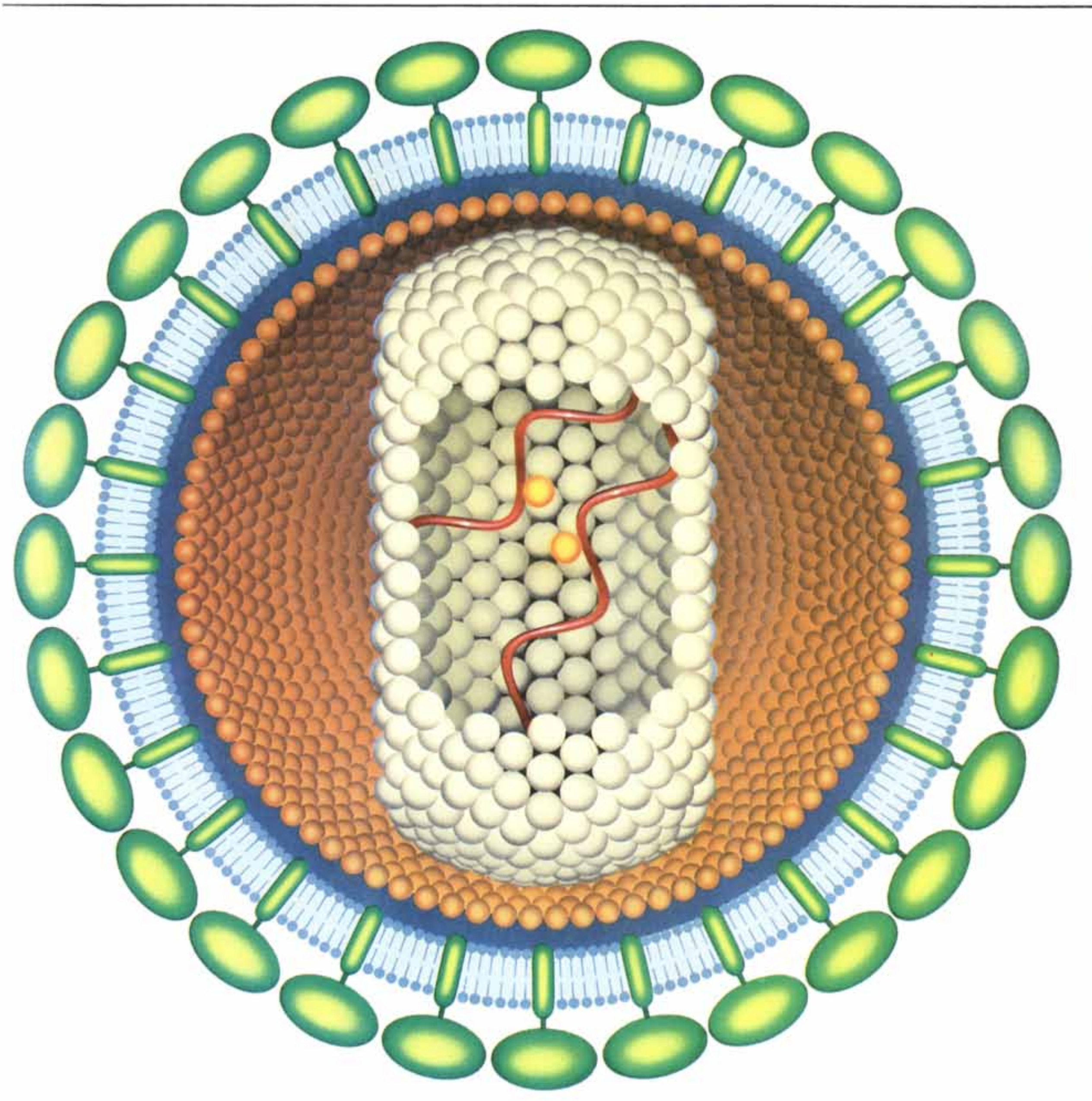
SCIENTIFIC AMERICAN



THE AIDS VIRUS

$2.50

January 1987

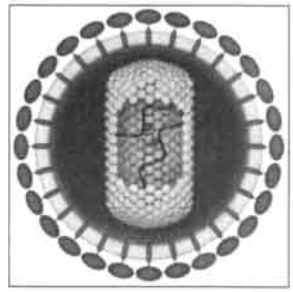

One good turn
deserves another.

Established 1845 **SCIENTIFIC AMERICAN** January 1987 Volume 256 Number 1

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $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.



THE COVER

The painting on the cover depicts the human retrovirus that causes AIDS: human *T*-lymphotropic virus III (see "The AIDS Virus," by Robert C. Gallo, page 46). The virus particle, about one ten-thousandth of a millimeter in diameter, is covered by an outer envelope consisting of a two-layer membrane (*blue*) studded with proteins (*green*); other proteins form a core structure (*light brown and ivory*). Like other retroviruses, HTLV-III carries RNA as its genetic material (*red*) along with an enzyme called reverse transcriptase (*yellow*). When the virus infects a cell, reverse transcriptase assembles DNA corresponding to the viral RNA. The DNA integrates itself into the chromosomes of the host, where it can later be activated in a burst of viral replication that kills the cell. Because the main host of HTLV-III is a white blood cell called the *T*4 lymphocyte, which has a crucial role in the immune response, infection by the virus often leads to a profound immune deficiency.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin, Science Graphics

Page	Source
15	Oliver Strimpel, The Computer Museum, Boston
16	Edward Bell
20	Edward Bell (*top*), Pat Macaluso (*bottom*)
30–37	Gabor Kiss
39	Jet Propulsion Laboratory/National Aeronautics and Space Administration
40–41	Hank Iken
42	Jet Propulsion Laboratory/National Aeronautics and Space Administration
43–45	Hank Iken
46	Zaki Salahuddin
48–56	George V. Kelvin, Science Graphics
55	Phyllis J. Kanki
75	Stephen Dalton, Photo Researchers, Inc.
76–78	Jerome Kuhl
79	Tom Prentiss
80–83	Jerome Kuhl
87–92	Thomas C. Moore
95	Paul Meakin
96–97	Andrew Christie
98	Andrew Christie (*top, bottom left*), Nancy Hecker and David G. Grier (*bottom right*)
99	Leonard M. Sander
100	David G. Grier (*top left, bottom right*), Eshel Ben-Jacob (*top right*), L. Niemeyer and H. J. Wiesmann (*bottom left*)
102	Pieter R. Cullis and Michael Hope, University of British Columbia
104–106	Patricia J. Wynne
107	The Liposome Company, Inc. (*top, bottom left*), Richard E. Pagano (*bottom right*)
108	Gabriel Lopez-Berestein
109	Carl R. Alving
110	Gabriel Lopez-Berestein
111	Patricia J. Wynne
113	Stefan Meyer, Animals Animals (*top*); K. Stepnell, Taurus Photos, Inc. (*bottom*)
114–116	Patricia J. Wynne
117	Michael L. Ryder
118	Thea Gabra-Sanders (*top*), Michael L. Ryder (*bottom*)
119	Ian R. Pitkethly (*top*), Michael L. Ryder (*bottom*)
121	Thomas Gold (*top*), Craig F. Bohren (*bottom*)
122–126	Michael Goodman

SCIENCE/SCOPE®

Advanced aircraft are being developed with the aid of computerized simulators that mimic aerodynamic and flight characteristics. Visual displays with 360-degree field-of-view projections create lifelike airborne situations, including tactical scenarios, weapons fire, and the effects of weather. Engineers use simulation data to improve aircraft design and performance. Hughes Aircraft Company supplies leading aircraft manufacturers with visual system components and other simulation hardware and software for flight simulators. Hughes recently installed a simulator subsystem at General Dynamics in Texas.

Direct communication between South Pole scientists and the U.S. is possible now for the first time through ATS 3, the third Applications Technology Satellite. NASA's Goddard Space Flight Center in Maryland installed a satellite antenna system at the South Pole to send and receive VHF signals from ATS 3. After over 18 years in geosynchronous orbit—well beyond the design life of seven years—the Hughes satellite is one of the last to use the low VHF frequency. Depleted of its positioning fuel, ATS 3 drifts daily into tracking ranges of ground stations at the South Pole and at the University of Miami in Florida. Communications are available for about four hours a day. The Antarctic project is a joint venture of NASA, the National Science Foundation, and private industry.

A new infrared viewer combines numerical temperature readouts and thermo-electric cooling to spot heat leaks and other energy losses more efficiently. The device is the latest model of Probeye® viewers from Hughes. As all units in the line, the Model 699 viewer sees heat the way a camera sees light and instantly converts it to a visual image. It can be used for pinpointing heating and cooling leaks and other maintenance problems in industry and commerce. A continuous digital display shows temperatures of objects in degrees Celsius or Fahrenheit. All-electric cooling eliminates the need for argon gas or liquid nitrogen, thereby cutting weight, making it easier to use, and removing restrictions by airlines and other common carriers on transporting pressurized devices.

A new infrared "eye" will help fight drug smugglers by letting U.S. Customs Service pilots see through darkness, smoke, and haze. The AN/AAQ-16 Hughes Night Vision System (HNVS), developed by Hughes, is a low-cost, computer-controlled infrared system that aids the pilot in navigation and surveillance. The AN/AAQ-16 measures minimal temperature differentials and produces a black-and-white TV-like picture for viewing on a cockpit display. The system's infrared sensor is housed in a turret mounted under the nose of the customs service's Piper Cheyenne 111A aircraft. Unlike a radar, it emits no energy and therefore cannot be detected by smugglers. The system is presently in use on both aircraft and helicopters.

Hughes Research Laboratories need scientists for a spectrum of long-term sophisticated programs, including: artificial intelligence knowledge-based systems and computer vision; applications of focused ion beams; electron beam circuit testing; liquid-crystal materials and displays; nonlinear optics and phase conjugation; submicron microelectronics; plasma applications; computer architectures for image and signal processors; GaAs device and integrated circuit technology; optoelectronic devices; and materials and process technologies for high-speed infrared detection and optoelectronic applications. Send your resume to Professional Staffing, Hughes Research Laboratories, Dept. S2, 3011 Malibu Canyon Road, Malibu, CA 90265. Equal opportunity employer. U.S. citizenship required.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068

HUGHES
AIRCRAFT COMPANY

Subsidiary of GM Hughes Electronics

LETTERS

To the Editors:

Recently an error in the August 1986 issue of *Scientific American* was brought to my attention. The first paragraph of the article "Progress in Laser Fusion," by R. Stephen Craxton, Robert L. McCrory and John M. Soures, refers to one of "the world's largest laser systems, the NOVA laser at the Lawrence Livermore National Laboratory in Berkeley."

To confuse the Government-owned facility for weapons research known as the Lawrence Livermore National Laboratory with the basic-science-oriented Lawrence Berkeley Laboratory, which is situated on a university-owned site overlooking the Berkeley campus of the University of California, is to give a very warped emphasis to Ernest O. Lawrence's scientific achievements. The confusion also propagates a most unfortunate distortion of scientific history.

This confusion has been increasing at an alarming rate since the start of today's accelerated nuclear-arms race, which has brought the Livermore laboratory an unusual degree of public notice and notoriety. At this point the Berkeley laboratory, where no classified work is done, is threatened with a loss of identity because of the prominence of the Livermore facility. It seems to have been forgotten that Livermore was established in 1952 as a subsidiary of Lawrence's famous Berkeley-based Radiation Laboratory, now called the Lawrence Berkeley Laboratory.

MARY B. LAWRENCE
(MRS. ERNEST O. LAWRENCE)

Berkeley, Calif.

To the Editors:

The opening paragraph of "Progress in Laser Fusion" describes the location of the NOVA laser as the "Lawrence Livermore National Laboratory in Berkeley." As the Representative of California's Eighth Congressional District, which includes Berkeley, I am compelled to advise you that the Lawrence Livermore National Laboratory is in Livermore.

This is a major concern for our area because of the military applications of Livermore's NOVA laser. The use of orbiting lasers is a linchpin of the Strategic Defense Initiative strategy to intercept ballistic missiles in the postboost phase and is a very dangerous extension of the arms race into outer space.

The Lawrence Berkeley Laboratory, which is widely known for its work on more peaceful laser applications, has often been the victim of misidentification as a weapons laboratory. In order to alleviate such confusion, and because of my deeply held belief that the research activities of the Livermore facility are a grave deviation from the peaceful applications of atomic energy advocated by Dr. Lawrence, I have drafted legislation to remove his name from the Livermore laboratory title. This bill, known in the 99th Congress as H.R. 1036, will soon be reintroduced. Hearings on the legislation may be conducted early in the 100th Congress before the Committee on Armed Services.

RONALD V. DELLUMS

Member of Congress
8th District, California
Washington, D.C.

EDITOR'S NOTE: The error was introduced during final editorial changes and not by the authors.

To the Editors:

I read with interest "The Colors of Things," by Philippe Brou, Thomas R. Sciascia, Lynette Linden and Jerome Y. Lettvin [SCIENTIFIC AMERICAN, September, 1986], and I thought I would relate an experience I had years ago. On a clothesline in her yard a neighbor had hung some laundry, which appeared to me to consist of navy blue bed sheets. The color caught my eye because navy blue is an unusual color for sheets. I saw them against a background of trees that were brightly lighted by the sun and assumed that the sheets were also sunlit. After some moments I suddenly realized that they were, in fact, shaded by the house, and it immediately became clear to me the sheets were white! The actual photons that were entering my eyes had not changed a bit, but my perception of the color had changed drastically—an apparent case of "discounting the illuminant."

It is stated that one cannot tell what light illuminates a surface, but it seems to me that to some extent at least one can. It does not take scientific instruments to tell that the rosy light of dawn is a different color from the yellow afternoon sun; this can be seen with the eye. Perhaps what happens is that one compares the *relative* colors of things. If everything in one's field of view has a bluish cast, for instance, perhaps one deduces that the illuminant is blue and somehow subtracts a degree of blueness from everything.

SHEILA BISHOP

Youngstown, Ohio

To the Editors:

Sheila Bishop's tale of the sheets reflects good observation. But the sudden emergence of "white" from "navy blue" as sheet color is not necessarily an act of mind, a cognitive change. This kind of effect can be had, for example, from seeing the sheets for a while by peripheral vision and then suddenly looking at them directly; the writer might try that as an experiment. She is in good company with the great Helmholtz in her notion of how to discount the illuminant. Still, we adhere to our view of the mechanism, however crudely it was put.

We owe readers an apology. An overlooked textual error led to improper captioning of the display, numbered 3, on page 88. The point of this display is that the colors of the four outlying hexagons are *not* changed by randomizing the annular order of the colored patches bounding them.

The last sentence of the third full paragraph on page 89 should read: "The gray hexagons do not differ much in their colors from what appears in the first display." In the next paragraph the last sentence should be: "Yet now the gray hexagons no longer differ much from one another in their color."

The caption for display number 3 should simply read: "Rearranging only the rings of colored hexagons that border each gray hexagon in the first display does not change the colors of the gray hexagons that are seen in that display."

PHILIPPE BROU

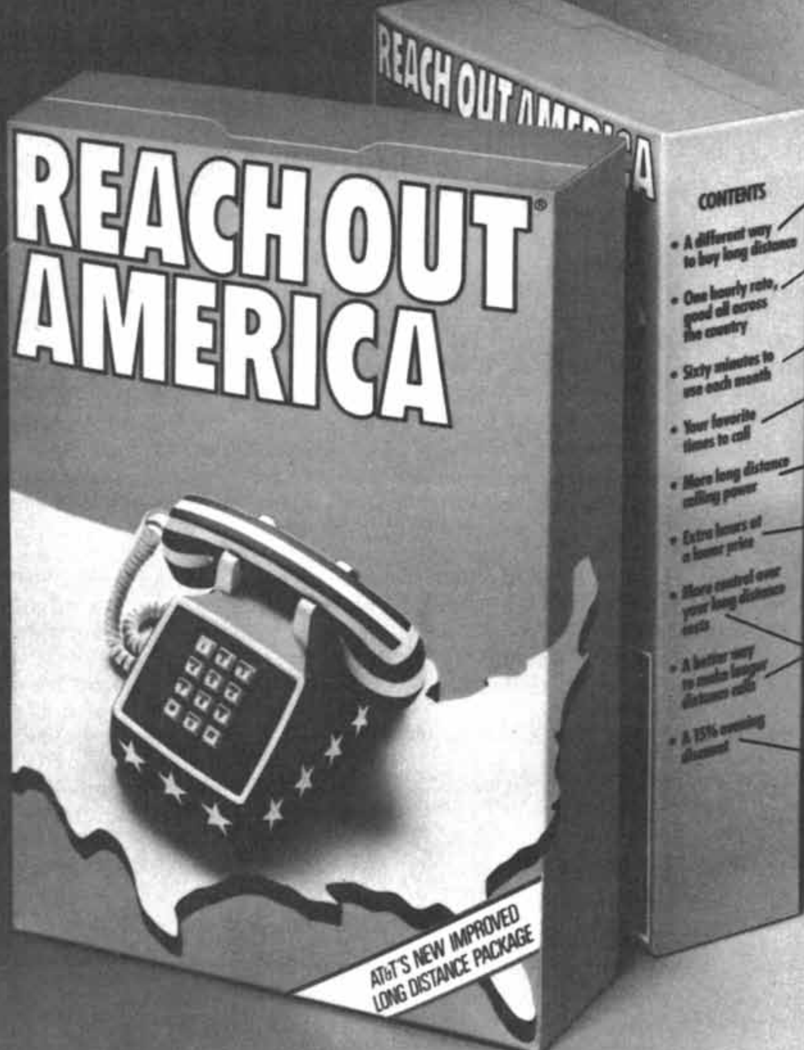
THOMAS R. SCIASCIA

LYNETTE LINDEN

JEROME Y. LETTVIN

EDITOR'S NOTE

The December "Books" department, in which books about science for young people were reviewed, was written jointly by Philip Morrison and Phylis Morrison, whose name was inadvertently omitted.



1. Make direct-dialed AT&T Calls to any other state—and pay by the hour, not by the mile.
2. Call all across the U.S.A.—even to Alaska and Hawaii—for one hourly rate.
3. Use your full hour, pay only $10.15. And get a special bonus, too!
4. Call all day Saturday, Sunday till 5 p.m., and every weeknight after 11 p.m.—till 8 a.m.
5. Make one 60-minute call...or as many as sixty 1-minute calls. Use your hour any way you like!
6. Talk for more than an hour, pay even less—just $7.80 for extra weekend and night hours.
7. Make your farther away calls weekends and nights—when you pay a flat hourly rate.
8. Keep track of the hours you talk, and you can keep track of your long distance costs.
9. Enjoy a special bonus: 15% off evening calls Sunday–Friday, from 5 p.m. to 11 p.m.
10. Clip and mail the coupon below for your valuable free guide, or call toll free: **1 800 CALL ATT, ext. 615**

10 WAYS YOU COULD GET MORE FOR YOUR MONEY WITH THIS AT&T PACKAGE

AT&T'S REACH OUT® AMERICA PACKAGE GIVES YOU AN HOUR'S WORTH OF CALLS—ALL ACROSS THE COUNTRY—FOR JUST $10.15. BUT ONLY WHEN YOU ASK FOR YOUR FREE GUIDE!

AT&T has packaged long distance for value. With "Reach Out" America, an innovative long distance package.

It's not for everyone. But if you're willing to spend $10.15 each month,* it could be right for you.

You'll get an hour of AT&T Calls any weekend, any night—to any other state. Plus a special bonus: a 15% discount on evening calls.

In fact, the more AT&T Long Distance calls you make, the farther you call, and the longer you talk...the better "Reach Out" America** could be for you!

AT&T
The right choice.

GET YOUR FREE GUIDE TODAY!

Our free guide will show you how you could get more for your money. To get your copy, just use the coupon below or call toll free: **1 800 CALL ATT, ext. 615**

*Plus a $10 order-processing charge when you sign up. A 3% Federal excise tax will be added to your monthly bill, and where applicable, a state surcharge.
**Not available in all areas. © 1987 AT&T

Clip and mail to:
AT&T Reach Out® Services
P.O. Box 418176
Kansas City, Missouri 64141-0176

☐ YES! I'd like to see how I could get more for my money. Send me my free guide, and all I need to sign up for AT&T's "Reach Out" America plan.

Name ______________________ Date __________
Address ______________________
City ____________ State ____________ Zip __________
Home Telephone (________) ______________________
area code

615

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JANUARY, 1937: "Motorists now may drive from Atlantic City directly into San Francisco without leaving the Lincoln Highway (U.S. 40). This is possible with completion of the 9.1-mile San Francisco–Oakland Bay Bridge. The bridge cost 77,200,000 dollars, which will be paid for from automobile, truck, interurban electric train and passenger tolls. Whereas vehicular traffic already is using the bridge, electric railway crossings will not commence until January, 1938."

"The total road mileage of the world is 9,268,397, or one mile of road to every 5.3 square miles of the total land area of 49,411,882 square miles in the world. The United States has one mile of road to every square mile; Japan, one to each .2 square mile; France, one to each .5; the United Kingdom, one to .5, and Germany, one to .8."

"The latest aviation altitude record was set by Squadron Leader F. R. D. Swain of the Royal Air Force when he reached a height of 49,967 feet. Such records serve not only to thrill the world but also to provide scientific information for stratosphere flying in the future."

"Color has reached greater perfection in movie cartoons than in real-life motion pictures. This is true not because of technical advances made by the cartoon manufacturers themselves but because on their stage, which consists of a series of celluloid sheets containing paintings superimposed over a water-color background, they can control both lighting and action with machine-like precision."

"In several recent instances the scientific prestige of old classical writings often thought of as possibly inaccurate has been enhanced when things described in them thousands of years ago have actually been dug up by modern archaeologists. Lucian, an ancient Greek writer, described a certain statue of Apollo at Athens. Recently the American School of Classical Studies at Athens, led by Prof. Leslie T. Shear of Princeton University, found more than 200 fragments of a statue, which they reassembled. It is evidently the statue Lucian saw."

SCIENTIFIC AMERICAN

JANUARY, 1887: "Sanitation may not show a progress equal to that achieved in locomotion or metallurgy over the past 50 years, but it is of far more general interest, as it comes home to immensely greater numbers. It matters to only a few that the journey can now be made from London to Edinburgh in nine hours instead of ninety, but everyone is interested in the fact that immunity can be obtained from smallpox, which formerly was as common as measles. Health is the basis of happiness, and the healthier people become, the happier they are, and thus every step gained in sanitary science distributes benefits to the entire community."

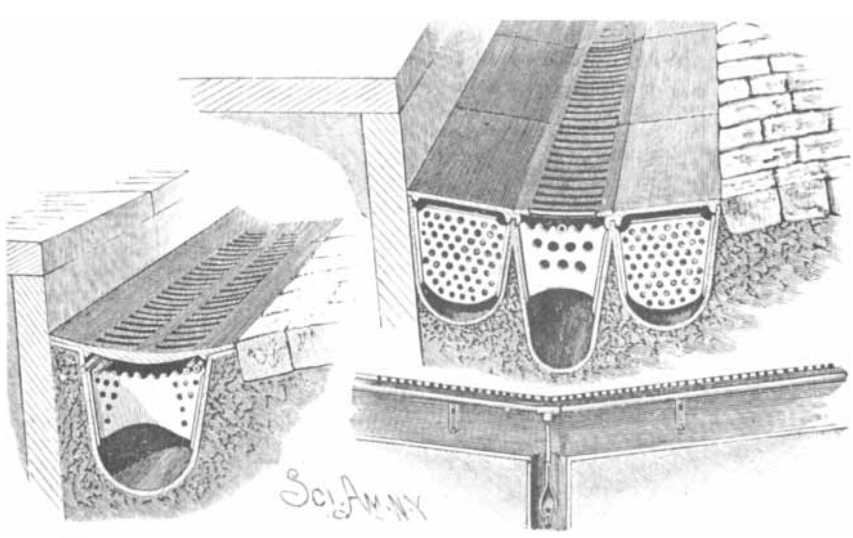
"The Russian government is engaged in one of the most extensive drainage enterprises ever undertaken in any portion of the world. The location is what is known as the Pinsk Marshes, in the southwest of Russia, near the borders of Galicia. About 4,000,000 acres have been reclaimed by means of the construction of several thousand miles of ditches and canals. When the task is finished, Russia will have effaced from the map of Europe one of the oldest and toughest regions of savage nature on the Continent."

"The question of submarine warfare would appear to be advanced an important stage by a new submarine torpedo boat which was lately tried in London. The great problem for solution in this kind of boat is a simple and ready means of effecting submersion quickly and of again rising to the surface as frequently as may be desired. The principle upon which the immersion and emersion of the new boat depend is simply that of displacement."

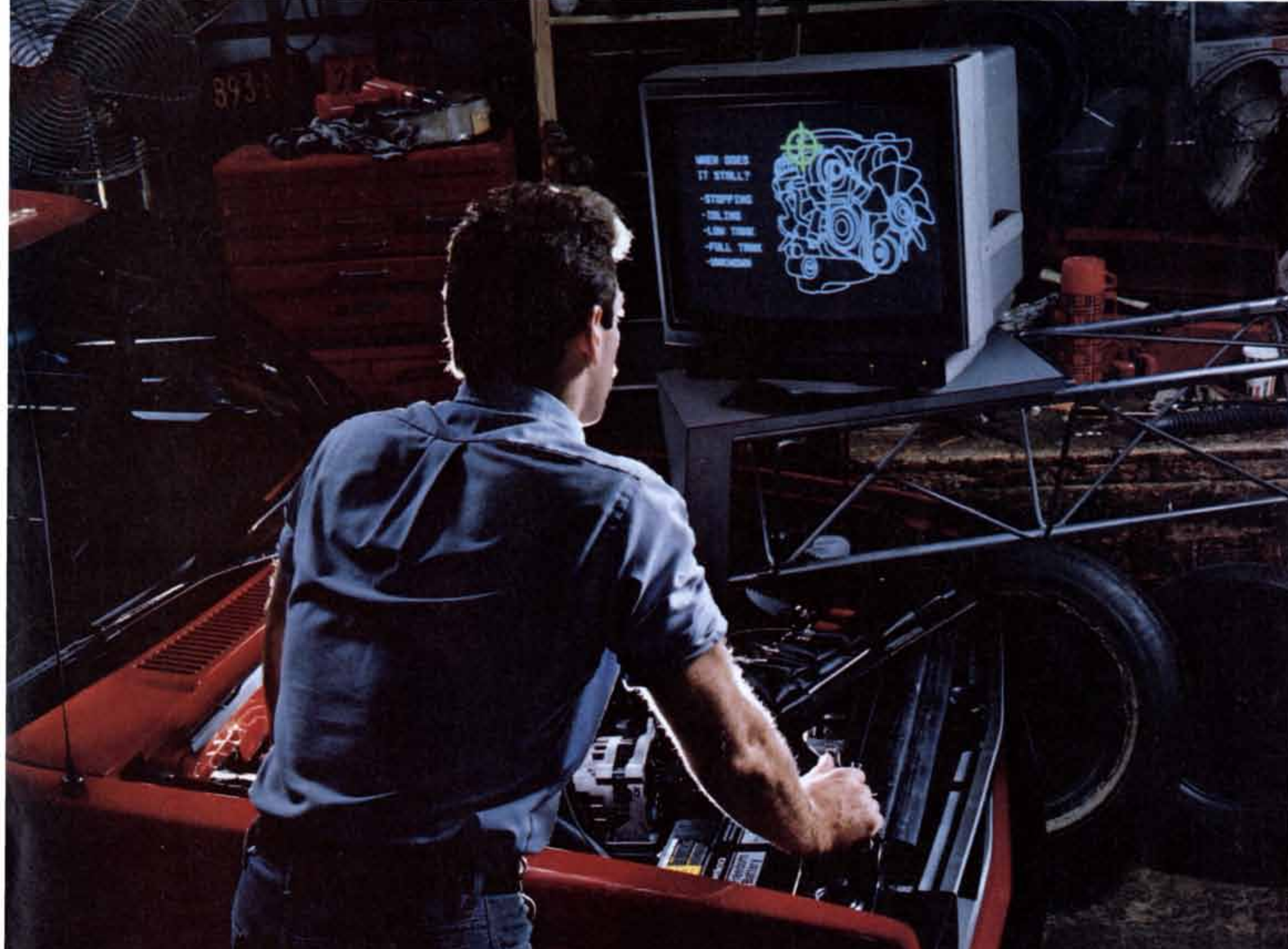
"One of the most striking mechanical works is the great crane Titon, which is now at work in the port of Leixões, Portugal, employed in placing the artificial stone blocks, 50 tons weight each, for the construction of the breakwater. The larger arm measures 46 meters from the axis of the machine, and the shorter 22¾ meters, making a total length of 68¾ meters. Its total weight is 450 tons."

"During the year ending Sept. 30, 1886, there were carried on the street railways of New York City a total of 325,427,015 passengers. We believe this is by far the greatest passenger traffic of any city in the world, even though New York is not the largest city."

"How to dispose of the snow: this is one of the most serious problems connected with the comfort and convenience of a great city in this latitude. A new method proposed by S. D. Locke of Hoosick Falls, N.Y., would utilize the steam plants existing in most cities to melt the snow. Underneath the surface gutter he proposes to construct a sub-gutter that connects directly with the sewer and that is covered with a grate, underneath which steam pipes are carried in racks. The snow can as quickly be moved by horse scrapers and brooms into the gutters as the streets can now be swept."

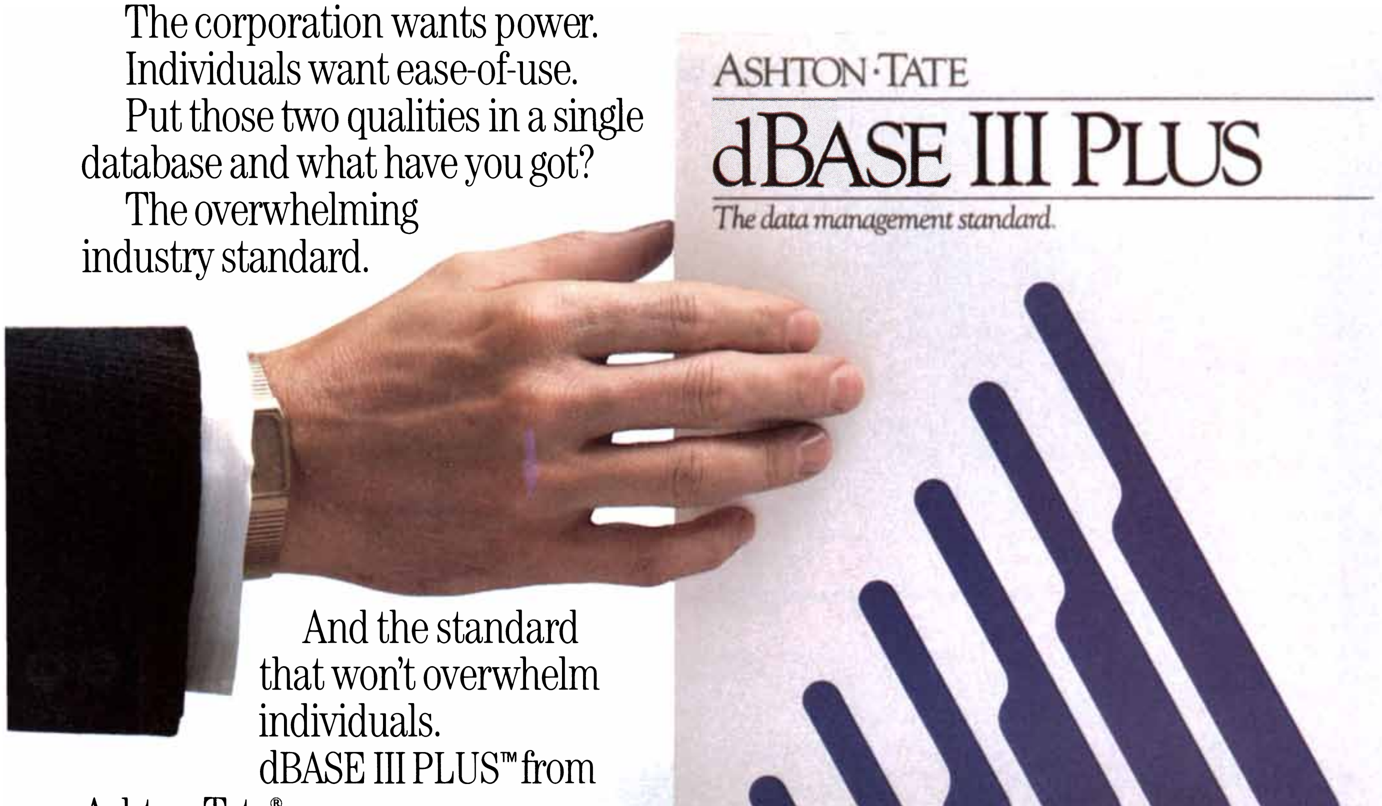


Arrangement for removing snow with steam



Every good b a double

The corporation wants power.
Individuals want ease-of-use.
Put those two qualities in a single database and what have you got?
The overwhelming industry standard.



And the standard that won't overwhelm individuals.
dBASE III PLUS™ from Ashton-Tate.®

The Industry Standard

More corporations choose it than all the other databases combined. Because they want broad compatibility across the whole company. And true networking capability. And the opportunity to unify their support and training resources. In short, they want one product that's flexible and powerful enough to meet the needs of all.

And with dBASE III PLUS, they get it all. Plus the broadest third-party support in the industry. You see, more applications are written for dBASE® than for any other database. So in many cases, the applications you need are already available.

More individuals choose dBASE as well.

Because we've made it easier for individuals to use. Particularly the sort who get queasy when they think they might have to program.

They can start using dBASE without

usiness needs standard.

The Individual's Standard

any programming experience. Because the Assistant feature, with its pull-down menus, guides them step-by-step through the creation and use of a database.

When they're ready to create applications, our Applications Generator writes the code.

And to make the complex even simpler, we have catalogs that let them recall relationships among files with a single command. Complex commands can also be stored for automatic use by our "history" capability.

Best of all, we've added these ease-of-use features without alienating the people who made dBASE the industry standard in the first place. The advanced users and programmers. Who still get all the power and programming features they need to develop their sophisticated applications.

So if you're looking at databases from either the corporate or the individual perspective, your choice is simple.

dBASE III PLUS.

Or dBASE III PLUS.

For the authorized Ashton-Tate dealer nearest you call (800) 437-4329, Ext. 2844.*

ASHTON·TATE®

*In Colorado and outside of the U.S., call (303) 799-4900, Extension 2844. Trademarks/owner: dBASE, dBASE III PLUS, Ashton-Tate/Ashton-Tate. © 1986 Ashton-Tate. All rights reserved.

THE AUTHORS

LYNN R. SYKES and DAN M. DAVIS ("The Yields of Soviet Strategic Weapons") are geophysicists who supplement their basic research with studies of underground testing of nuclear weapons and its implications for arms-control parleying. Sykes, who is Higgins Professor of Geological Sciences at Columbia University, has studied problems of detection and identification in nuclear testing for the past 25 years and was a member of the U.S. delegation that negotiated the Threshold Test Ban Treaty in Moscow in 1974. After receiving a B.S. and an M.S. at the Massachusetts Institute of Technology in 1960, he earned his Ph.D. from Columbia in 1964; he has held his present position since 1968. A major contributor to plate-tectonic theory, Sykes now studies earthquakes at Columbia's Lamont-Doherty Geological Observatory. Davis was a postdoctoral fellow and research geophysicist at Lamont-Doherty before joining the department of earth and space sciences at the State University of New York at Stony Brook last September. He received his A.B. at Princeton University in 1978 and his Ph.D. in geophysics from M.I.T. in 1983. Davis' primary interest is the mechanics of crust deformation.

ANDREW P. INGERSOLL ("Uranus") became involved in the *Voyager 2* mission even before its first encounter with Jupiter in 1979. Currently a professor of planetary science at the California Institute of Technology, he has devoted his career to atmospheres and climates, an interest first sparked by undergraduate projects at Amherst College. He studied at Harvard University for both his A.M., awarded in 1961, and his Ph.D., which he received in 1965. In addition to the Uranus flyby, he has taken part in several Pioneer planetary missions and is particularly proud of his work with VEGA, a Soviet space project supported by American and French scientists.

ROBERT C. GALLO ("The AIDS Virus") was first drawn to cancer research when, as a 14-year-old, he watched his sister die of leukemia. He is now head of the Laboratory of Tumor Cell Biology at the National Cancer Institute, where he has worked since 1965. He has made several substantial contributions to cancer research, the most notable being his discovery of the link between retroviruses and human leukemia and lymphoma. In 1981 he assembled a team to spearhead AIDS research and in 1984 he identified HTLV-III, the causal virus. He holds a B.A. from Providence College and earned his M.D. from Jefferson Medical College in 1963. His work has won him many distinctions, including two Lasker awards for medical research in the past five years. This month's article follows Gallo's report in the December issue on the first human retrovirus, HTLV-I.

JAMES L. GOULD and PETER MARLER ("Learning by Instinct") are respectively professor of biology at Princeton University and professor of animal behavior at Rockefeller University. Gould got a B.S. in molecular biology at the California Institute of Technology in 1970, then did research on animal behavior for his Ph.D., which he received from Rockefeller University in 1975. He joined Princeton's faculty the same year. Although his primary interests are the communication, navigation and learning of honey bees, Gould is also investigating innate recognition in birds and sexual selection in guppies. Marler studied botany at the University of London for a B.Sc. (awarded in 1948) and a Ph.D. (1952), and in 1954 he earned a second Ph.D., in zoology, from the University of Cambridge. After nine years on the faculty of the University of California at Berkeley, he went on to Rockefeller in 1966. Marler now studies animal communication at the university's field research center in Millbrook, N.Y.

TERRY R. PENNEY and DESIKAN BHARATHAN ("Power from the Sea") are aerospace engineers at the Solar Energy Research Institute (SERI) in Golden, Colo. Penney, a section manager in the thermal-sciences research branch, joined SERI in 1979 after completing an investigation of two-phase flow in natural-draft cooling towers for the Electric Power Research Institute. He was graduated from Purdue University in 1973 with a B.S., then went on to the University of Tennessee for an M.S. in mechanical engineering, awarded in 1975. He also worked as a wind-tunnel project engineer at the Arnold Engineering Development Center in Tullahoma, Tenn. Bharathan was born in Kumbakonam, India, and studied at the Indian Institute of Technology in Madras before receiving a Ph.D. from the University of Virginia in 1976. For four years he was research associate professor at Dartmouth College, where he examined the safety aspects of nuclear power. He moved to SERI in 1980 and now directs research in ocean thermal energy as a senior engineer. Bharathan is cochairman of the solar-energy group in the American Institute of Chemical Engineers' heat-transfer and energy-conversion division.

LEONARD M. SANDER ("Fractal Growth") became curious about fractals and nonequilibrium growth while indulging in a favorite hobby: computer hacking. As professor of physics at the University of Michigan he ordinarily conducts experiments in solid-state theory and statistical physics. He was an undergraduate at Washington University in St. Louis and earned a Ph.D. in physics from the University of California at Berkeley in 1969. Sander spent a year as a postdoctoral fellow at the University of California at San Diego before going to Michigan in 1969.

MARC J. OSTRO ("Liposomes") was the first employee of the Liposome Company, Inc., and now holds the position of vice-chairman and chief scientific officer. He received his B.A. in biology at Lehigh University and his Ph.D. in biochemistry from Syracuse University. He worked as a postdoctoral fellow and associate professor of microbiology and immunology at the University of Illinois Medical Center before leaving in 1981 to direct the company's start-up. Ostro maintains his ties to academia as adjunct associate professor at the University of Medicine and Dentistry of New Jersey in Piscataway.

MICHAEL L. RYDER ("The Evolution of the Fleece") became interested in sheep while working for the Wool Industries Research Association in Leeds, England, in 1951. Since then he has traveled to more than 40 countries and written 200 articles, combining a career in biology with historical and archaeological studies of livestock. He got a B.Sc. in 1951 at the University of Leeds, then did research on wool follicles for nine years, in the process earning an M.Sc. in 1954 and a Ph.D. in 1956 from Leeds. After two years at the University of New England in Australia, he moved to Edinburgh in 1962 to take a position with the Agricultural Research Council's Animal Breeding Research Organisation, eventually becoming principal scientific officer. In 1984 he joined the staff of the Hill Farming Research Organisation. Ryder plans to create an international research institute and museum of sheep and wool history when he retires later this year.

COMPUTER RECREATIONS

A program called MICE *nibbles its way to victory at the first Core War tournament*

by A. K. Dewdney

Core War—the game in which specialized programs do their level best to destroy one another—was in the spotlight late last year at the first international Core War tournament held at the Computer Museum in Boston, Mass. Of 31 programs entered, three emerged as most robust. The ultimate victor was a program called MICE. Its author, Chip Wendell of Rochester, N.Y., received a handsome trophy that incorporated a core-memory board from an early CDC 6600 computer.

Core War has already appeared twice in this department in recent years [see "Computer Recreations," May, 1984, and March, 1985]. Written by human beings, the Core War programs are on their own as they spar in the arena of a computer's memory. The section of memory reserved for the struggle is called the core, after an obsolete form of memory constructed from miniature ferromagnetic rings known as core elements. The game has generated so much enthusiasm that it has sparked the formation of the International Core Wars Society. The game was recently modified by the society; the new version lays out the format that players will follow for now.

The basis of Core War—and the ammunition of the recent tournament—is a battle program written in a special, low-level language called Redcode. A set of 10 simple instructions enables a program to move information from one memory location to another, to add and subtract information, to alter the order in which its instructions are executed and even to have several instructions executing simultaneously [*see illustration on page 16*]. One basic instruction, for example, is the move command MOV. It consists of three parts—an instruction code and two addresses—that all occupy the same location in the core. The command is most generally written as MOV *A B*. If *A* happens to be 102 and *B* is −5, the computer will go forward 102 addresses and copy what it finds there into the location five addresses behind the MOV instruction.

The simplest Redcode program consists of just one MOV instruction: MOV 0 1. The program, which is called IMP, causes the contents at relative address 0 (namely the MOV instruction itself) to be transferred to relative address 1, just one address ahead of itself. Redcode instructions are normally executed consecutively. This means that after the MOV 0 1 instruction is executed the computer will try to execute an instruction at the next address. There is, of course, now an instruction occupying that address: the MOV 0 1 command just copied there. As a consequence IMP patters from address to address through the core, mindlessly destructive. It leaves a trail of MOV 0 1 instructions behind it.

An IMP can even steal an enemy program's very soul, its execution. To see how this can happen, imagine that a battle program is being executed in the usual manner, in the order of its instructions. An IMP enters the program from the top, overwriting the code with an endless sequence of MOV 0 1 instructions. Sooner or later the subverted program will probably transfer execution back to the overrun section. At such a point the program becomes a new IMP. It flies the same flag but is now doomed to follow in the tracks of the enemy IMP until the battle is over.

To avoid being overrun a Core War program must at the very least contain an IMP-STOMPER. The safeguard consists of two instructions executed cyclically:

```
MOV #0 −1
JMP −1
```

The first command moves the integer 0, symbolized by #0, to the relative address −1; in other words, every time the MOV command is executed the location just above it (the only direction from which IMP's can attack) is filled with a 0. The second instruction is the JMP command. When it is executed, it transfers the stream of execution, or flow of control, to the address at relative location −1, namely the address just above the JMP. Each execution cycle of the program causes a 0 to be slammed down on any IMP that may have arrived just above the IMP-STOMPER. Consequently the IMP is erased.

There are two basic rules in Core War. The first rule is that the two competing programs must take turns executing their instructions. The alternation is governed by MARS, the Memory Array Redcode Simulator. As the somewhat strained military mnemonic suggests, MARS simulates the action of a computer. It continually updates the contents of the core array in accordance with the instructions being executed. In doing so, it allows just one instruction per side to be executed per turn. The second rule is that if a program ceases to run, it loses.

As a program runs, it can have more than one stream of execution. If execution encounters the command SPL *A* in a Redcode program, it splits into two streams. One stream goes to the instruction that immediately follows SPL *A* and the other jumps to the instruction at relative address *A*. Unfortunately the MARS system cannot execute both instructions simultaneously; it executes one of the instructions on the next turn and the other instruction on the turn after that. What might be thought an incredible advantage is somewhat adumbrated; the more concurrent streams of execution a program has, the slower each stream proceeds. This is only fair, however. In the case of multiple streams of execution a battle program is declared the winner when all its opponent's streams have died out. At such a point MARS, which would still expect to find an executable instruction, can find only the computational equivalent of shell holes and bomb craters.

To illustrate the SPL command, here are the first five instructions of my own entry in the Core War tournament. It is called COMMANDO for reasons that will soon become clear.

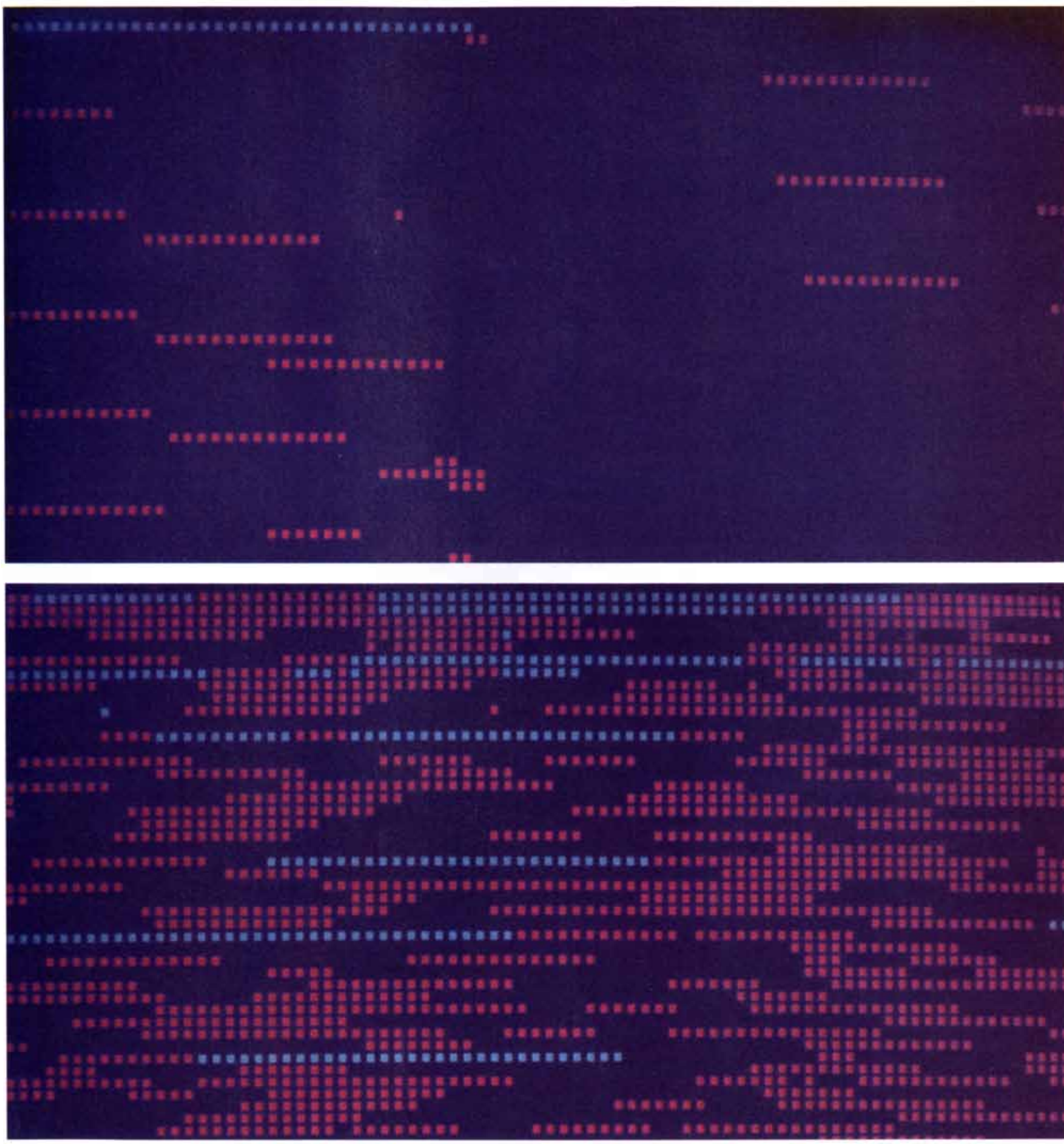
```
MOV #0 −1
JMP −1
SPL −2
MOV 10 113
SPL 112
  .
  .
  .
```

Readers will recognize an IMP-STOMPER in the first two instructions. Execution of the actual program begins at the third instruction, SPL −2. On COMMANDO's next two turns the first and fourth instructions will be executed. On the two turns after those, the second and fifth instructions will be executed. Each stream proceeds independently of the other and at half the speed, so to speak. In the code above, COMMANDO sets the IMP-STOMPER running on its own. Then it moves another IMP (patiently waiting 10 addresses beyond the second MOV instruction) to a distant location (113 addresses beyond). The second IMP is activated by the second SPL command.

COMMANDO's remaining instructions copy the entire program into a new segment of the core, 100 addresses beyond its present location. The new copy, like a commando just parachuted into enemy territory, is activated by a JMP command in the original program. The old copy of COMMANDO, except for the IMP-STOMPER, ceases to run. Then the entire cycle of copying begins again.

How would COMMANDO fare against its competitors? The tournament was organized to provide for as many engagements as possible between the 31 entries. A complete round robin, in which every program fought all others in turn, would have required 465 battles, more than time would allow. Consequently the entries were divided arbitrarily into two nearly equal groups, division I and division II. A round robin was then held within each division.

Imagine the strange mixture of emotions I felt when COMMANDO emerged as the winner of division II. On one



Early and late stages in a battle between MICE **(red)** *and* CHANG1 **(blue)** hand I was proud that my cybernetic child had done so well. At the same time I was somewhat mortified at the prospect of winning the tournament overall. Since I had consented to serve as a commentator for the finals, my objectivity (and credibility) would undoubtedly be strained.

The top four programs from each division were then entered in a new round robin. Three programs emerged victorious from the fray, CHANG1 by Morrison J. Chang of Floral Park, N.Y., and two entries by Chip Wendell, MIDGET and MICE. My COMMANDO fell by the wayside, mortally wounded. The final win by MICE came oddly; MIDGET and MICE both fought CHANG1 to a draw, but MICE captured the deciding point by beating MIDGET.

The contest between each pair of finalists consisted of four consecutive battles. The time limit on each engagement was 15,000 instructions per side, or approximately two minutes of real time. In each case the two battle programs were placed in random, nonoverlapping positions in the core and allowed to have a go at it. As it happened, each battle between a given pair of programs always had the same result. In the case of MICE versus CHANG1 the result was four draws.

It is fascinating to watch a Core War in progress. The display used at the tournament shows the core as a succession of cellular strips [*see illustration on preceding page*]. Each cell represents a single address in the core, and the last cell in the bottom row is contiguous with the first cell in the top row, in keeping with the circular structure of the core. The program that has the first move initially occupies address 0 and subsequently fills consecutive core locations. Its color is light blue. The opposition occupies a randomly selected segment of locations not overlapping those assigned to the first program. The color given to the second program is bright red. The color of a cell in the display is determined by the last program to alter the address it represents. In this way one has an engaging overview of the action.

Against a dark blue screen MICE and CHANG1 crept about, launched IMP's, hurled bombs and reproduced (parthenogenetically). One of the contests was typical: CHANG1 began as a strip of blue cells in the upper left-hand corner of the screen and the birth of MICE was heralded by a red strip that appeared less than halfway down the screen. Immediately MICE began to proliferate rapidly.

One of the shortest self-replicating programs I know, MICE has just eight instructions, two of which create a new copy of the program some 833 addresses beyond its present location in the core [*see top illustration on page 20*]. The two instructions demonstrate a few additional features of the Redcode language:

loop MOV @*ptr* < 5
DJN *loop ptr*

The word *loop,* which is simply a label that stands for an address, makes Core War programs easier to write. The DJN (short for decrement and jump on nonzero values) command causes execution to jump to the instruction labeled *loop* if the value stored at another address (labeled *ptr*) is not yet zero. The @ sign indicates a system of reference known as indirection; when the MOV command is executed, it does not move the contents of the location labeled *ptr* but instead moves the contents of the contents, so to speak. The number stored at *ptr* is the address of the datum to be moved. In this case the datum is one of MICE's instructions.

The number stored at *ptr* continually changes owing to the decrementing function of the DJN command. The number starts at the last program address and steadily decrements to zero, at which point the copying loop is finished. In a similar manner the address at which the instructions are to be stored is also given by indirection. The relative address 5 initially holds the number 833 and the first instruction moved by MICE lands 832 addresses beyond the MOV command; as indicated by the < sign, the target address is decremented and MOV is executed. MICE copies itself tail first.

An SPL (split) command immediately following the loop transfers execution to the new copy of MICE. But following this successful birth the parent program begins anew. There is no limit to how many progeny a single program of this type may produce. And each new program does the same thing. MICE, indeed!

So it was that in a typical contest with CHANG1, MICE bred with incredible rapidity. Soon the screen was full of little red strips. In the meantime CHANG1 had activated a kind of IMP factory at its downstream end. The factory was achieved with only three instructions:

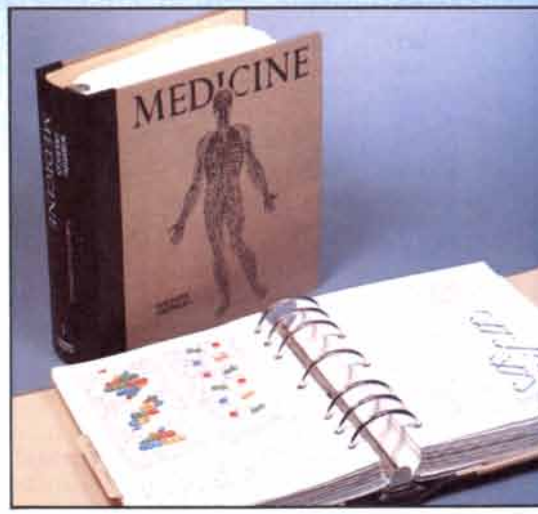
SPL 2
JMP −1
MOV 0 1

When execution arrives at the SPL command, it splits into two branches. One of them transfers execution to MOV 0 1. The other executes the JMP −1 instruction, which begins the process anew. In the meantime one IMP has already left the assembly line on a mousing mission. One problem with profligate IMP production is that a large number of independent streams of execution slows down every process executed; 1,000 IMP's move 1,000 times slower and more painfully than a single IMP. In any event, the fateful horde emerged slowly at the top of the display screen as an ever lengthening

INSTRUCTION	MNEMONIC	ARGUMENTS		EXPLANATION
Data statement	DAT		B	A nonexecutable statement; *B* is the data value
Move	MOV	A	B	Move contents of address *A* to address *B*.
Add	ADD	A	B	Add contents of address *A* to address *B*.
Subtract	SUB	A	B	Subtract contents of address *A* from address *B*.
Jump	JMP	A		Transfer control to address *A*.
Jump if zero	JMZ	A	B	Transfer control to address *A* if contents of address *B* are zero
Jump if not zero	JMN	A	B	Transfer control to address *A* if contents of address *B* are not zero
Decrement: Jump if not zero	DJN	A	B	Subtract 1 from contents of address *B* and transfer control to address *A* if contents of address *B* are not zero.
Compare	CMP	A	B	Compare contents of addresses *A* and *B;* if they are equal, skip the next instruction.
Split	SPL	A		Split execution into next instruction and the instruction at *A*.

A summary of Redcode, an assembly language for Core War

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.



A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.

Each month, six to nine replacement chapters to update your text *plus* new references, a four- to five-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the floppy-disk version for your IBM® PC/PC Jr., Macintosh™, or Apple® IIe or II+ (or compatible).

Your subscription includes:
the two-volume, 2,300-page loose-leaf text
and, each month, six to nine replacement chapters,
a newsletter, new references,
and a completely revised index.

Or call toll-free: 1-800-345-8112
(in Penn. 1-800-662-2444)

CINE

SCIENTIFIC AMERICAN MEDICINE
415 MADISON AVENUE, NEW YORK, N.Y. 10017

☐ **Yes,** I'd like to try the SCIENTIFIC AMERICAN *Medicine* system. Please enter my subscription at a first-year price of US$245*.
☐ Also enroll me in the CME program. I'd prefer:
  ☐ the printed version at **no additional charge.**
  ☐ the floppy-disk version at an additional $US97.*
  Computer type or compatible:
   ☐ IBM® PC/PC Jr. (128K RAM, DOS 2.0/2.1)
   ☐ Macintosh™ (512K RAM)
   ☐ Apple® IIe (64K RAM)
   Apple® II+ (64K RAM) 80-col. card by: ☐ Apple® ☐ Videx®
☐ Enroll me in the floppy-disk CME **only** at US$152.*
  (Note model above.)
☐ Check enclosed* ☐ Bill me

Signature ______

☐ VISA ☐ MasterCard Exp. Date ______

Account No. ______

Name ______

Specialty ______

Address ______

City ______ State ______ Zip Code ______

*Add sales tax for Calif., Ill., Mich., or N.Y. Allow 8 weeks for delivery. Add US$10 for shipping to Canada. IBM is a registered trademark of International Business Machines Corporation. Apple is a registered trademark of Apple Computer, Inc. Videx is a registered trademark of Videx, Inc. Macintosh™ is a trademark of McIntosh Laboratory, Inc. and is used by Apple Computer, Inc. with its express permission.

7G

CHANG1

```
        MOV   #0     -1
        JMP   -1
        DAT          +9
start   SPL   -2
        SPL   4
        ADD   #-16   -3
        MOV   #0     @-4
        JMP   -4
        SPL   2
        JMP   -1
        MOV   0      1
```

MICE

```
ptr     DAT   #0
start   MOV   #12    ptr
loop    MOV   @ptr   <5
        DJN   loop   ptr
        SPL   @3
        ADD   #653   2
        JMZ   -5     -6
        DAT   833
```

Contenders for the Core War championship

solid blue strip. Would they be able to subvert MICE?

While the IMP's were reproducing, some of the MICE copies were killed by data bombs from CHANG1. A data bomb usually consists of a 0 that is launched by a MOV command into what one hopes is enemy territory. The key instruction in Chang's program is MOV #0 @ − 4. The 0 is moved to an address contained in a location that is four instructions above the MOV command. The location is continually incremented by 16 to ensure a well-spaced barrage.

As some MICE were dying in this manner, the IMP's began to exert their destructive influence. But each copy of the original MICE program carries with it a suicide option; it continually checks on whether its first instruction (which is a data statement consisting of 0 alone) is still zero. If it is not, MICE allows execution to proceed to a (non-executable) data statement and dies quietly rather than lose its soul to a tiny fiend.

If some copies of MICE were being killed by data bombs and others were executing their own execution, so to speak, to avoid capture, how did MICE survive? The answer surely lies in its profligate spawning of new copies. Many of these, after all, landed on enemy IMP's. Indeed, before time was called one copy of MICE had landed on CHANG1's home program and destroyed it. CHANG1, however, had created enough IMP's to tide it over until the closing buzzer sounded. The battle was a draw.

The art of Core War programming is surely still in its infancy. Progress will be incremental and cumulative. Some intrepid programmer will discover infallible remedies against IMP's and another will discover simple means of self-repair. Readers wanting to keep abreast of the latest developments may subscribe to *The Core War Newsletter* by writing to William R. Buckley at 5712 Kern Drive, Huntington Beach, Calif. 92649. Readers who want to write battle programs should probably join the International Core Wars Society. Mark Clarkson currently directs the society and would welcome new members. He lives at 8619 Wassall Street, Wichita, Kan. 67210–1934. One does not have to join the society, however, to order the all-important "Core War Standards" document from Clarkson. It precisely describes the syntax and semantics of Redcode programs; its cost is $4. One cannot be a Core Warrior without it.

Battle programs of the future will perhaps be longer than today's winners but orders of magnitude more robust. They will gather intelligence, lay false trails and strike at their opponents suddenly and with determination. Such trends may already be in evidence at the second international Core War tournament to be held at the Computer Museum this fall. In the meantime readers have ample opportunity to express their cleverness and cunning in Redcode language.

Last fall's tournament owes much of its success to Mark and Beth Clarkson as well as to Gwen Bell, president of the Computer Museum, and Oliver Strimpel, its associate director and curator. It seems worthwhile to conclude with a brief note on the museum itself.

The Computer Museum in Boston is apparently the only museum in the world devoted entirely to computers. Housed in a renovated (and now chic) warehouse on the Boston waterfront, it features old vacuum-tube monsters, PC's for personal play, walls adorned with stunning graphics, a complete NORAD SAGE computer system and a host of exhibits that entertain and educate. Readers visiting a certain famous old ship in the Boston harbor can have their computational cup of tea right next door.

In this department last October I described a program called FACEBENDER, inspired by the work of Susan E. Brennan of Hewlett-Packard Laboratories in Palo Alto, Calif. As input the program takes the digitized version of a face to be caricatured, which it then compares with an average reference face, similarly digitized, in memory. The program then distorts, or exaggerates, each feature of the input face by an amount that is proportional to its distance from the corresponding feature in the reference face; an ear that is moderately large compared with the reference ear will be enlarged still further by multiplying all the differences by an exaggeration factor f.

Readers who want to implement the FACEBENDER program may have been daunted by the prospect of digitizing their own face from a photograph. Pat Macaluso of White Plains, N.Y., uses the reference face as the basis of its own caricature. "The key," says Macaluso, "is to scale the range of variation to the size of each feature. Thus an ear receives more absolute variation than the chin cleft. Simply calculate the enclosing 'box' by calculating the maximum and minimum of the x and y coordinates for each feature." Within this framework the amount of distortion is governed by random numbers selected by the program. In this way an endless variety of faces can be produced by Macaluso's self-referential version of FACEBENDER. One of the caricatures so produced resembled Leonardo da Vinci. It is shown below.

A reader known only as DMI from Pasadena, Calif., has a suggestion for avoiding "facelessness," the dreaded state that occurs when the exaggeration factor is too large; all features degenerate into a wild and unrecognizable bird's nest of polygons. Imagine that the face to be caricatured is superposed on the reference face and that corresponding points are connected by springs. The distortion process now attempts to displace the points of the input face, but in doing so it encounters resistance from the springs. Small distortions are thereby hardly affected, but large ones are pulled up short of the faceless state.

Self-caricature of the average face

BOOKS

Guano and its successors, fuel for fusion, the elegance of fractals and of gemstones

by Philip Morrison

FERTILIZER IN AMERICA: FROM WASTE RECYCLING TO RESOURCE EXPLOITATION, by Richard A. Wines. Temple University Press ($34.95). FERTILIZER TECHNOLOGY AND USE, THIRD EDITION, edited by O. P. Engelstad. Soil Science Society of America, Inc., 677 South Segoe Road, Madison, Wis. 53711 ($40). By 1850 New York had for a generation been the largest city in the country; it was the bustling commercial and manufacturing center we glimpse in the pages of Melville and Whitman. The hay to power its horses, the fresh produce on its tables and the fuelwood for its fires came from the intensively worked farms of nearby Queens and Nassau counties, celebrated among the "gardens of America."

The newly cleared lands of America could be farmed for about a century on the nutrients stored in the soil under vanished forest, and so the task of replenishment could be postponed. European observers were astonished that the pioneer farmers could universally neglect this arduous and traditional rite. Perhaps the first macronutrient to be missed from most soils is phosphorus, then nitrogen and finally potassium. By about 1800 the long-worked farms of the settled East began to show loss of yield just as the urban markets started to grow. From the 1830's on the farm journals praised the virtues of composts and manures along with the rotation of crops.

Near the city there was no way to rebuild the soil from the farm's own resources, as a self-sufficient landholder might hope to do. Long Island farmers sold off half of their hay to keep city traffic on the move and would not even allot many acres to pasture livestock, because fresh produce was so much in demand. They were therefore obliged to look outside the farm for other sources of plant nutrients to restore their sandy soils. The sea at hand offered inedible but plentiful alewives, horseshoe crabs and seaweed. These were liberally spread along the furrows, with many other organic expedients as well, all of them bulky, wet, dilute and heavy materials, demanding plenty of labor.

Yet plainly New York City "was a veritable manure factory." A great ecocycle was set turning across the East River before 1840; the very boats and trains that took out the farmer's cash products returned bearing stable manure from dray horses and horsecar teams, ashes from spent city fires, ground bones from the busy slaughterhouses, even odorous street sweepings. The best farmers now were held to be those who bought the most manure. The nutrient cycle was regarded as being closed not farmyard by prudent farmyard but through an intricate maze of technical and market provisions for gathering organic fertilizer of whatever kind in the city and doling it back to the farms.

Yes, New Yorkers tapped the most obvious urban source as well: night soil. Every night the honey buckets were carted off from a multitude of privies and cesspools to the wharves, where in less frugal mood the stuff would be dumped forthwith; a foul mess perpetually surrounded Manhattan. But for two decades a company whose long sheds adjoined the Hackensack River contracted for its boats to receive a portion of that municipal bounty, which it transformed into an unoffending dry powder mainly by mixing in plenty of peat. It was no success; the acclaimed poudrette they touted and sold was too dilute to return much nutrient value to the farm.

Enter guano (the word comes from the Incas, and its agricultural use in Peru is 2,000 years old). By 1845, barely five years after the first barrels of guano were sold in Liverpool, a "perfect mania" was abroad in the U.S. as well for this most powerful of natural manures: guano enough to bring a strong crop response required less weight spread on the land, compared with usual manures, by one or even two orders of magnitude.

The stuff was first collected from three islets washed by the Humboldt Current 10 miles off the arid coast south of Lima. It is the consequence of an almost unique natural setting: upwelling nutrient-rich cold waters blossom with plankton to nourish great schools of fish, colonies of gregarious seabirds batten on the fish and the utterly rainless regime allows the droppings of the replete birds to heap up over the centuries, all soluble ingredients still in place.

Year after year the Chincha Islands yielded up entire sea cliffs of guano (some of the layers were 200 feet deep) to a rotating fleet of merchantmen that waited by the hundreds to load the fertilizer. Ten million tons had been taken to Europe and North America in 30 years. Although they were sought, no such profitable sources of guano were ever found except along that same coast. (Smaller quantities were harvested in the analogous cold current off southwestern Africa.) The speculative frenzy led to bizarre claims and rapacious deeds around many a scattered island. Some 70 guano islands in the Pacific and the Caribbean were claimed for the U.S. before 1880. Many islands were claimed twice; some were quite imaginary. In the end American guano fever left us such enduring tiny dependencies as Swan, Howland and Baker.

Seabird guano was in due course shipped from many places far from Peru, but it was always more or less rain-leached; soluble nitrates were low and the less soluble phosphates were high. The substitute fertilizer was nonetheless commercially useful, although on a smaller scale. Almost odorless, inorganic, concentrated, the phosphatic guano prepared the farmers for more change still.

This surprising chapter of technological history is at once outrageous and subtle. For Richard Wines it becomes a persuasive story of evolutionary preadaptation. Phosphate rock was found in South Carolina in 1867. It was so well received that by 1880 its production was an order of magnitude greater than all imported phosphatic guanos. Yet phosphate rock would never have been so rapidly adopted if the system of the farmers, the market, the transport, the periodicals, the farming practices and all the rest had not evolved step by step toward use of a mineral-like concentrate from afar. For such an evolution to unfold, the old conception of a cycle closed on the farm itself—the land sustained through hard work using familiar barnyard manures and farmyard composts—had to be broken.

Superphosphates, pulverized bone treated with acid to render it soluble and hence more available to plant roots, were proposed by the eminent Justus Liebig, if rather unemphatically. By the 1870's many American entrepreneurs were offering mixed fertilizers of useful grade. They were 2 or 3 percent nitrogen, 6 to 8 percent soluble phosphoric acid and (imported from Germany) about 2 percent potash. Gross adulteration had been ended by law and by the producers. Even the analyses on the bag were credible.

A map derived from the 1880 census shows general use of commercial fertilizer in the counties of the Eastern Seaboard to be heaviest around the major cities, from Boston to Norfolk. In the South its use was widespread as well, more as a short-term solution to the labor and credit problems of postwar agriculture than as the outcome of a long history of soil improvement. Fertilizer began to move inland during the same decade; in the Finger Lakes the wheat farmers of the time all applied it, although it is reported to have had no appeal whatever for farmers in Ohio, Wisconsin and Kansas. Their time was to come. In America 100 years ago "the fertilizer industry had evolved to its modern shape."

Today fertilizers can be said to feed rather more than a billion human beings the world around. After growing steadily for the past 20 years, production has reached a plateau of about 50 pounds per person per year, about half of it nitrogen. Production is also shifting from the developed countries, old seat of the chemical industries, to the developing ones, in particular to those that possess the raw materials.

A rough sketch of the industry at present looks like this. Everywhere atmospheric nitrogen is fixed as ammonia, at high pressure under catalysis, in plants that are thoroughly modernized but still depend on the reactions Fritz Haber developed at Karlsruhe before World War I. Ammonium phosphates supply the majority of the phosphorus; potassium is spread as the chloride. Nitrogen fixation requires plenty of energy. It is done most cheaply with natural gas, which also provides the needed hydrogen. The big natural gas fields are beginning to attract the nitrogen-fixation plants, although right now China and the U.S.S.R. outfix the rest of the world. Phosphates are strip-mined from sedimentary rock outcrops. Their dominant deposits are those in the southeastern U.S. and in Morocco. Sulfuric acid treatment yields phosphoric acid, which is then ammoniated. The potassium salt is mined underground, mostly in Saskatchewan and in the Ukraine; it is freed from sodium chloride by an ingenious froth-flotation process.

The experts assembled by the Soil Science Society have prepared an ample text and reference volume. About half of the book treats of the production, marketing and use of each major fertilizer type. The rest touches a wide spectrum of technical topics.

A general reader is caught by more than one surprise. There is a pipeline net that brings liquid anhydrous ammonia from the Gulf Coast north to the corn and soybean belt. The largest of these lines joins synthesis plants in the Louisiana delta to a fan of terminals that spreads from Indiana to Nebraska. There railroad tank cars and tank trucks load up with a million tons of ammonia each year and deliver their cargo to the fields.

Nowadays sulfate-laden rainwater is acid rain. Yet the contribution rainfall makes to the sulfur needs of crops (the element is absorbed as sulfate ion) has always been important; increased use of prepared fertilizers low in sulfur impurities coupled with high hopes of cleaner air implies that more attention should be paid to this nutrient. Here and there in the U.S. and Canada overt signs of sulfur deficiency in crops are seen. Although natural sulfate sources are complex, the balance sheets clearly show a net loss of sulfur sent off with the crops. Some good seems to be blowing in that ill wind.

Yet modern farming continues the strategy of leaving as little as possible to nature. Those circular fields under center-pivot sprinkler irrigation now gain fertilizer along with water; new measuring and mixing systems make the scheme efficient. New minimal soil-working techniques, the chisel plow and no-till planting, show some waste of fertilizer nitrogen after such shallow application. Still, crop yield does not decrease with no-till farming: other advantages of the technique compensate for the loss of nitrogen.

Even in the well-fertilized agriculture of the U.S. the bacterial symbionts of legumes, crop residues and animal manures return a little more nitrogen to the croplands than fertilizer does. Sewage sludge—the night soil of an industrialized society—supplies little of the need. (If all of it were returned to the land, it could contribute only 1 or 2 percent.) The nutrient cycle will remain open until the genetic engineers can contrive a nitrogen-fixing symbiosis for our major crops.

HYDROGEN PROPERTIES FOR FUSION ENERGY, by P. Clark Souers. University of California Press ($65). Since the time of Alfred Nobel nitrates have been seen as both explosives and essential nutrients for our crops. Contemporary technology too offers such a paradox, but one a millionfold more

Trademarks of some late 19th-century fertilizer companies energetic. If only we human beings can somehow continue to evade the effects of the manmade explosive reaction of deuterium with tritium, we shall within decades be nurtured by that same reaction in its controlled version. The idea has ancient roots. For these past eons all life has fed on the much slower natural proton-proton reaction burning deep within the sun.

By page 2 of this authoritative text the reader finds the graphs from which this promise has been inferred, inconclusively so far. The curves show that the deuterium-tritium thermonuclear reaction is (over a range of three hard-won orders of magnitude in temperature) the easiest of all to achieve. Although there are faint hints of other possible paths, this one alone is well trodden in the foothills. No trail at all has yet been broken up the col to the economic summit.

This authoritative engineering text is a fine interim report from one of the indispensable support parties for the future climb. It is a "materials science description of hydrogen," the simplest and most abundant of substances, in the domain of pressures that are not too high and at temperatures that range downward from room temperature. Most attention is paid to the cryogenic regime where without heavy pressure vessels we can directly engage the condensed forms of hydrogen, the liquid and the solid (not to forget hydrogen snow). The book is a critical compilation of the nature of a substance in bulk; it pays minimal attention to the cryogenic engineering that makes most of the work possible.

The heavier isotopes of hydrogen were first found during the pioneer years that encompassed the discovery of the neutron; their bulk properties have long been reviewed as exemplars for many of the subtle insights of quantum physics. The same phenomena still dominate, but now it is the knowing engineer who needs both to understand and to control in depth.

The author bears the title of Tritium Technology Leader at the Lawrence Livermore Laboratory. He has assembled, digested and reviewed in sprightly tone the state of his extraordinary art, a state achieved over recent decades by an international effort of considerable scale. At first a single professor and his group would carry out the taxing initial studies. Then some national laboratory would undertake to extend and improve the data.

Now a worldwide engineering community is seized with the eventual applications, from the gas fillings of the big plasma fusion reactors to a myriad of tiny frozen pellets made of mixed deuterium and tritium, expected targets for some exotic, powerful but routinely generated beam of the future. It is at least a good omen that among the dozens of collaborators who sent the author opinions and unpublished fresh data is the group at the University of Kharkov in the Soviet Ukraine. Its members are the only workers who produced good isotope comparisons among the three molecular species: the H_2 of the high school chem lab, of balloons and of large-scale chemical industry; its heavy-isotope counterpart D_2, and the mixed hydrogen pair HD.

That molecular gas is the starting point. It is elegantly understood: the only table in the book complete for all six of the molecules built of two clinging hydrogens gives molecular rotational properties with high accuracy. Spectroscopy can work its wonders with small samples, and the motivation to do so is ample because of the remarkable properties of the rotations of the homonuclear, or symmetric, molecules. Those dumbbells (a misnomer for the spherically symmetric lowest state of zero rotation) cannot easily change between odd and even quantum states of rotation. They behave like a mixture of two nonreacting gases unless external forces, in the form of catalysts, act on the weak electrical or magnetic moments to coax equilibrium. A little oxygen, radiation, metal catalysts, "almost anything" extraneous, will work. Other hydrogen molecules in solid or liquid form will work in time, although they are the slowest of all aids.

The isolated molecule in its initial state of rotation at absolute zero will relax to the ground state by internal interactions only after some 10^{12} years. The cold gas should have very little of the rotational spin-1 state, whose kinetic energy clearly raises it above the minimal-energy nonrotating ground state. Cooling from room temperature catches the homonuclear species with their rotations on; they cannot leave odd states for even ones, and so they retain a thermal-energy content when they are cold that dismays incautious cryogenists. Their cold hydrogen has an unexpected internal source of heat, whose release is dependent on wall compositions and other environmental circumstances not easy to foresee. There are now industrial tools available to solve the problem set by this subtle consequence of nuclear identity. You can buy a paramagnetic catalyst (nickel-silica) and promptly get the heating reaction over with by dropping a few pellets of it into your cryogenic tank of rocket fuel, provided the liquid circulation is good.

The liquids are full of interest and the data are far from complete. The tritium-containing species, for example, are known mainly by extrapolations. (That radioactive gas can be bought for about $10,000 a gram; it is usually managed within a set of Chinese boxes "of varying degrees of expense.") While it remains molecular hydrogen, tritium is much less hazardous than other radioisotopes: the body does not much accept it. Once it is oxidized to water, however, tritium becomes 10,000 times more hazardous; its 10-day residence time in the body is said to be shortened by "drinking beer in quantity" (a handy tip?). Liquid hydrogen wets every ordinary material, since its molecules like one another less than they do steel or aluminum or even Teflon. The stuff will "run up" container walls and across the ceiling. The layer may be no thicker than a monolayer, but it will be there.

The solid is extreme in many properties. Its crystal structure can be either body- or face-centered cubic. It can also exist in disordered phases. Solid hydrogen is hard to see because the speed of light in this material of low electron density falls below the speed in a vacuum by only one part in six. That is half the effect found in ice or in the light organic substances that usually head lists of refractive-index values. The mass density is even lower: for the pure proton solid it is a tenth of that of water. "There is a considerable art to growing hydrogen crystals" of quality. They are always grown slowly from the melt; rapid freezing precipitates snow. Strangely enough, a millimeter-thick frozen layer of hydrogen crystallites can look an "opaque brown-black." Cracks do it; light scatters and rescatters at the boundaries, and there is a minimal absorption demanded by the theory.

The presence of a little helium gas is "devastating" to the hope of growing single crystals. A tiny impurity forms a bubble (helium does not freeze) that "twists and turns like a grotesque worm." Tritium, of course, decays steadily to helium, and the bubbles formed are a hassle to the flow of the material, since the helium is very sparingly soluble in the hydrogen liquid. The heat of decay per unit volume is "truly surprising"; tritium concentration is therefore normally measured calorimetrically. Fortunately for experimenters, liquid convection moves heat well in small samples. The radiation effects set free plenty of ions in bulk tritium; the electrical conductivity of liquid deuterium is four orders of magnitude larger than that of liquid proton hydrogen. The assumption is that the added free charge was made by a tiny unexpected tritium impurity, one part per billion! The charge

Permanent pleasures.

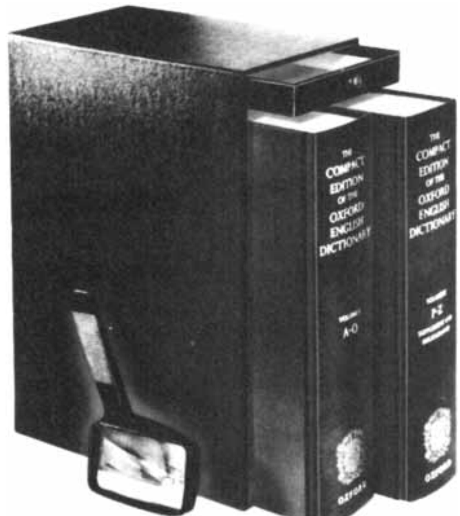
Choose one of these fine works and save up to $305⁵⁰

You simply agree to buy 4 additional books within the next two years.

The Times Atlas of the World Seventh Comprehensive Edition

for **$29.95** (Pub. price $139.95)

"The best atlas now available in English"—*The New York Times.* With 122 double-paged, 8-color maps for clear reading of geographical and political features, this latest edition (1985) also contains a fully updated index-gazetteer with over 200,000 entries and many other unique features. Size: 18" x 12"; 520 pages.



The Compact Edition of The Oxford English Dictionary

for **$27.95** (Pub. price $175)

"The most complete, most scholarly dictionary of the English language"—*The Christian Science Monitor.* Through photoreduction, the original 13-volume set has been reproduced in this two-volume *Compact Edition.* A Bausch & Lomb magnifying glass is included.

The Encyclopedia of Philosophy

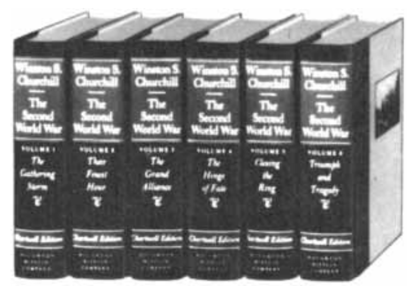
for **$24.95** (Pub. price $225)

Regarded as the most comprehensive encyclopedia of philosophy ever published, this superb set—compiled with the world's foremost philosophers—encompasses vital aspects of ancient, medieval, modern, Eastern and Western thought. The four volumes represent an essential aid for students and a rewarding reference source for home libraries.

The Story of Civilization by Will and Ariel Durant

for **$29.95** (Pub. prices total $335.45)

For almost half a century Will and Ariel Durant traced the continuity of world history—the religions and philosophies, the political and economic tides, the arts and sciences, the customs and conquests—to show the foundations of society today. A Book-of-the-Month Club exclusive for almost 50 years, the Durants' illustrated masterwork is history come alive.



The Second World War by Winston S. Churchill

for **$27.50** (Pub. price $295)

New Chartwell Edition of this six-volume Nobel Prize-winning masterwork, in genuine leather quarter-bindings, with a foreword by William L. Shirer. Churchill's Olympian wisdom, passion and wit—and his unique role as a wartime leader—make his account of World War II unforgettable reading. Frontispiece photographs, 3-color battle-map endpapers, 195 additional maps, charts and diagrams.

Facts About Membership. As a member you will receive the *Book-of-the-Month Club News®* 15 times a year (about every 3½ weeks). Every issue reviews a Selection and more than 125 other books, which are carefully chosen by our editors. If you want the Selection, do nothing. It will be shipped to you automatically. If you want one or more other books—or no book at all—indicate your decision on the Reply Form and return it by the specified date. *Return Privilege*: If the *News* is delayed and you receive the Selection without having had 10 days to notify us, you may return it for credit at our expense. *Cancellations*: Membership may be discontinued, either by you or by the Club, at any time after you have bought 4 additional books. Join today. With savings and choices like these, Book-of-the-Month Club is where book lovers belong.

BOOK-OF-THE-MONTH CLUB®

Book-of-the-Month Club, Inc., P.O. Box 8803, Camp Hill, PA 17011-8803 **A170-1**

Check one box only.

- ☐ 946. Times Atlas of the World **$29.95**
- ☐ 905. Compact OED **$27.95**
- ☐ 917. Ency. of Philosophy **$24.95**
- ☐ 913. The Story of Civ. **$29.95**
- ☐ 959. The Second World War **$27.50**

Please enroll me as a member of Book-of-the-Month Club and send me the item I have checked at left, billing me for the appropriate amount, plus shipping and handling charges. I agree to buy four more books during the next two years. A shipping and handling charge is added to each shipment.

Name____________________ 7-64
(Please print plainly)

Address__________________ Apt.______

City________________________

State________________ Zip______

Prices shown are U.S. prices. Outside the U.S. pr ces are generally higher.
All orders are subject to approval.

density in the pure liquid hydrogen is thought to be due mainly to cosmic radiation; that fluid is about as good an insulator as can be found.

There is much more, from curiosities such as mu-meson chain catalysis of fusion to the use of bread-and-butter instrumentation of the latest kind, such as nuclear-magnetic-resonance systems. Clark Souers has worked well and hard without losing his sense of humor. He has embraced the SI metric system and professes to have "no mercy to show on this subject." Actually he has tempered his units of time and length when needed.

There is about this volume a pleasant hint of a mature technology to come; indeed, the wide use of liquid hydrogen (or perhaps solid hydrides) as combustion fuel with air—in which the normal isotope mix will rule unchallenged—is quite likely to accompany or even precede the fusion uses. The condensed matter of the giant planets begins to look more and more like a terrestrial commodity of the future, a real strangeness becoming a commonplace with passing time.

On Growth and Form: Fractal and Non-Fractal Patterns in Physics, edited by H. Eugene Stanley and Nicole Ostrowsky. Martinus Nijhoff Publishers, distributed in the U.S. by Kluwer Academic Publishers, 190 Old Derby Street, Hingham, Mass. 02043 ($44.50; paperbound, $14.95). Intricately branching, clotted, spongy or deckled patterns abound in ordinary matter, from colloids to copperplating to cell colonies. It is this classical macroscopic but complex structural category that has attracted the attention of the physicists from many lands who, with some eager students, made up the informal summer session reported in this up-to-date bargain of a book. Half of the volume consists of tutorial reviews and is labeled "the course"; half, "the seminars," consists of specialized accounts.

The pleasure of this topic for the nonspecialist is the simplicity of the problems that are posed, a delight it shares with the theory of numbers. Just as in that theory, the sophistication of the mathematical search for answers runs quickly beyond the grasp of a general reader. Here, however, there are simple and striking experiments to ponder, and plenty of intriguing computer simulations are graphically presented.

Number theory celebrates Euclid and Diophantus; this is a specimen of a young literature. A paper of 1960 seems in the context to be almost of Ionian antiquity. A couple of such pioneer articles are cited repeatedly; the way was opened in the mid-1970's and most of the main results are dated in the 1980's. The dozens of contributors to the volume represent the bibliography of the field well. This is a type specimen of a subdiscipline witnessed in its happy days of growth.

Begin with one experiment of 1984, reported by its coauthor in a brief review chapter. A deposit of solid copper was electroplated out of a blue-green gel. Records of the current as time went by allowed calculation of the mass of the deposit as the radius grew. A power law fits the data well over a range of four or five powers of 10 in mass. But the measured mass does not increase with the third power of the radius, as it would for any self-respecting pure metal sphere; instead it scales according to a smaller power, 2.4. Under these carefully arranged conditions the copper is not a solid crystalline deposit but a spongy compaction of feathery dendrites. The effective unit of deposit was a tiny metal grain; the process was limited by the slow diffusion of the copper ions inward to the cathode, so that either "a grain was seeded and grew or it did not." That submillimeter ball of copper is a fractal: a structure similar to itself at a wide range of scales, whose randomized geometry can be understood as a kind of fractional dimensionality.

The theorists (and their computers) were ready for it. Here is their model. Start with a single seed particle. Then allow another to diffuse toward it in a random walk, the ordinary Brownian motion in three dimensions. Apply the simplest rule for aggregation: The particle sticks if and where it hits a previous particle and stays there. Continue the growth—the computer algorithms to do this cheaply and well are by no means simple—for 50,000 or 100,000 model particles. Repeat, say 64 times, and plot the average results of mass versus radius. The power law fits, and the best-fit exponent is estimated, to within a few percent, at about 2.5.

One of the classic studies of the 1960's was a related model in which the infalling grain moves from the boundary toward a single seed, not in the random walk of diffusion but along a randomly aimed straight line. If it strikes any predecessor, it sticks where it hits. That mode, which is called ballistic growth, generates patterns in the plane that after a few thousand trials look like faintly porous, deckle-edged and roughly circular spots. Simulation shows that if growth goes on for long enough—say a quarter-million tries—the laciness dwindles steadily. Then the area covered grows with the square of the radius, and the edge becomes progressively better defined. In this case the fractal dimension has approached the ordinary Euclidean value of 2.

Diffusion-limited aggregation (DLA in the jargon) is only one among the successes in this new province of statistical mechanics. The copperplating example suffers from its electrical elegance; we do not see much of that dendritic coppery geometry. Try a more photogenic case, the slow injection of blue-dyed water into a thin, transparent layer of a gummy fluid, all held between two narrowly spaced flat plastic plates. The water fingers out from its point of entry into radial patterns, randomly branching and rebranching, down to little twigs.

Once again the optically measured area of the infiltrating blue shows a power-law growth related to the radius of the branch tips: a fractal again, now with the exponent 1.7. The spatial environment of the layer in this example is two-dimensional, of course, whereas the copperplating was in 3D; the DLA model for this case predicts 1.68 for the fractal exponent, again to an accuracy of a few percent. This model too is studied chiefly by simulation experiment, and so the theorists concede statistical error just as the experimenters do. (In this young community the theorist and the experimenter are still partners—sometimes one and the same physicist.)

Physics is always richer than even this dazzling geometry; there is much to be said about the material properties of fractal structures, in the real world and in the model one. Suppose the patterns are made of linked resistors; what is the resistance from end to end? How likely to appear is a closed connecting path between borders when clusters grow from many seeds within a large region? How strong is a fractal structure under mechanical stress? What is its spectrum of acoustic vibrations? In his tutorial summary editor Stanley, a physicist at Boston University, lists and discusses no fewer than 10 characteristic dimensions associated with physical properties of model fractals. At this point there enters the most advanced part of the subject: its relation to the deep universal theories of phase change, not really open to the general reader.

The forked-spark patterns of dielectric breakdown are also encompassed by the theory. The electric field obeys the same averaging requirements that enter the random walk, and so the similarity can be justified. A certain liberty must be taken with the fully defined DLA model; we still do not know how to relate the growth rate at the tip of a discharge to the electric field there.

The relation is parametrized with a power law; with that loophole for adjustment the beautiful branching discharge forms (the so-called Lichtenberg figures) yield the expected fractal dimension.

In the summer of 1986 (here this column digresses from the book) DLA theory—and the editor's group—made a good start on reproducing the most familiar self-similar material of all, the wonderful snowflake. Its sixfold nature is easily built in, of course, by letting the particles diffuse on a triangular lattice. A noisy quality is added by a lottery that decides which of many equally probable sites should actually grow at each step. This amounts to adding random elements in the dielectric-breakdown model. The amount of noise can be controlled by stacking the odds. Then anisotropy is added by favoring certain axis directions in the lattice. The adjusted results show convincing model drawings of snowflakes, although the correlations between side branches of different arms, so plain in many of the snowflakes of real winter, are still missing. Perhaps they are the result of some defect sources of crystal twinning that propagate from the center of the flake.

The quick and inexpensive publication of this report is praiseworthy; little was lost in appearance and in a few editorial glitches. But stealing your title from a great classic of science is not a good idea, even under cover of a glowing tribute. An allusion to chance might have allowed the editors an evocative title all their own. What they sing is certainly one new canto in the epic of the endless conflict between the random and the necessary.

PHOTOATLAS OF INCLUSIONS IN GEMSTONES, by Eduard J. Gübelin and John I. Koivula, with the cooperation of Henry O. A. Meyer, Edwin Roedder and H. A. Stalder. ABC Edition, Zurich, distributed in the U.S. by the Gemological Institute of America, 1660 Stewart Street, Santa Monica, Calif. 90404 ($175). Share a dream: on a long bench are arrayed a collection of hundreds of different faceted gemstones of incredible size, beyond the treasure of princes, bigger than your doubled fists, each lustrous, gleaming, ruby red, emerald green, topaz blue or diamond white. We look deep into each of them, through a polished window about the size of a visiting card, to view in those entrancing depths a scene of strange and varied inclusions in the gemmy host.

Reality is different only in scale. The gems are fine, costly examples of the size we expect, about that of a fingernail. The scenes are millimeter-size; our vision is given us through the low-powered microscope, magnified a few tens of times, in color photography of beauty, fantasy and meticulous detail. The optical techniques are powerful and varied, using interference, phase contrast, tiny optical-fiber spotlights, dark field and more. An account of techniques is followed by a general explanation and by guest chapters on the genesis of diamond and quartz and of fluid inclusions. Then the atlas displays inclusions ordered by their mineral nature, from amphibole to zircon; next a collection ordered by the host gem, and finally a special array emphasizing synthetic gems.

The author-photographers are gemmologists. The senior author brings 40 years of study to the work, begun in his father's gem laboratory in Lucerne. His younger colleague is a leading U.S. practitioner. The most straightforward purpose of the volume is as an aid to the authentication of gems, since it is not the host but the inevitable inclusions that best distinguish the natural from the synthetic—so far. It may well be that new synthetics will "bring about an end to the deeply rooted feeling that, in the world of gemstones, flawless is best."

There is more than appraisal here: there is a powerful diagnostics of process and the contagious appeal of sheer aesthetic delight. Three cool blue photographs display a negative crystal, a faceted void, in a Sri Lanka sapphire. It is really no void at all; there dwells within it a small mobile bubble. The cavity is filled with a fluid and a little of its vapor. That fluid becomes plain as the bubble vanishes, once the gem is heated to the temperature at which liquid carbon dioxide and the gas lose their mutual interface and merge.

An entire company of little negative crystals parade within a blue topaz from Zimbabwe, each with its two-phase filling, a diagnostic as tight as a hallmark. There are voids containing liquid methane, voids containing both carbon dioxide and water and voids containing fluorescing oil. Entire scaffoldings of crisscross needles, minuscule submarine gardens of flower-like growths, cracks and rosettes and much more are all clues to the nature and order of the processes of deposition. Often they are clues we cannot yet quite puzzle out. Starred gems, cat's-eyes, moonstones and the rest of that glittering lot are examined.

Fifty photographs record organic inclusions in the natural polymer amber: tree gum that has had 30 million years in which to form cross-links. The star among the amber inclusions is surely our unusual vertebrate cousin, a little lizard three inches long, beautifully embalmed. The skeleton has already been thoroughly examined by X ray, through the amber and the residue of the flesh. The studies were made by the paleontologists of the Natural History Museum in Basel. The DNA analysts may well be next.

Finally, the superb synthetic opals and emeralds and lapis of more than one company disclose for us beautiful and telltale textures of their genesis. The newest kind of synthetic hydrothermal emeralds show the feathery helixes of fissure healing, once held to be valuable features unique to natural emeralds but in fact the consequences of process. Like causes, said Newton, like effects.

A fossil of a lizard in amber some 30 million years old


TOYOTA CRESSIDA
CRESSIDA ENTERTAINMENT
A Technics** AM stereo/FM/MPX tuner with cassette and acoustic flavor/tone equalizer is standard.
CRESSIDA. THE MOST TROUBLE-FREE LUXURY CAR YOU CAN OWN.
CRESSIDA LUXURY
It's lavish, with automatic climate control, power windows, power door locks, remote power outside mirrors and rich velour or optional leather interior.
A nationwide survey rated Toyota Cressida the most trouble-free new luxury car you could own.* And it's every bit a luxury car. Its interior is elegantly appointed with the finest materials. The exterior is beautifully finished, aerodynamically superb. To delight the driver in you, Cressida has impeccable road manners and the power of a sophisticated 6-cylinder, twin cam engine. Inside and out, Cressida compares with cars costing thousands more.
TOYOTA
WHO COULD ASK FOR ANYTHING MORE!
TOYOTA
Get More From Life... Buckle Up!
*Based on problems encountered in first 90 days of ownership— J.D. Power & Associates 1985 New Car Initial Quality Survey.
**Technics is a trademark of the Matsushita Electric Industrial Co., Ltd.
© 1986 Toyota Motor Sales, U.S.A., Inc.

Established 1845 SCIENTIFIC AMERICAN January 1987 Volume 256 Number 1

The Yields of Soviet Strategic Weapons

They have been generally overestimated. Recalculated yield estimates of Soviet weapon tests indicate that U.S.S.R. compliance with treaty limits has restrained its development of strategic nuclear warheads

by Lynn R. Sykes and Dan M. Davis

A common perception, extending even into the higher levels of the Federal Government, holds that the potential destructive power contained in the Soviet Union's strategic-warhead inventory greatly exceeds that of the U.S. Opponents of arms control argue that the Soviet Union has managed to attain this advantage by exploiting loopholes in arms-control treaties or perhaps through outright noncompliance, whereas the U.S. has constrained itself by holding to the letter if not the spirit of the treaties. Indeed, the Reagan Administration alleges that the U.S.S.R. has more than likely violated the Threshold Test Ban Treaty (TTBT), which sought to restrain the destructive potential of nuclear weapons by limiting the explosive energy, or yield, of underground nuclear-weapon tests.

We have recalculated the yields of all underground Soviet nuclear-weapon tests by a revised method that is similar to the one reportedly adopted by the Government only last year but that had been advocated by many seismologists over the past decade. Our results indicate that the yields of Soviet underground nuclear-weapon tests had been consistently overestimated—by analysts both inside and outside the Government. The reduced yields we arrive at show that, contrary to the Government's allegations, the record of Soviet underground tests has been in accordance with the TTBT. The reduced test-yield estimates imply that yields for deployed Soviet warheads have also been overestimated and that, in fact, a rough overall parity (in terms of total destructive power) has prevailed for the past decade in the strategic balance between the two superpowers. Furthermore, we found that Soviet nuclear-warhead designs are significantly less efficient than U.S. designs, producing a much smaller explosive yield for the same weight.

Hence the limitations imposed by the TTBT, along with other arms-control treaties drawn up under the Strategic Arms Limitation Talks (SALT), have ultimately hampered the Soviet Union more than the U.S. in the development and deployment of high-yield warheads for strategic weapons. These results compel reconsideration of the role that arms-control agreements have played and may still play in shaping the strategic balance between the two superpowers.

The Limited Test Ban Treaty, signed by the U.S. and the U.S.S.R. in 1963, prohibits nuclear explosions in the atmosphere, under water or in outer space. Although the treaty did reduce the radioactive pollution of the biosphere, it did not halt the nuclear-arms race; it merely drove the detonation of nuclear devices underground. Because underground explosions generate vibrations that propagate as seismic waves through the crust and mantle of the earth (in which case they are known as body waves) and along the earth's surface (in which case they are known as surface waves), seismology provides the main tools for detecting and characterizing nuclear explosions. Seismologists can distinguish the seismic signals of an underground nuclear explosion from the signals of earthquakes; they can also make accurate estimates of the explosion's yield.

Programs for developing and testing nuclear warheads and their delivery systems are carried out in great secrecy in the Soviet Union. The U.S.S.R. has released yield values and other pertinent information only in connection with a few underground "peaceful nuclear explosions," which presumably were not related to weapon development. Yet because the seismic waves radiated by a nuclear test travel great distances, seismographs far removed from the test site can record the telltale signal. Such data, as well as the methodology for yield estimation we employ, are all in the public domain.

Seismologists characterize the size of a seismic event—whether it is an explosion or an earthquake—by means of magnitudes. The magnitude of a seismic event is the logarithm of the amplitude of the event's seismic waves normalized, or adjusted, for the distance between the source and the observing station. In our calculations we used two magnitude scales, one based on certain body waves, called *P* waves, and the other based on surface waves. Of these, *P* waves have most often been relied on for yield estimates since they have been extensively recorded even for explosions of low yield.

A magnitude value by itself cannot

reveal the explosive yield of a nuclear test. In order to have meaning a magnitude must first be calibrated against the magnitudes measured for underground explosions of known yields. Yields from underground nuclear explosions in six regions of the U.S. have been published; the yields for French underground nuclear tests in the 1960's in southern Algeria are also publicly available. If the body-wave and surface-wave magnitudes recorded for these tests are plotted as a function of the logarithm of their respective yields, a straight line can be drawn through the points [*see illustrations on page 32*].

Once such a line has been constructed, it would appear that estimating the yield of a Soviet weapon test is relatively straightforward: the magnitude of the test's seismic waves is measured and, by noting where on the line relating magnitude to yield the measured magnitude falls, the yield can be determined. In practice, however, complications arise.

Chief among them is the varying effect the local geology of the test sites can have on seismic waves. Another complication is the mingling of explosion-generated seismic waves with the waves produced by tectonic release: the relieving of natural geologic stresses in the vicinity of underground explosions at or soon after detonation. Finally, the temperatures in the lower-crust and upper-mantle regions below the test site must be considered, since they affect *P*-wave amplitude.

The U.S.S.R. has conducted most of its nuclear-weapon tests at three main sites: one is in eastern Kazakhstan, near the city of Semipalatinsk, and the other two are on the arctic island of Novaya Zemlya. The U.S. has carried out some underground tests on Amchitka Island, a remote island of the Aleutian chain, but most U.S. tests have taken place at the more accessible Nevada Test Site (NTS).

It turns out that if surface-wave magnitudes are used in plotting lines relating magnitude to the logarithm of yield, the line does not shift appreciably according to the type of rock in which the explosion took place—as long as it is not dry, porous alluvium. Dry alluvium is essentially deposits of sand and gravel, which contain large amounts of air space. Both surface- and body-wave magnitudes for explosions in dry alluvium are smaller than those for similar explosions in hard rock or below the water table since much of the explosive energy is dissipated in closing the air spaces.

Such alluvial deposits are not thick enough at the Soviet test sites to be a significant factor for contained underground explosions of yields greater than a few kilotons. (One kiloton is the explosive energy released by the detonation of 1,000 tons of TNT.) Hence a nearly universally applicable calibration curve based on the surface-wave magnitude can be used to calculate yields of Soviet explosions.

The seismic signature of a release of tectonic stress, which had tended to muddle seismic signals of explosions in the past, can now generally be removed from the surface-wave signal of a nuclear test by means of modern analytical techniques. In any event, tectonic release is usually not a major complicating effect for explosions at the Nevada or Amchitka sites in the U.S. or the Novaya Zemlya site in the U.S.S.R. In these cases either the natural tectonic stresses are low or the tectonic-release signal is such that its effect on surface-wave magnitudes is negligible when the magnitudes are averaged from measurements taken at several seismograph stations around the world. In the case of eastern Kazakhstan the orientation of the natural stresses is such that their tectonic-release signal cannot be averaged out. We have corrected for the residual effect in calculating yields from surface waves for explosions at that site.

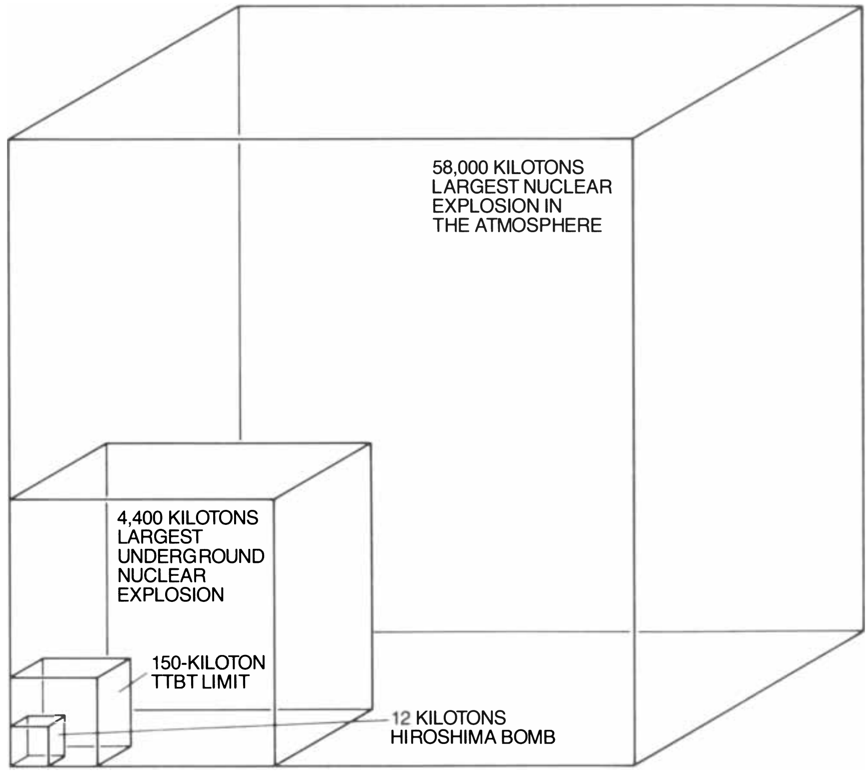
Explosions in hard rock such as granite typically generate *P* waves whose amplitudes are between one and a half and two times larger than those of *P* waves from explosions in softer rocks such as tuff (a volcanic rock) and shale. Hence the type of rock in which a nuclear test takes place is an important factor in estimating yields from body-wave magnitudes. Available geologic data indicate that a hard-rock calibration curve is appropriate for most explosions in eastern Kazakhstan. The assumption can best be checked by comparing the yields estimated from surface-wave magnitudes with those obtained from body-wave magnitudes for 150-kiloton explosions at the Kazakhstan site. The yields calculated on the basis of the two magnitudes turn out to be nearly identical.

For the Novaya Zemlya sites, where the largest Soviet underground explosions have taken place, yields can be calculated by applying a calibration line constructed from the body-wave magnitudes of U.S. tests that had yields greater than about 200 kilotons. Such large explosions can be contained only by detonating the nuclear devices between 500 and 2,000 meters

MAIN NUCLEAR-TEST SITES of the Soviet Union are on the island of Novaya Zemlya (at two separate locations) and in eastern Kazakhstan, near the city of Semipalatinsk. Although the largest Soviet nuclear explosions took place at the Novaya Zemlya sites, the majority of tests have been conducted at the Kazakhstan site. Nuclear devices have also been set off elsewhere in the U.S.S.R. for reasons unrelated to weapon development.

underground. At such depths the rocks are more likely to be water-saturated and both stronger and less porous than those close to the surface. These factors contribute to the result that *P* waves from deep, high-yield explosions are similar, regardless of whether they are generated by a test in Nevada or on Novaya Zemlya—as long as they are corrected for attenuation.

Body waves have a major disadvantage for estimating yields in that *P* waves, unlike surface waves, can be absorbed or attenuated, depending heavily on the temperature from 25 to 150 kilometers below the test site. The NTS, from which most of the published yield information comes, is in an area that experienced a major episode of deformation and heating within the past several million years. *P* waves from an underground explosion in Nevada will therefore be more attenuated and will exhibit smaller body-wave magnitudes than those generated by an identical explosion on Amchitka, in central Asia or on Novaya Zemlya, all of which are areas that have not undergone geologically recent heating. If *P*-wave and yield data from Nevada are used to calculate yields of Soviet tests (or for that matter yields of tests in other regions of low attenuation like Amchitka) without correction for that important effect, yields will be overestimated by a factor of from two to four. Until last year Government analysts had not appropriately accounted for this effect in their yield-estimate calculations.



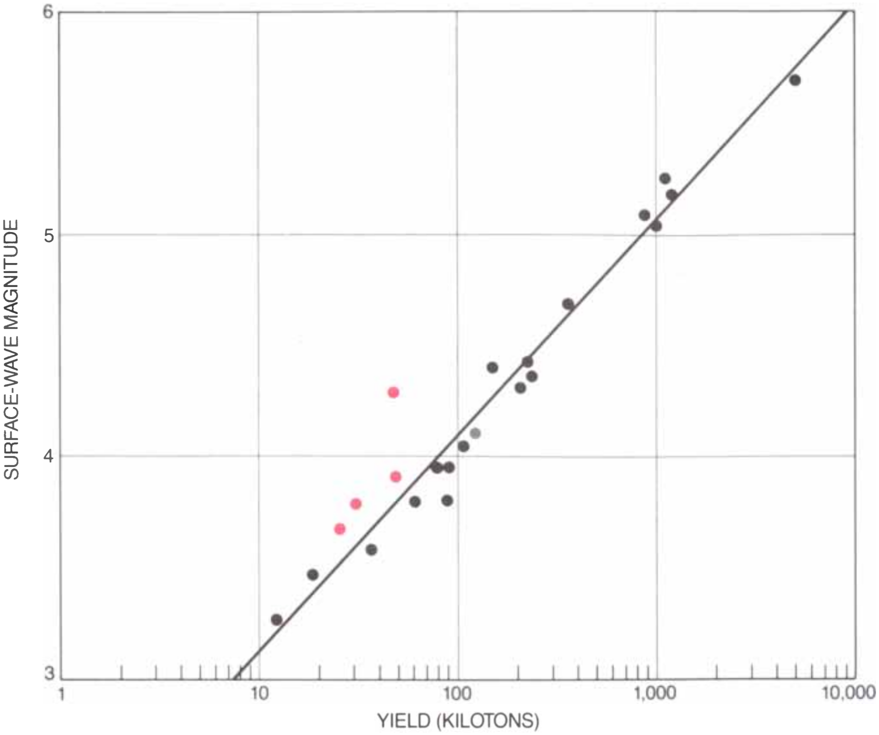
RELATIVE YIELD, or explosive energy, of various nuclear devices is depicted by means of boxes whose volume is proportional to the device's yield. The boxes represent the yield of the most powerful devices ever detonated in the atmosphere (done by the U.S.S.R. in 1961) and underground (done by the U.S. in 1971), the 150-kiloton testing limit set by the Threshold Test Ban Treaty (TTBT) and the yield of the fission bomb that destroyed Hiroshima in 1945. One kiloton is equal to the energy released from 1,000 tons of TNT.

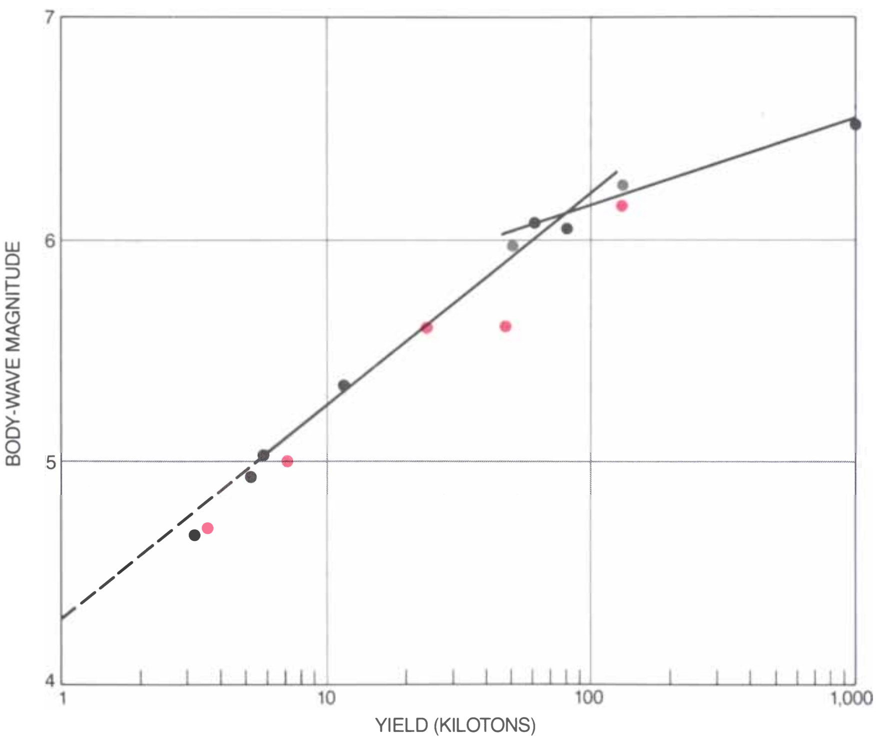
The first question we considered after we had computed the yields for all known Soviet underground tests from appropriately adjusted magnitude-calibration curves was whether, as the U.S. Government alleges, the U.S.S.R. had violated the TTBT. The treaty (which has not yet been officially ratified by the U.S.) sets a limit of 150 kilotons on the maximum yield of an underground weapon test. It also includes a protocol that assists verification by detailing technical data to be exchanged and by restricting testing to certain designated test sites.

Our results show that no explosion since March 31, 1976, when the TTBT went into force, has had a yield clearly greater than 150 kilotons, although seven explosions (all of them taking place at the Kazakhstan site) were very close to the limit. Consequently there appears to be no scientific basis for the claim that the U.S.S.R. probably has violated the TTBT. A few of the calculated yields for those seven explosions are several percent greater than 150 kilotons, whereas others are somewhat smaller. In the three cases for which both body- and surface-wave magnitudes were measured, a yield somewhat larger than 150 kilotons obtained from one magnitude was balanced by a yield somewhat smaller than 150 kilotons obtained from the other magnitude. This kind of dispersion is what is expected from measurement errors, and the data are therefore consistent with the assertion that the seven largest tests all have an identical yield very close to 150 kilotons.

On the other hand, we cannot rule out the possibility that one or two Soviet tests since 1976 may have exceeded the limit by several percent. In any event, the U.S. has also conducted a number of tests since 1976 with yields near the limit stipulated by the TTBT. Indeed, yields calculated from seismic waves for three U.S. tests are somewhat larger than 150 kilotons. Once again, this is probably a consequence of uncertainties in the seismic measurements, although minor violations on the part of the U.S. cannot be ruled out by the seismic data. In this light the congressional testimony of Donald M. Kerr, former director of the Los Alamos National Laboratory, is particularly trenchant. He stated that a nuclear-weapon test would have to be twice as large as 150 kilotons for the test to be militarily significant.

When recorded underground explosions are grouped according to estimated yield [*see illustration on page 33*] certain interesting testing patterns can be discerned. After the Limited Test Ban Treaty went into effect in 1963 and before the TTBT took effect in 1976 there were certain large yields that appeared to be favored by the Soviets in their underground tests: 500, 1,000 and 2,000 kilotons. (The choice of round numbers in warhead design is not surprising for a country such as the U.S.S.R. that operates on the metric system.) All these large-yield tests were conducted on Novaya Zemlya. During the same period a prominent gap between yields of 150 kilotons and yields of slightly less than 400 kilotons is evident. It was also Soviet practice during this period to conduct numerous tests of about 18-kiloton yield. The devices tested in these cases were more than likely the fission triggers that serve to ignite the much more powerful fusion reactions in Soviet thermonuclear weapons. The explosions set off in the 25-to-75-kiloton range were likely to have been full-





CALIBRATION LINES that relate the magnitude of certain types of seismic waves to the yield of an underground nuclear test can be drawn from the points representing underground explosions of known yield that took place in the U.S. (*black dots*), the U.S.S.R. (*colored dots*) and Algeria (*gray dots*). Surface waves, which travel along the upper crust of the earth, provide a nearly universally applicable calibration line (*top*) over the entire range of yields because they are not significantly affected by the geology of the test site. Body-wave magnitudes for explosions in hard rock (*bottom*) are derived from the amplitude of *P* waves: compressional seismic waves that propagate through the mantle and crust of the earth. *P* waves radiated by an underground explosion are susceptible to attenuation or absorption, depending on the temperature of the rocks from 25 to 150 kilometers below the explosion point. The points plotted here have been corrected for this and other distorting effects so that the calibration lines can be used to estimate yields of explosions at the eastern Kazakhstan testing area. Most of the seismic measurements in the diagrams are from P. D. Marshall of the British Ministry of Defence and Donald L. Springer and Howard C. Rodean of the Lawrence Livermore National Laboratory.

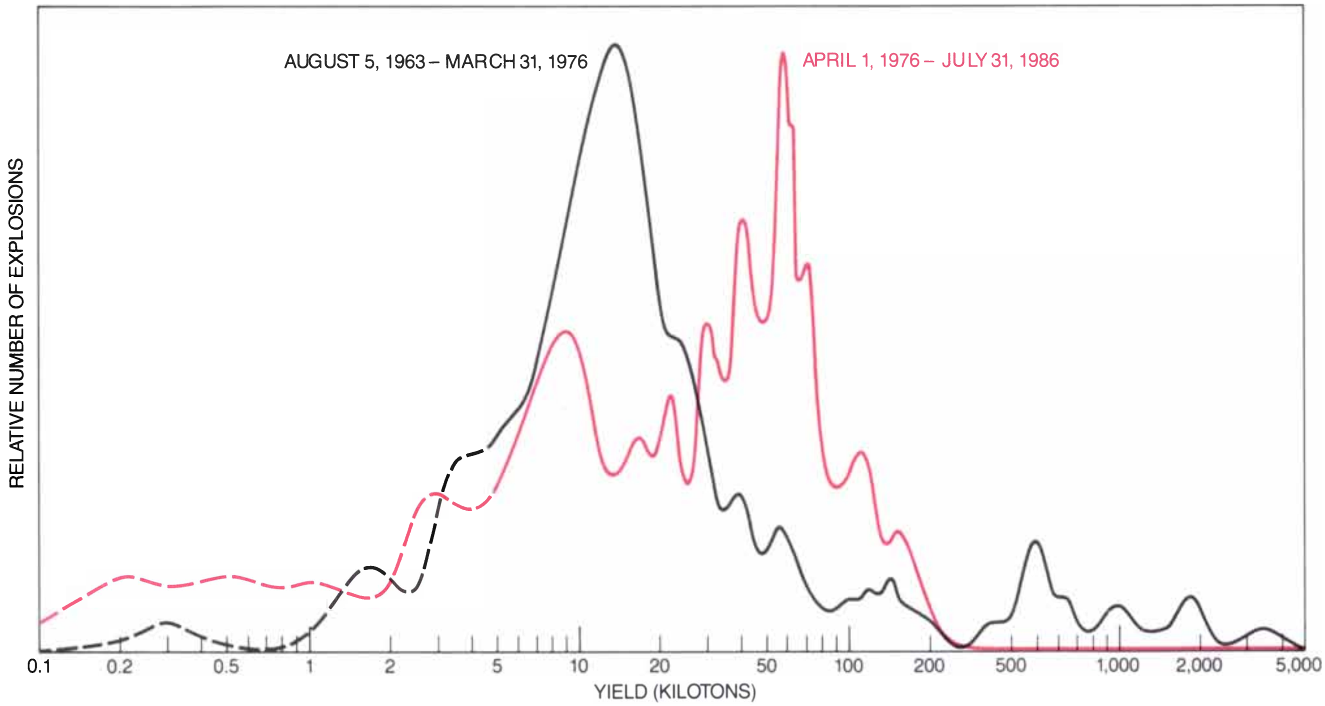
yield tests of smaller warheads, as well as a few partial-yield tests of larger weapons.

Since the TTBT went into force it has been Soviet practice to carry out many more tests in the range of 25 to 75 kilotons. These tests are probably a combination of full- and partial-yield tests. Although 150 kilotons is larger than the yields of many nuclear weapons, including most of those carried on submarines, it is less than the yields of many weapons that were tested underground before 1976. Because of the TTBT, neither country has been able to test these large-yield warheads at full yield for more than 11 years.

Large-yield warheads are likely to be carried by strategic delivery vehicles, which can reach targets in the U.S. from their normal stations. Such vehicles include intercontinental ballistic missiles (ICBM's), submarine-launched ballistic missiles (SLBM's) and long-range bombers equipped with bombs, cruise missiles or short-range air-to-surface missiles. A missile warhead is delivered to a target in a reentry vehicle (RV), which shields it from the extreme heat generated on entry into the atmosphere at high speed. Some missiles (including all those deployed before the late 1960's) carry only one RV. A few carry two or three multiple RV's (MRV's) that disperse before detonating and thereby destroy a larger area. Other strategic missiles carry up to 14 multiple independently targetable RV's (MIRV's). Each MIRV can be directed to a different target by releasing it after slightly altering the ballistic trajectory of the "bus" on which the MIRV's are carried into space. The throw weight of a missile is its effective payload and includes the weight of warheads, decoys and the bus.

In order to determine what the likely explosive yield of a given missile's warhead is, we must have some idea of how large a warhead can be carried by a missile of a given throw weight. An approximately linear relation (much like the one by which we determined weapon-test yields from the seismic-wave magnitudes) can be drawn between the yield of a given warhead and the throw weight of the missile carrying the warhead—with one important difference. The ratio of the yield of a missile's warhead to its throw weight has tended to improve in the course of time as the result of advances in nuclear-weapon and RV design.

On the average, U.S. yields for a given throw weight have gone up by about 4 percent per year for single-RV missiles and by about 5 percent per year for MIRVed missiles. When an ad-



SOVIET TESTING PRACTICES changed markedly after the TTBT went into force on March 31, 1976. High-yield underground nuclear tests before then (*black*) clustered around certain specific yields: 500, 1,000 and 2,000 kilotons. The prominent peak around 18 kilotons is probably the result of frequent testing of the fission triggers for the much more powerful thermonuclear warheads. Since the TTBT went into force (*color*) the Soviet Union, contrary to U.S. Government assertions, has not detonated any nuclear explosives with a yield clearly greater than 150 kilotons. Although all yield estimates plotted here are subject to error, the yields for nuclear tests smaller than about five kilotons (*broken curves*) are particularly susceptible to uncertainties in the test-site geology.

justment based on these annual rates is made to compensate for the year of a missile's deployment, a simple linear relation matches the publicly available data for U.S. missiles remarkably well [*see illustration on page 36*].

If the logarithm of warhead yield is plotted as a function of the logarithm of missile throw weight per RV, the points should fall along a line of slope equal to 1 (that is, a line that forms an angle of 45 degrees with the horizontal axis) if the weights of all components of a warhead scale up in proportion to the warhead's yield. In fact, however, many components of a missile do not have to scale up in proportion to yield. For example, the fission triggers of thermonuclear warheads are probably close to the same size regardless of the warhead's total yield. That would give the yield-to-throw-weight line a larger slope, implying that a warhead of a given yield weighs less than two warheads of half the yield. Actually the warhead yields and throw weights of U.S. missiles deployed at about the same time describe lines with slopes somewhat greater than 1.

We now have the tools to tie a deployed Soviet missile warhead to what must have been its underground test. Given the published estimates of the missile's throw weight and the date the missile was first deployed, the diagrams of warhead yield versus throw weight can be consulted to estimate the likely warhead yield. Any tests in this yield range that took place at about the time of the first flight test of the missile are therefore likely to have been tests of the missile's warhead. Since we seek to determine the restraining effect of the TTBT on the design and deployment of Soviet strategic weapons (in contrast to tactical or short-range nuclear weapons), we concentrated on tests with yields equal to or greater than 100 kilotons: they are more likely to be full-yield tests of warheads carried by strategic delivery vehicles.

There is still a catch, however. How can one plot a yield-to-throw-weight diagram for Soviet missiles when one does not know the yields of Soviet missile warheads to begin with? (Although public estimates do exist, we wanted to avoid having to rely on data that were likely to be inaccurate.) The way out is to assume that the largest Soviet tests between 1963 and 1976 were done at full yield. The fact that no arbitrary limitation on yield had yet been set by the TTBT argues that our assumption is a valid one. The repeated detonation of weapons of only a few yields between 1963 and 1976 also gives credence to our assumption. The U.S.S.R. cannot detonate devices of yields larger than about 150 kilotons at its Kazakhstan test site, since seismic waves from explosions of greater yield can cause damage to aboveground structures there. The Soviet Union consequently was forced to carry out full-scale tests of high-yield strategic weapons on Novaya Zemlya, where weather and remoteness allowed only one or two tests per year.

A general correlation between the rate of nuclear testing and the rate of deployment of new strategic systems also helped to get us started. A look at the historical record will show that a spate of large tests always precedes the deployment of several missile systems. Hence when the Soviets carried out numerous nuclear tests in the atmosphere in 1961 and 1962, they deployed new missiles a few years later. Conversely, when there is a lull in nuclear testing, as was the case after the Limited Test Ban Treaty went into effect in 1963, few new strategic systems are deployed in the next several years. Similarly, because the U.S.S.R. did not test any devices with yields between 150 and just under 400 kilotons between 1963 and 1976, it is unlikely that the U.S.S.R. developed warheads of that yield from 1963 to, at the earliest, 1976. We took advantage of that conspicuous gap in tested yields and deployed hardware and began our initial speculation there, knowing that warhead yields of less than 400 kilotons could automatically be excluded.

In the mid-1960's the Soviets had just started underground testing and most of their tested yields were quite small. The lone 400-kiloton test of 1966 is, by a factor of almost three, the largest Soviet underground test of the 1960's. The device may also have been tested in the atmosphere before 1963 or it may have been developed as a warhead for a missile whose design was not very successful and so was not deployed in large numbers. Our assignment of that weapon to the SS-13 ICBM, which was first flight-tested in late 1965 and first deployed three years later, fits the latter category. The development timetable for the SS-13 is somewhat anomalous, since the missile was tested (but not deployed) before its warhead. More than likely the Soviets would have carried out the SS-13 warhead test a year or two earlier, in keeping with their typical schedule, had it not been for the limitations imposed by the new underground-testing technology.

The three largest underground tests before the TTBT went into effect had yields of well over a megaton (1,000 kilotons). Two of them, one of which took place in September, 1973, and the other in November, 1974, had yields of about two megatons; the third test, in October, 1973, had a yield of about 3.5 megatons. It is reasonable to assume that these were tests of two of the three Soviet single-RV ICBM "mods," or modifications: the SS-17 mod 2, the SS-18 mod 1 and the SS-19 mod 2. The 3.5-megaton warhead would seem to be appropriate for the SS-19 mod 2, since the mod 1 and mod 3 versions of the missile (which had somewhat larger throw weights) accommodated six MIRV's of variously 400- and 500-kiloton yields. A two-megaton warhead would clearly have been too light. The two-megaton warhead seems to be appropriate for the SS-17 mod 2, which has half the throw weight of the SS-19 mod 2.

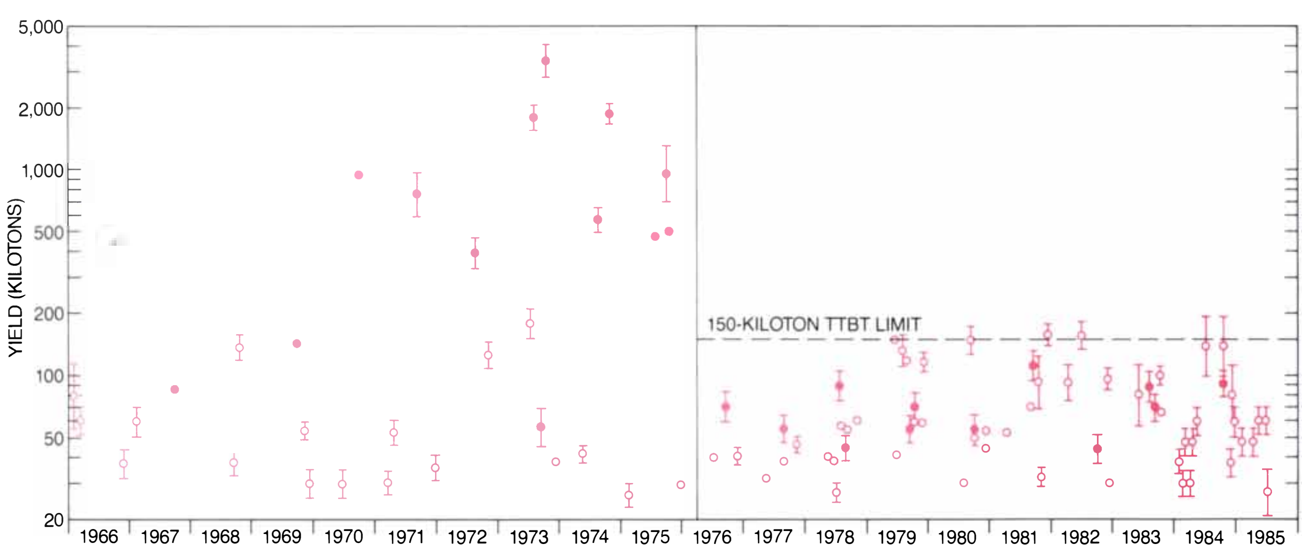
It would seem that warheads were tested underground for the single-RV versions of the SS-17 and SS-19, but not of the large SS-18 missile. This makes sense since the SS-18 replaced the SS-9, which was first deployed in 1966. Apparently the SS-18 mod 1 was designed not only to occupy SS-9 silos but also to carry the SS-9 warhead payload, which had been tested in the atmosphere in the early 1960's. The TTBT intervened before the Soviets could replace the 1960's vintage warhead inherited from the SS-9 (for which we estimate a yield of about 10 megatons) with a newer single warhead, which might well have had a considerably higher yield.

The 400-kiloton test in 1972 was early enough to have been a test of the warhead for the first group of MIRVed Soviet missiles: the SS-17 mod 1, the SS-18 mod 2 and the SS-19 mod 1. These missiles were first flight-tested soon after the warhead test and were deployed between 1974 and 1976. The warhead might have been the same type as the one tested in 1966 for the SS-13 or an upgraded version of it.

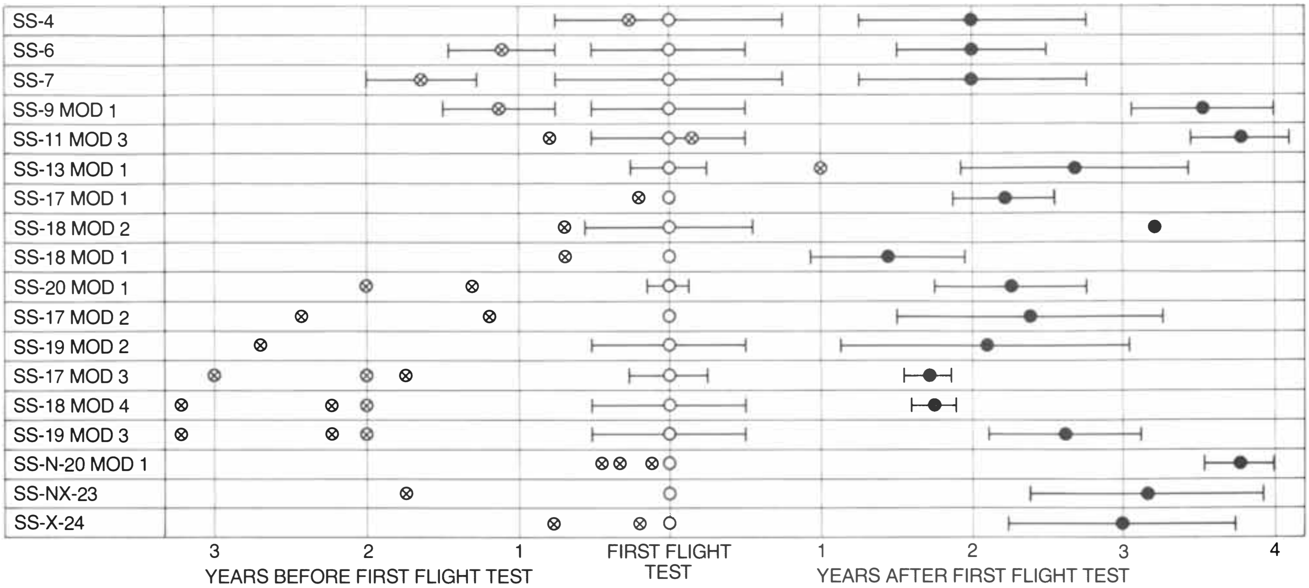
The series of 500-kiloton tests in 1974 and 1975 took place too late to have been tests of the warhead for the first MIRVed Soviet ICBM's. Their timing, although somewhat earlier than usual, makes it likely that they were tests of the warhead for the MIRV's carried on later mods of those same missiles: the SS-17 mod 3, the SS-18 mod 4 and the SS-19 mod 3. These missiles were flight-tested in 1977 and first deployed two years later. The warhead tests were probably rushed to finish them before the TTBT became effective. The fact that three of the five largest Soviet tests in 1974 and 1975 had yields of 500 kilotons argues that the development of weapons of such a yield was of the highest priority to the U.S.S.R. at the time.

These matchings of warhead yield with missile type, based solely on testing chronology, suffice for the construction of diagrams relating warhead yield to missile throw weight for both MIRVed and nonMIRVed Soviet missile systems. From these diagrams we can, in "bootstrap" fashion, deduce other warhead-to-missile pairings. The paramount requirement is that our assignments be consistent with one another and with the historical record of warhead tests and of missile flight tests and deployments.

For example, given that both the SS-11 mod 2 and the SS-N-8 mod 1 (an SLBM) have single warheads and a throw weight of about 900 kilograms, our plot of warhead yield versus throw weight for single-RV Soviet missiles (adjusted for an early 1970's period)



CALCULATED YIELDS of all recorded Soviet nuclear-weapon tests of yield greater than 25 kilotons are plotted chronologically from 1966 through 1985. Yield values for all explosions on Novaya Zemlya (*solid dots*) larger than 110 kilotons are averages based on body-wave and surface-wave magnitudes, as are those for several eastern Kazakhstan tests (*circles*) with yields of between 80 and 150 kilotons from 1976 through 1981; the remaining yields were calculated from body-wave magnitudes alone. The Soviets commonly detonate devices of nearly the same yield within the space of one or two years. The vertical lines show the range of error.



TIMETABLES for the development and deployment of various Soviet missiles reveal a pattern. Anywhere between one year and three years before the first flight test of a missile (*circle*), the warhead designated for the missile is tested at full yield (*circle with cross*) as many as three times. After both the missile and the warhead designs have been judged satisfactory the two components are mated and the missile is finally deployed (*dot*). Deployment generally takes place between one year and four years after the first flight test. Approximate-error bars are included where appropriate. A sense of the typical Soviet pattern of missile testing and deployment helped the authors to pinpoint warhead tests of certain missiles from the record of Soviet underground nuclear tests.

indicates that they are capable of carrying a warhead of a one-megaton yield at most. Indeed, 900-kiloton tests took place in October, 1970, and September, 1971. The fact that the missiles underwent their first flight test before the warhead was tested, as in the case of the SS-13, may be the result of technical difficulties in conducting such large tests so early in the development of the instrumentation for underground testing.

Given our results, we can now compare the range of yields of Soviet strategic weapons with that of U.S. strategic weapons. The smallest warheads in the strategic arsenals of the two countries have yields of about 40 to 50 kilotons and are deployed on MIRVed SLBM's: the Soviet SS-N-18 mod 3 and the U.S. Poseidon C3. The largest nuclear weapon in the U.S. inventory is the nine-megaton warhead for the Titan II missile, which is being phased out. Published reports indicate that single-RV mods of the SS-17, SS-18 and SS-19 ICBM's, all of which contained large-yield warheads, have been replaced by mods carrying smaller-yield, multiple warheads. Once the Titan II is retired (along with a bomb of similar yield) from the operational strategic force, the yields of the largest warheads will be about one megaton for the U.S.S.R. and 1.5 megatons for the U.S. (A possible exception may be a relatively small number of older, high-yield nuclear charges in bombs carried by Soviet aircraft.)

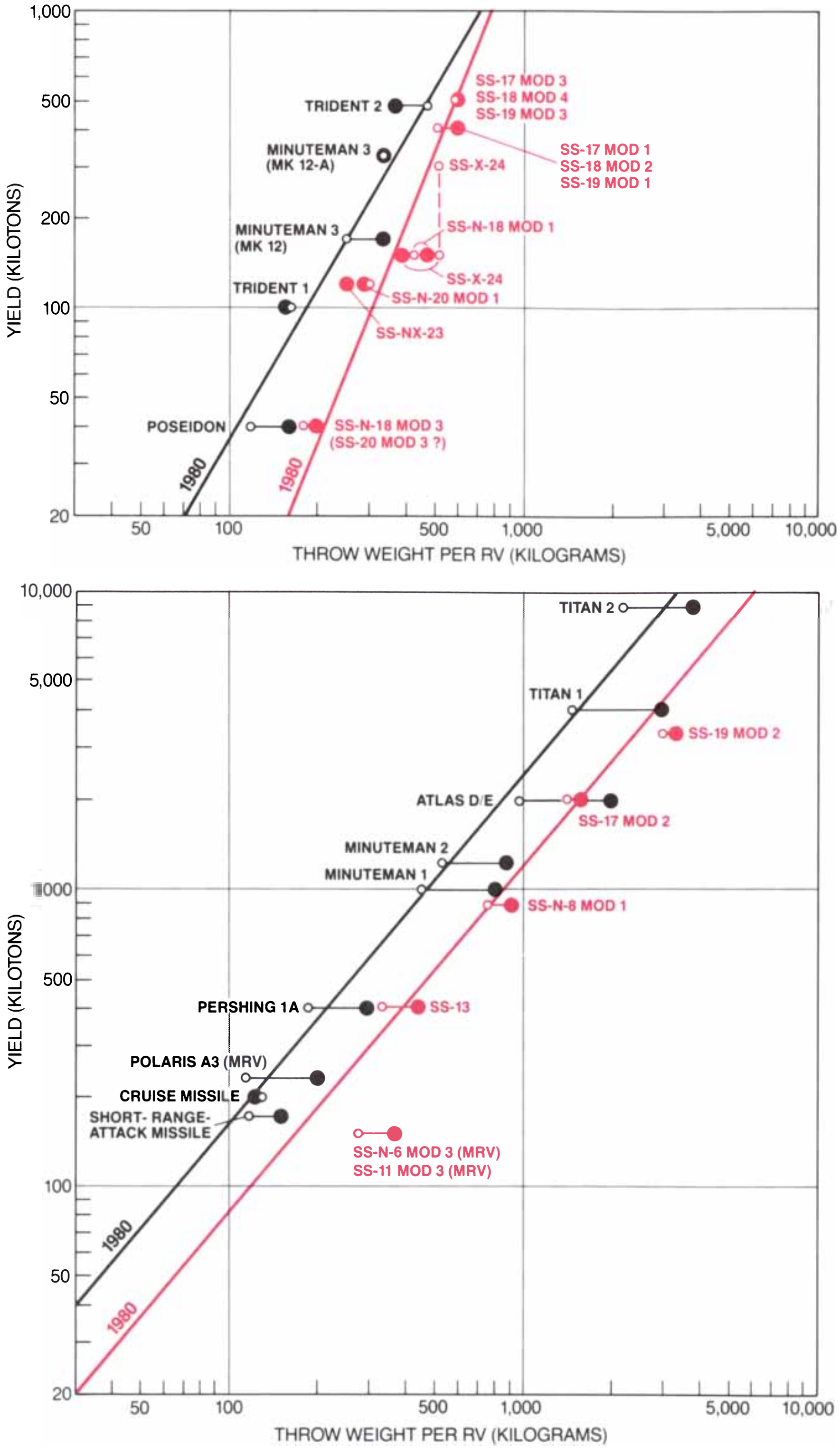
Our yield estimates therefore indicate marked similarities in the warheads of the strategic weapons of the two countries. Other estimates of the yields for Soviet strategic warheads have ranged from 200 kilotons to as much as 50 megatons. The inflated yield estimates, which assign yields much greater than 150 kilotons to all Soviet strategic weapons, would seem to be at odds with the fact that the U.S.S.R. did agree to a 150-kiloton limit in the TTBT. Our estimates of warhead yield make much more sense, since it is highly unlikely that the Soviets would have signed a treaty that essentially prohibited them from testing all operational warheads in their inventory at full yield, particularly since the U.S. was not similarly restrained from further testing and refinement of all its strategic warheads.

Combining our yield estimates with published estimates of the number of Soviet missiles and the number of RV's each missile carries enables us to compile various aggregate measures of the strength of the Soviet strategic arsenal, including total megatonnage. The idea that the U.S.S.R. can count on its strategic systems to deliver a vastly greater number of megatons than the U.S. can deliver with its systems has been widely held for many years. Our revised and more accurate yield estimates indicate, however, that this has not been the case. We should nonetheless point out that estimates of total megatonnage are particularly dependent on assumptions made about the deployment of bombs on long-range aircraft, about which relatively little has been published.

The total megatonnage carried by Soviet strategic weapon systems was substantially less than the megatonnage carried by U.S. strategic weapon systems until the mid-1970's. At that time a dramatic expansion of the Soviet strategic arsenal that had begun in the late 1960's, coupled with a switch to more accurate and smaller weapons on the part of the U.S. strategic forces, resulted in the attainment of rough parity in terms of total megatonnage, number of warheads and ability to destroy such hardened military targets as missile silos.

We should stress that no numerical aggregate is more than a very imprecise measure of strategic capability. The diversified strategic arsenal of the U.S. is not easily compared with the ICBM-dominated arsenal of the Soviet Union. Moreover, we should note that our calculations included only U.S. and Soviet systems capable of striking at intercontinental distances. They do not include weapons aimed at the U.S.S.R. by Britain, France or China. Nor do they count shorter-range systems such as the U.S. intermediate-range ballistic missiles in Europe that threaten targets in the western Soviet Union.

We conclude that although the U.S. and Soviet strategic arsenals are different in their delivery systems, the warheads carried by those systems have comparable yield ranges. Per-



MEASURE OF WARHEAD EFFICIENCY is the ratio of the warhead's yield to the throw weight (per reentry vehicle) of the missile carrying it. In order to formulate a general relation between warhead yield and throw weight, warheads of different vintages must be brought to a common temporal reference frame because newer warhead designs are more efficient than older ones. In these diagrams dots plot the actual yield-to-throw-weight ratio and smaller circles show estimated ratios based on 1980 warhead technology. Straight lines describe the time-corrected relation between yield and throw weight remarkably well. The ratios for Soviet strategic missiles are in color; those for U.S. strategic missiles are in black. Missiles that carry multiple independently targetable reentry vehicles (MIRV's) (*top*) have throw weights per reentry vehicle (RV) that are about twice as great as those of single-RV or multiple-RV missiles (*bottom*) because of the added weight of the bus, the MIRV-dispensing mechanism. The tested yield of the warhead for the new SS-X-24 appears to be too small by 50 percent. The Soviets may not have exploited the missile's capabilities because they were constrained by arms-control treaties.

haps most relevant is the fact that both arsenals are huge and in rough parity.

One clear difference between U.S. and U.S.S.R. systems that does emerge from our analysis is that Soviet warheads are less efficient than U.S. ones, as measured by their ratio of yield to throw weight. In fact, it appears that since the 1960's Soviet warheads have been comparable in efficiency to U.S. warheads of roughly a decade earlier. For example, the RV for the U.S. Minuteman I (deployed in 1962) is quite similar in yield and throw weight to that for the Soviet SS-N-8 mod 1 (deployed in 1973). Likewise, the warhead for a MIRV on the U.S. Poseidon missile (deployed in 1971) is similar in efficiency to that for the Soviet SS-N-18 mod 3 (deployed in about 1978). In calculating warhead yields it is therefore imperative not to assume that the Soviets are on a par with the U.S. in warhead technology.

The TTBT has prevented the Soviet Union from carrying out the testing needed to take more than modest advantage of the larger throw weight of its ICBM's. In addition the SALT II Treaty has limited the number of Soviet delivery systems and the maximum number of RV's on ICBM's and SLBM's. In the absence of both treaties the U.S.S.R. could well have deployed many more missiles and either larger numbers of warheads or warheads of greater yield.

Consider, for example, the dilemma facing the Soviet Union in connection with its new ICBM, the SS-X-24. Kerr, speaking as director of the Los Alamos National Laboratory before a congressional committee in 1985, stated that "one can reasonably safely extrapolate in yield by about a factor of two" the results of a partial-yield test. His testimony suggests that the U.S.S.R. could not have tested within the 150-kiloton limit of the TTBT and still have developed reliable weapons with a yield much larger than 300 kilotons. If the yield-to-throw-weight ratio of Soviet warheads has continued to improve at its historical rate, the Soviets could feasibly arm the SS-X-24 with from eight to 10 RV's, each carrying a 300-kiloton warhead.

Another possible option for the Soviets would be to develop a 150-kiloton warhead so that it could be tested at full yield. In such a case the SS-X-24 might be outfitted with more than 10 RV's, perhaps as many as 14 or 15. Yet that option is precluded because testing and deploying more than 10 RV's on an ICBM is prohibited by SALT II.

The U.S.S.R. must deal with a difficult choice: either trust computer modeling of higher-yield warhead explosions and test at only a fraction of full yield (paying a price in warhead reliability) or develop a 150-kiloton warhead that can be tested at full yield but no more than 10 of which can be deployed on an SS-X-24. Past Soviet testing practices and their lag in computer technology suggest that the Soviets have probably chosen to play it safe and have equipped the SS-X-24 with 150-kiloton warheads even though they would thereby fail to exploit the missile's throw weight. Such a choice, forced by the combined strictures set forth by the TTBT and SALT II, is to the U.S.'s benefit.

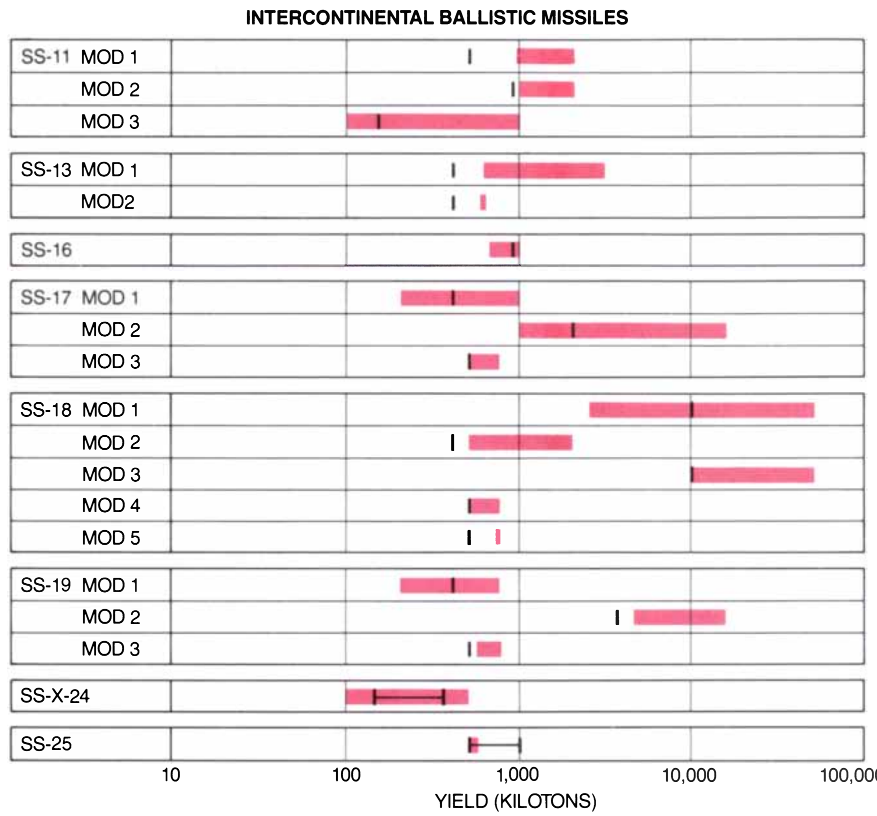
It is even more interesting to consider the potential effect of a lower test-ban threshold, say one of 15 or 20 kilotons. Such a threshold would force the superpowers to test their lower-yield SLBM warheads at partial yield. Since the U.S.S.R. lags far behind the U.S. in the development of SLBM warheads, future Soviet SLBM deployment would be constrained much more severely by a low testing limit than future U.S. deployment would. A lowering of the threshold to about 10 kilotons—the size of tests that could be reliably verified by seismograph stations outside the U.S.S.R.—would prevent full-scale tests of strategic weapons altogether [see "The Verification of a Comprehensive Nuclear Test Ban," by Lynn R. Sykes and Jack F. Evernden; SCIENTIFIC AMERICAN, October, 1982].

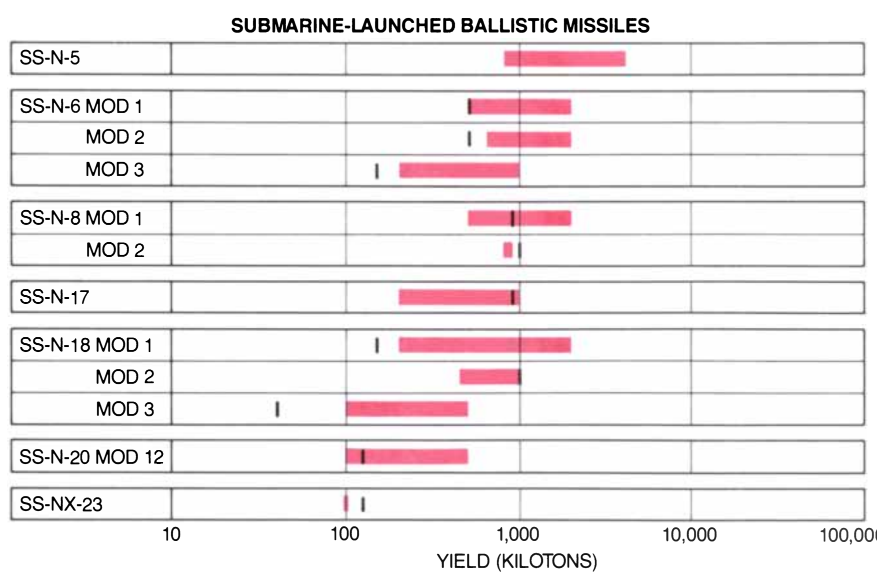
A network of U.S. monitoring stations within the U.S.S.R. would make possible the identification of nuclear testing at a much lower level—a level low enough to halt essentially all nuclear-weapon testing. In this connection the success of the Natural Resources Defense Council in setting up seismograph stations near Semipalatinsk could indicate growing willingness on the part of the Soviet leadership to entertain such a possibility.

It is important to ask why it has taken so long to institute an accurate methodology for estimating yields of Soviet underground explosions. Scientists who work in this area have repeatedly stated over the past 15 years that the U.S. was seriously overestimating Soviet yields. The proper corrective procedure was actually worked out more than 10 years ago, but it was not officially accepted until last year. Recent seismological research has only strengthened our conclusions about the correct methodology for estimating the yield of underground nuclear explosions.

Although there is some logic in the position that it is better to overestimate rather than to underestimate Soviet military capabilities, incorrect yield estimates may well have thwarted attempts to make overall sense of what the U.S.S.R. has in fact been doing. In that sense the price for inaccurate yield estimates may well have been high. Inflated estimates undoubtedly contributed to the perception that the U.S.S.R. was way ahead of the U.S. in various measures of strategic strength. Even U.S. officials were misinformed about the size of Soviet weapons, hampering their efforts to negotiate arms-control agreements.

We need to come to terms with the reality that the U.S.S.R., like the U.S., possesses a huge nuclear arsenal and that limiting the number of Soviet weapons is in our national interest. This can be accomplished most readily by maintaining the limitations set by SALT II and the Threshold Test Ban Treaty and by building on them in future arms-control agreements.





WARHEAD-YIELD ESTIMATES by the authors (*black marks*) are generally lower than the range of estimates in the public literature (*colored bars*). The authors' warhead-yield values imply that the total destructive potential of the Soviet strategic arsenal does not greatly exceed that of the U.S. strategic arsenal, as has been commonly thought.

Uranus

The giant blue-green planet visited last January by Voyager 2 *has one pole pointing toward the sun, and its magnetic field is askew. Its atmosphere is dense and icy, yet its winds resemble the earth's*

by Andrew P. Ingersoll

Uranus, in contrast to all other planets of the solar system, spins on its side, that is, its axis of rotation lies nearly in the plane of its orbit; the spin axes of the other planets are all nearly perpendicular to the orbital plane. How did the anomalous orientation of Uranus arise, and how does it affect the atmospheric circulation? Does the planet have a magnetic field, and if so, what is the field's orientation? These are a few of the questions *Voyager 2* set out to answer during the months around its closest approach to Uranus on January 24, 1986 [see "Engineering *Voyager 2*'s Encounter with Uranus," by Richard P. Laeser, William I. McLaughlin and Donna M. Wolff; SCIENTIFIC AMERICAN, November, 1986].

The encounter provided more new questions than answers, although fortunately it provided both in great numbers. That this is the way things ought to be is sometimes difficult to convey to an impatient public. During the joyful press briefing held by Voyager investigators on January 27, one reporter was bothered that scientists were still bewildered by the new findings and demanded to know why it was taking so long to explain them. Edward C. Stone of the California Institute of Technology, the project scientist for the Voyager mission, replied for all of us. "We are happily bewildered," he said. "We learn the most when we see things that we can't readily explain. If you see things you can explain right away, then you probably haven't learned very much. That means you probably already knew it."

In the months since the flyby, part but by no means all of the initial confusion over the Voyager data has given way to coherent interpretations. In what follows I shall recount some of what has been learned so far about Uranus. I shall focus on the planet itself; *Voyager 2*'s equally rich and fascinating findings concerning the Uranian moons and rings will be described in future issues of *Scientific American* by workers who are experts on these subjects.

Visually the most striking thing about Uranus, whether it is observed from the earth or from the vantage of *Voyager 2,* is just how unstriking it seems: it is a nearly featureless blue-green ball. When *Voyager 2* finally detected patterns in the Uranian clouds, the patterns turned out to be much smaller than the diameter of the planet and only approximately 5 percent brighter than their surroundings. From the earth Uranus' angular diameter is only about four seconds of arc. Since the earth's atmosphere blurs features smaller than about one arc-second regardless of the size of the telescope, ground-based observations can distinguish nothing at all on Uranus. One cannot even see that it is rotating.

Nevertheless, long before *Voyager 2* flew within 80,000 kilometers of Uranus investigators were aware of the strange orientation of the planet's rotational axis. The knowledge came from observing the orbits of the major moons and the rings. The orbits are all nearly circular, and they lie nearly in one plane. This observation suggested that early in its history the Uranian system had settled into a minimum-energy state, in which the moons and rings orbit in the planet's equatorial plane. Several forces conspire to push a planetary system toward this state, including gravitational interactions among its components, collisions with interplanetary debris and drag exerted by gas left over from the formation of the planet.

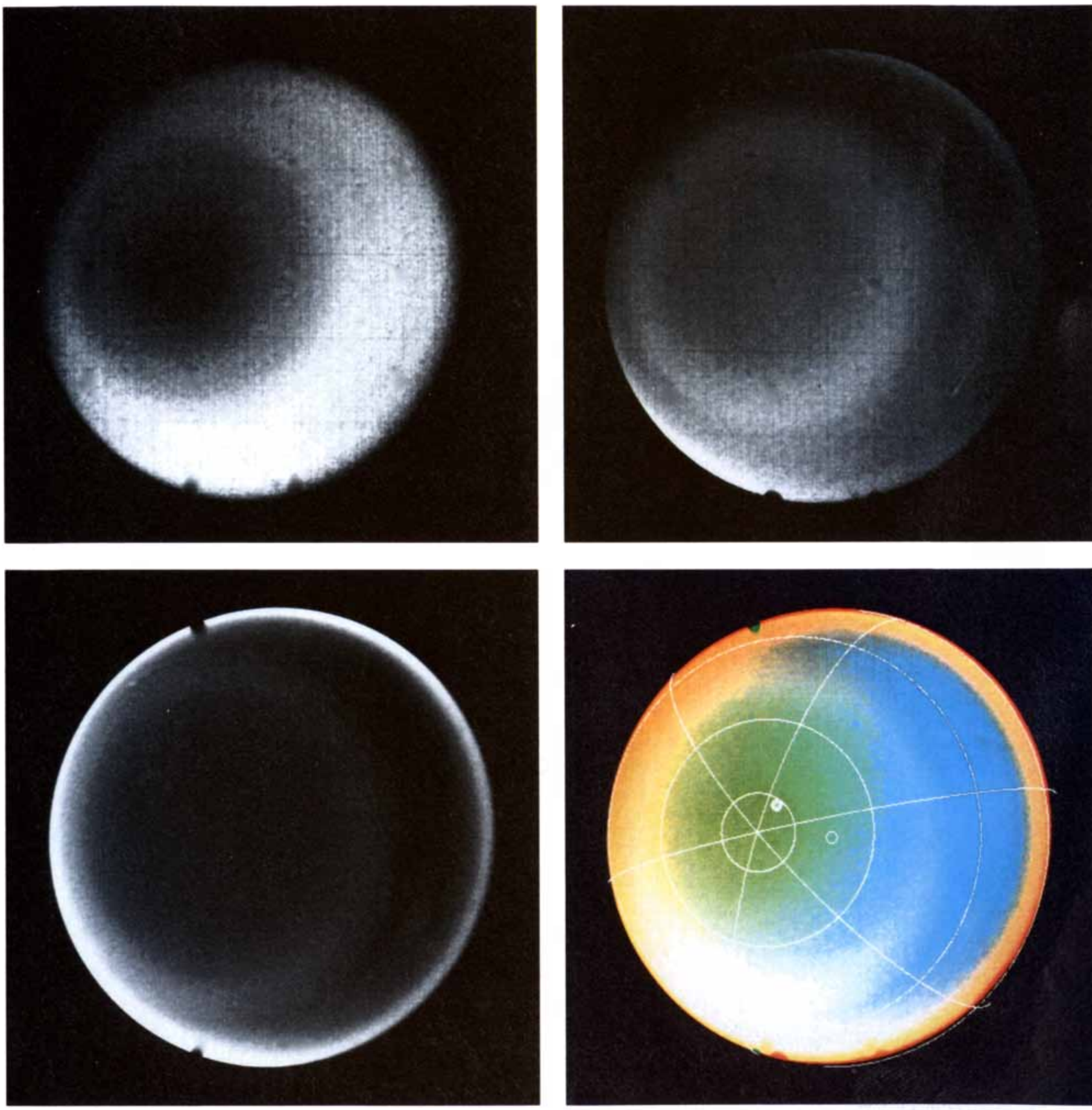
Precise observations have shown that the pole of counterclockwise rotation (which on Uranus, in contrast to the earth, is the south pole) is tilted 98 degrees with respect to the pole of the planet's counterclockwise orbit around the sun. At present the south pole points almost directly at the sun and the earth.

The mass, radius, temperature and atmospheric composition of Uranus were also known before the Voyager mission. The mass had been inferred from the orbital periods of the satellites; it is equal to about 14.5 earth masses. The radius, about 25,600 kilometers (four times that of the earth), had been determined by timing how long stars stay hidden behind the planet. The temperature at a pressure level of .4 earth atmosphere, some distance above the cloud tops, had been calculated by observing infrared (heat) emissions from Uranus; it turned out to be about 59 degrees Kelvin (−214 degrees Celsius). Finally, the atmospheric composition at the cloud tops had been deduced from the infrared spectrum, which includes the signatures of molecular hydrogen (H_2) and methane (CH_4). Selective absorption of reddish sunlight by the methane gives Uranus its blue-green color.

Although the top layer of the atmosphere is mostly gaseous hydrogen, the bulk of Uranus is made of heavier stuff. This conclusion is based on the planet's density, which is 1.27 grams per cubic centimeter. (Liquid water has a density of 1.) The density suggests that Uranus consists mostly of "ices," that is, of substances that would be frozen at the surface of the planet. Specifically, it must consist primarily of water, ammonia and methane, which, because they are compounds of the four most abundant reactive elements in nature (hydrogen, oxygen, carbon and nitrogen), are the most abundant ices in the solar system. At the low temperatures prevailing near the top of the Uranian atmosphere these compounds condense to form clouds of ice crystals. Methane freezes at the lowest temperature, and so it forms the top cloud layer. The methane clouds obscure the underlying ammonia and water clouds, which explains why the signatures of these substances are absent from the planet's infrared spectrum.

The picture that emerged from the ground-based observations put Uranus in a separate class with Neptune, somewhere between hydrogen- and helium-rich Jupiter and Saturn and the rocky, metal- and oxygen-rich planets of the inner solar system. It is worth noting that the classification is not what one would expect based on the most straightforward model of how the solar system formed. According to the model, the preponderance of the light elements hydrogen and helium should increase progressively with distance from the vaporizing heat of the sun. Yet Uranus, with its concentration of ices, contains more relatively heavy elements than Jupiter and Saturn do, and Neptune is heavier still. The icy components of Uranus and Neptune may have come from comets, which are stabler in the outer solar system, but the question remains open.

Uranus bridges a second gap between the Jovian and the terrestrial planets. It seems to have lost most but not all of the internal heat it had when it formed. As much as 30 percent of the heat radiated by the planet may come from its interior rather than

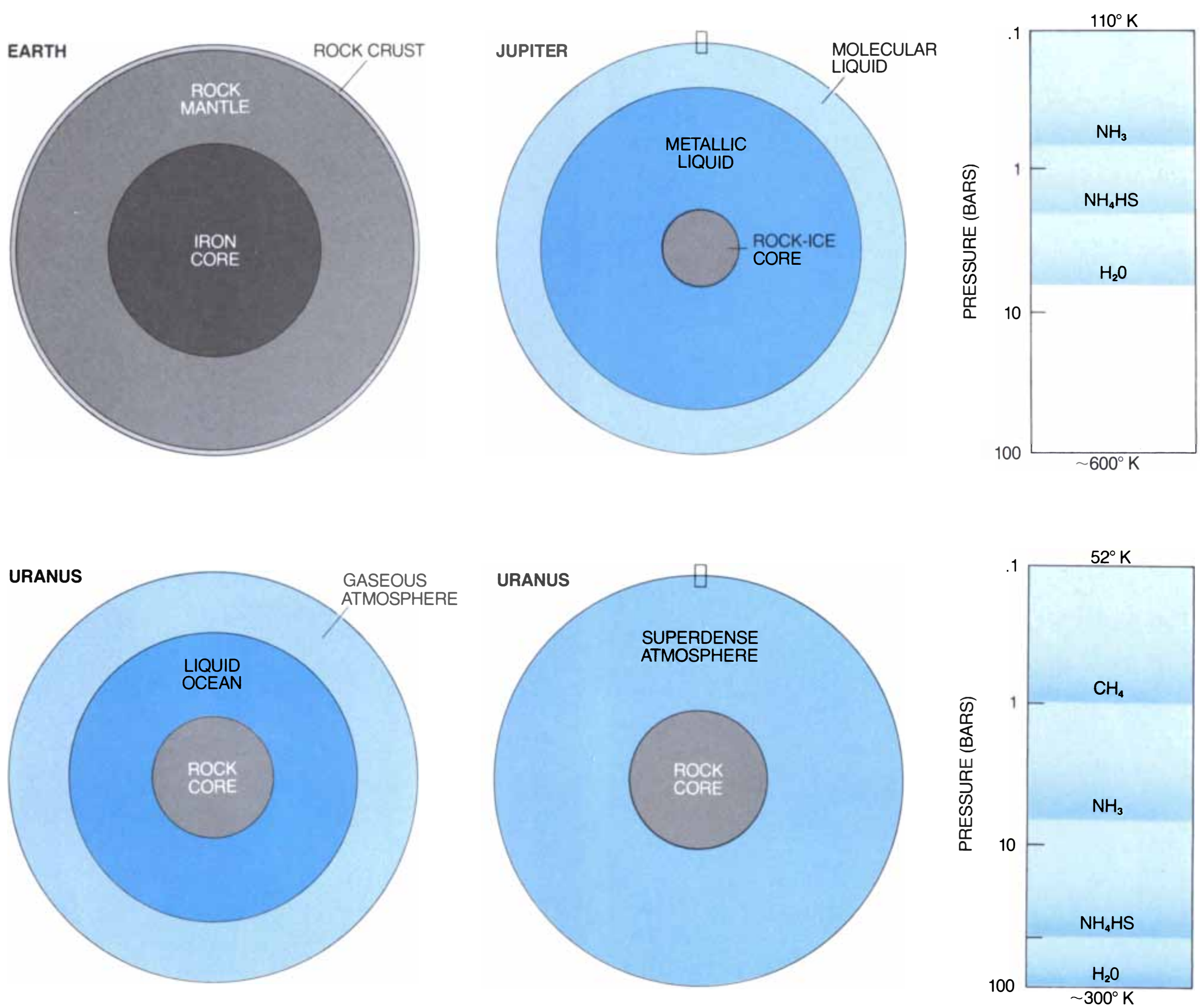


URANUS would be virtually featureless to an observer in space, but Voyager images in which the contrast has been drastically enhanced reveal cloud bands. The black-and-white images were made through three filters: violet (*top left*), orange (*top right*) and methane orange (the color selectively absorbed by methane gas) (*bottom left*); the false-color image is a composite of the three. The latitude-longitude grid on the false-color image shows that the cloud bands are centered on the pole and not on the point directly below the sun (*white dot*). The white circle indicates the point that was directly below the spacecraft when it made the images.

from the sun. The corresponding figure for the earth is .01 percent; for Jupiter and Saturn, which are much more massive and have therefore retained more of their internal heat, the figure is at least 70 percent. The strength of a planet's internal heat source is an important clue to its evolution. Atmospheric circulation patterns are also influenced by the extent to which the atmosphere is heated from below. One of the goals of the *Voyager 2* mission was to refine the rather uncertain estimate of Uranus' internal heat. The calculation requires the analysis of a large number of observations, however, and it has not yet been completed.

My own role in the Uranus flyby was in planning and analyzing the atmospheric observations, particularly those pertaining to clouds and winds. At first it was discouraging work, owing to the featurelessness of the planet. Commands to the spacecraft had to be programmed long before anything was visible on Uranus. Nevertheless, I and other workers interested in the Uranian atmosphere proposed that the spacecraft make many images of the planet. Our colleagues on the imaging team (which was then waggishly called the "imagining team") agreed—partly because they were sympathetic and partly because, until the day before the encounter, the planet was the only thing large enough to fill the field of view of *Voyager 2*'s narrow-angle camera.

During the last few months of 1985 the planet grew steadily in the field of view, but it remained as dull as ever. The problem was that variations in sunlight dominate the variation in brightness across the planet's disk; real atmospheric features are washed out by the solar glare. Fortunately the sun-



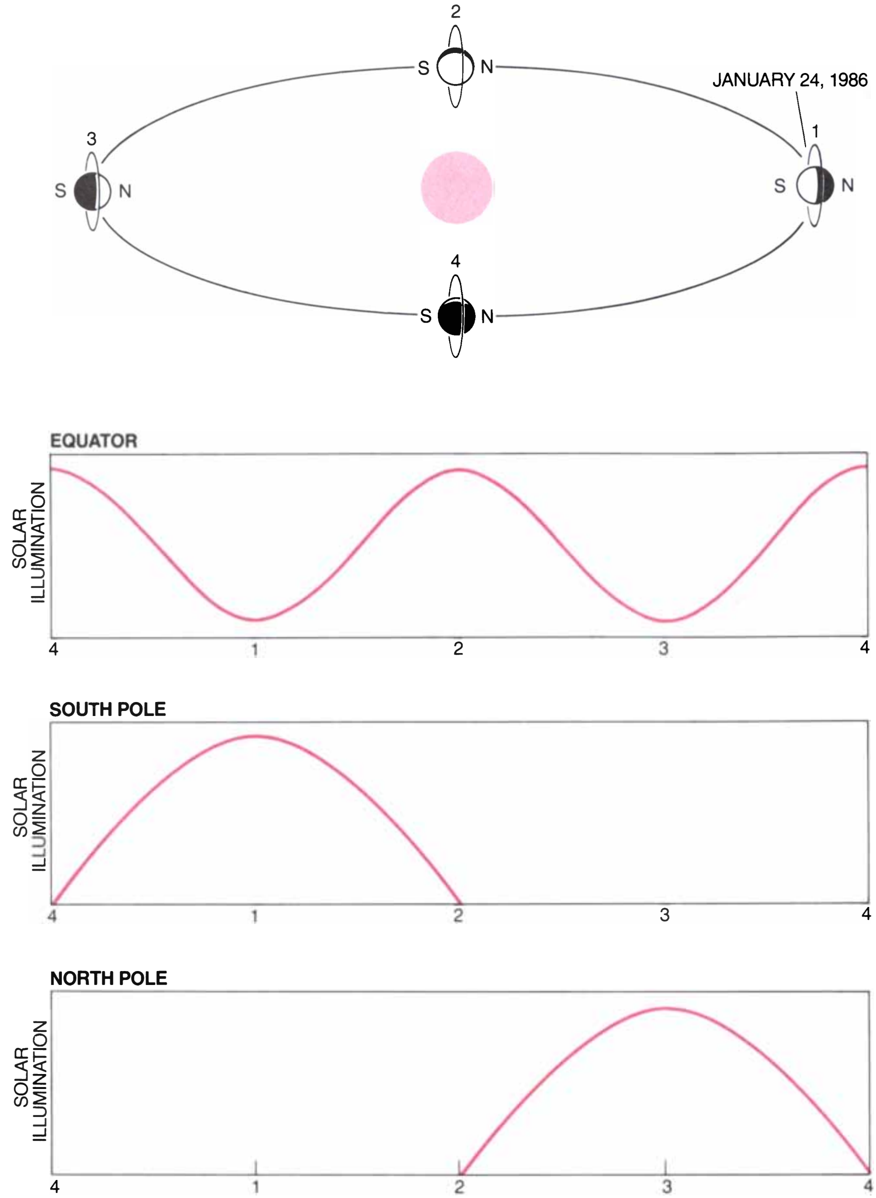
INTERNAL STRUCTURES of the earth, Jupiter and Uranus are compared. (Their relative sizes have been distorted; Uranus is about four times and Jupiter more than 11 times the size of the earth.) The earth, like Mercury, Venus and Mars, is a dense, rocky planet consisting mostly of metals and their oxides. Jupiter, like Saturn, consists mostly of hydrogen and helium, which form a molecular liquid in the outermost layer and a metallic liquid—that is, a mixture of protons and free electrons—under the intense pressures prevailing at greater depths. Uranus, like Neptune, is intermediate between the terrestrial and the Jovian planets, consisting mostly of water, ammonia (NH_3) and methane (CH_4) "ices." In the three-layer model of Uranus (*left*) melted ices form a liquid "ocean" between the rocky core and the gaseous hydrogen-helium atmosphere. Voyager data, however, favor a two-layer model (*right*) in which the gases and ices are mixed in a dense atmosphere. Near the visible surface layers of the atmosphere of Jupiter and Uranus, ammonia, ammonium hydrosulfide (NH_4HS) and water are thought to condense (in a sequence determined by their condensation temperatures) to form icy clouds. Uranus is cold enough to allow methane to condense above the other clouds.

light distribution is regular—the planet is brightest at the subsolar point near the south pole and grows progressively darker toward the equator—and so it can be modeled mathematically. Not long after Thanksgiving, Charles Avis, Robert H. Brown and Torrence Johnson of the Jet Propulsion Laboratory worked out a mathematical expression for the brightness variation on a hypothetical planet similar to Uranus but truly featureless. When the model was subtracted from the observed brightness variation on Uranus, faint cloud features became visible.

As I noted above, the features are only a few percent brighter than their surroundings, but the contrast could be exaggerated and displayed in false color. The enhanced images revealed a series of cloud bands concentric with the pole of rotation. Superposed on the bands and between them were smaller features that circled the pole counterclockwise, each at a constant latitude. Features at different latitudes circled at different rates, with periods ranging from 14 to 17 hours. Clearly, therefore, the clouds were not simply being carried by the planet's rotation. We had observed winds.

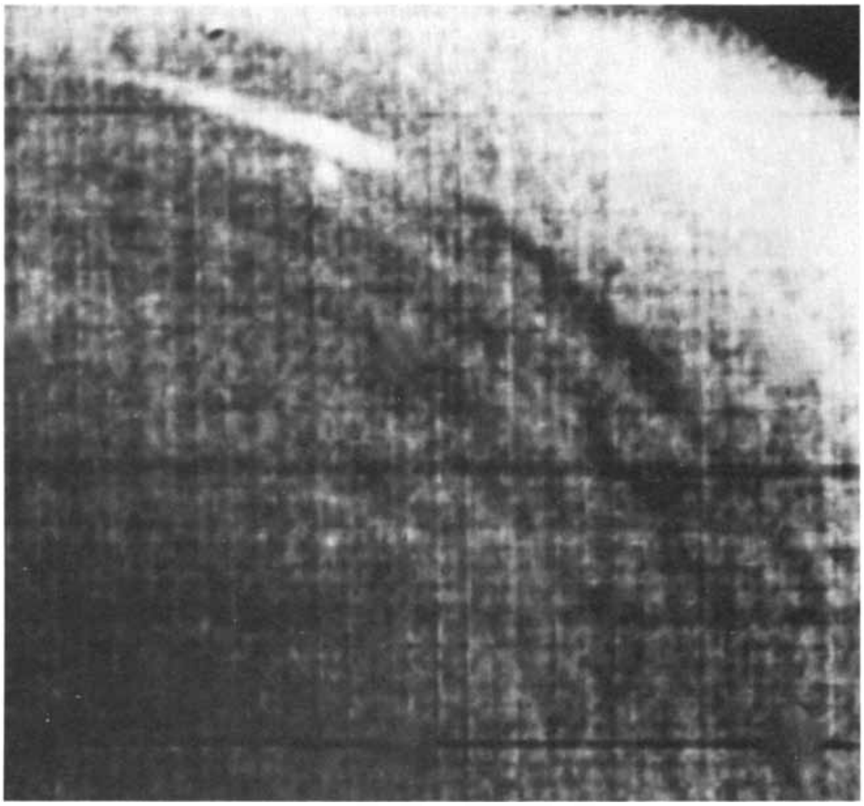
At last, after months of expectation, we had a scientific result, and a significant one at that. First, the cloud bands did not have to be concentric with the pole, but they were. Second, the winds did not have to blow in an east-west direction, but they did. Both the banding and the east-west winds are similar to what one finds on Venus, the earth, Jupiter and Saturn. Before *Voyager 2*'s encounter with Uranus one might have expected its atmospheric circulation to be different. On Uranus as on the other planets the sun supplies most of the energy that drives the circulation. At the time of the encounter the sun was almost directly above the south pole, the north pole had been in darkness for about 20 years and the equator was in constant twilight. The sunlight distribution was thus completely different from that on the other planets, whose axes are much less tilted. And yet the atmospheric circulation was similar.

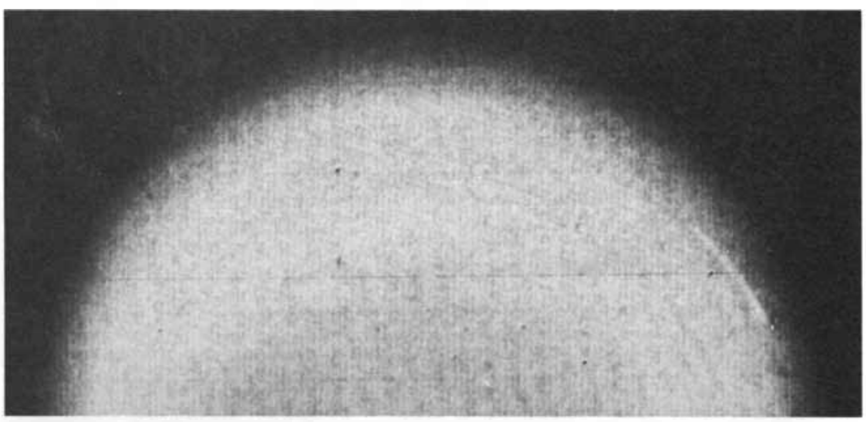
Apparently the sun, although it supplies the energy that drives a planet's atmospheric circulation, does not determine the pattern of the circulation. Instead the pattern is dominated by the effects of a planet's rotation. The rotation gives rise to what is called the Coriolis force, which steers the winds into zones of constant latitude. If an atmospheric parcel curves away from its latitude zone, the Coriolis force sends it spinning back. Spacecraft observations of the planets, particularly Uranus, have shown just how important the Coriolis force is. We have learned that atmospheric circulations are not so much forced by the sun as they are coasting under their own inertia.

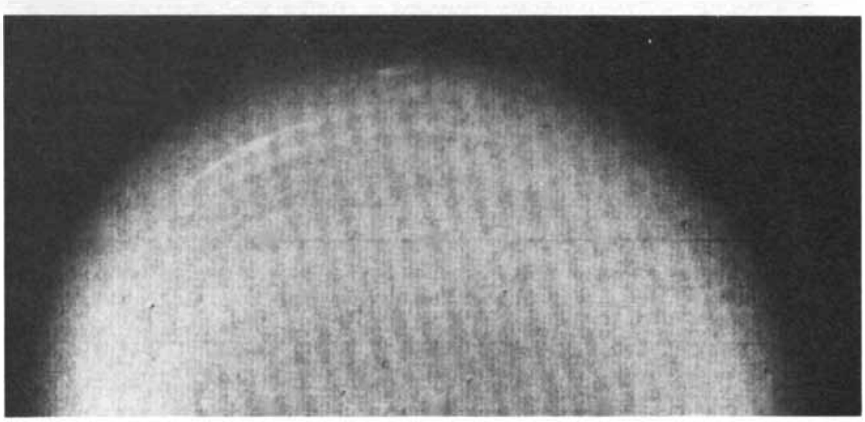


SEASONAL CHANGES in solar illumination on Uranus are vastly different from the changes on other planets because the planet lies on its side. The Uranian year is about 84 earth years long. At present the south pole is pointed toward the sun and the north pole is in constant darkness; 42 earth years from now the conditions will be reversed. The equator, now in constant twilight, has two such winters and two summers every year.

By early January of 1986, when observations of cloud features on Uranus were already in hand, the investigators involved in *Voyager 2*'s magnetic field and charged-particle experiments were still anxiously awaiting their first results. If Uranus had no magnetic field, these investigators would have nothing to observe but the solar wind of charged particles streaming past the planet; the planet's only effect would be to create a wake in the flow. There would be nothing to say about dynamo processes and electrically conducting regions inside the planet, other than that apparently there were none. Furthermore—and this was of concern to atmospheric scientists as well—there would be no way to determine the planet's internal rate of rotation. On the giant planets such as Uranus, which lack a solid surface, an internally generated magnetic field provides the only fixed reference frame against which atmospheric motions can be measured.







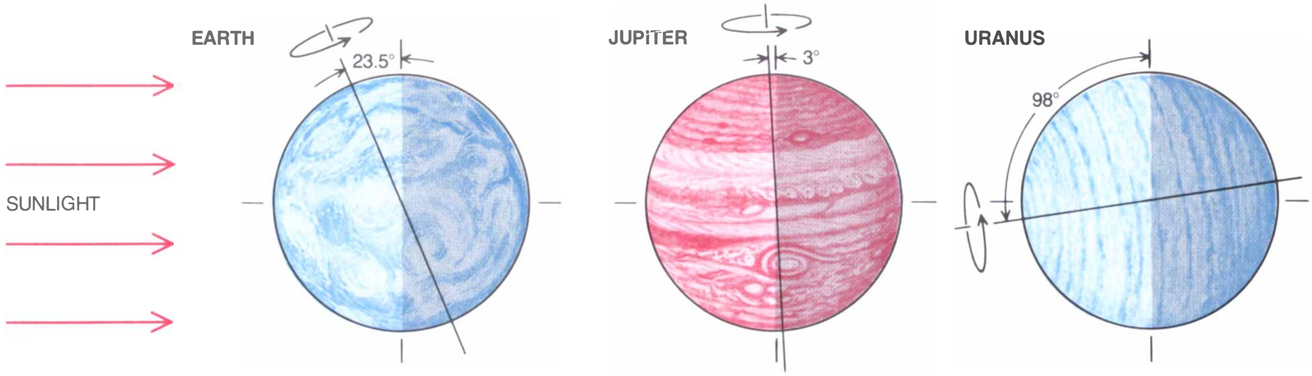
BRIGHT CLOUD FEATURES on Uranus were observed by *Voyager 2* within and between the concentric cloud bands. The clouds move counterclockwise around the pole of rotation; features at different latitudes move at different speeds, indicating they are being carried by east-west winds whose intensity varies with latitude. The feature in the closeup view (*top*) is probably the plume of a convective updraft. By tracking it in a time-lapse sequence one can measure the wind speed. In the middle photograph the feature is at the two o'clock position; in the bottom photograph it has moved to the 11 o'clock position.

Dejection grew among the members of the *Voyager 2* team as the spacecraft passed the point at which theoretical models had predicted that the effects of a magnetic field would begin to be observed. Then, only five days before the closest approach, the spacecraft detected radio signals and charged-particle streams emanating from Uranus. By analogy with other planets the radio emissions had to come, either directly or indirectly, from charged particles spiraling around magnetic field lines. The modulation of the emissions suggested that the Uranian magnetic field was tilted in relation to the planet's axis and was therefore wobbling as the planet rotated.

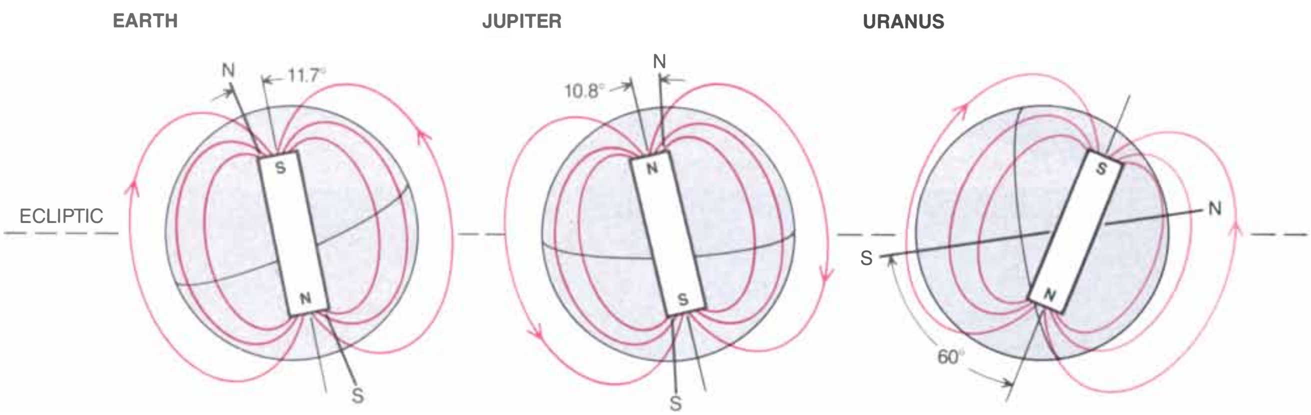
At this point *Voyager 2* had still not entered the magnetic field. On the sunward side of the planet the field is confined to a comparatively small region by the solar wind, which deforms the field and sweeps it into a long tail behind the planet. Right outside this region (the magnetosphere), where the field can just hold its own against the solar wind, a bow shock forms. (The bow shock is analogous to the shock wave that precedes an airplane traveling at supersonic speeds, but it is an electromagnetic disturbance rather than a pressure wave.) *Voyager 2* crossed the bow shock on January 24, only 10 hours before its closest approach to the planet. The magnetometer team, under the leadership of Norman F. Ness of the Goddard Space Flight Center, was soon mapping the shape and strength of the Uranian magnetic field.

It was not as simple a field as had been expected. Every other planetary magnetic field is dominated by a dipole component, the equivalent of a small but powerful bar magnet at the center of the planet. On the earth, Jupiter and Saturn the hypothetical bar magnet is nearly parallel to the planet's rotation axis. (The tilt angle of the earth's field, 11.7 degrees, is the largest.) Quadrupole and octupole components describing irregularities in the dipole field are important mainly inside these three planets, near the electrically conducting core whose fluid motions are thought to generate the field. On Uranus, in contrast, the bar magnet is tilted by 60 degrees with respect to the rotation axis, and the other field components are almost as strong as the dipole. To suppress these components and make a pure dipole model fit the data one has to move the dipole off center by 30 percent of the planet's radius.

Might the uncommonly large tilt of Uranus' magnetic axis be associated with the unusually large angle between the rotation axis and the orbit axis?



EAST-WEST WINDS dominate the atmospheric circulation on Uranus, much as they do on the earth and Jupiter, even though the distribution of sunlight is entirely different on Uranus as a result of the unusually large tilt of its spin axis. The similarity suggests that the Coriolis force arising from a planet's rotation has a dominant influence on its atmospheric circulation pattern.



MAGNETIC FIELD of Uranus is tilted by 60 degrees with respect to the planet's spin axis, and the hypothetical dipole magnet that models the field most accurately is offset from the center of the planet. (The south end of the dipole is the one toward which a compass needle would point.) In contrast, the dipole fields of the earth and Jupiter are tilted only slightly, and they are not offset.

For that to be the case the interior of Uranus, where the magnetic dynamo resides, would have to "know" where the sun is, because the sun defines the orbit. One possibility is that the interior is affected by the sun's differential gravitational pull: the pull is stronger on the day side than on the night side of the planet. My colleague David J. Stevenson of the California Institute of Technology has shown, however, that this tidal effect is much too slight (Uranus is 19 times as far from the sun as the earth is) to have a significant impact on the dynamo region inside the planet.

Ness and his co-workers have speculated that Uranus may be undergoing a magnetic field reversal, which might explain both the tilt of the magnetic axis and the offset of the dipole. (On the earth there is geologic evidence of numerous field reversals, although the fraction of time spent going from one polarity to the other is small.) Alternatively, the offset may simply indicate that the dynamo region is closer to the surface on Uranus than it is on other planets. Uranus contains a lot of water and ammonia, which become good electrical conductors at lower pressures (and hence at shallower depths) than the hydrogen and helium that predominate on Jupiter and Saturn.

In truth, however, neither the tilt of Uranus' magnetic axis nor the offset of the dipole has been adequately explained. Yet one should remember that no planetary dynamo, including the earth's, is well understood, primarily because observational data on the interior of a planet are so hard to come by. The Uranian magnetic field appears strange; on the other hand, if a larger sample of planets were available, one might find that a substantial number have fields inclined by 60 degrees or more.

The magnetosphere of Uranus extends to an altitude of at least 590,000 kilometers on the day side of the planet and to about six million kilometers on the night side. Like the magnetospheres of other planets, it is filled with an ionized gas, or plasma, composed of equal numbers of positive ions (primarily protons) and electrons. The particles are trapped in the magnetic field, and they oscillate between the north and south magnetic poles. Their average energy increases toward the planet. Indeed, the *Voyager 2* charged-particle team headed by Stamatios M. Krimigis of Johns Hopkins University found that Uranus has radiation belts (regions of high-energy particles) similar to the earth's Van Allen belts.

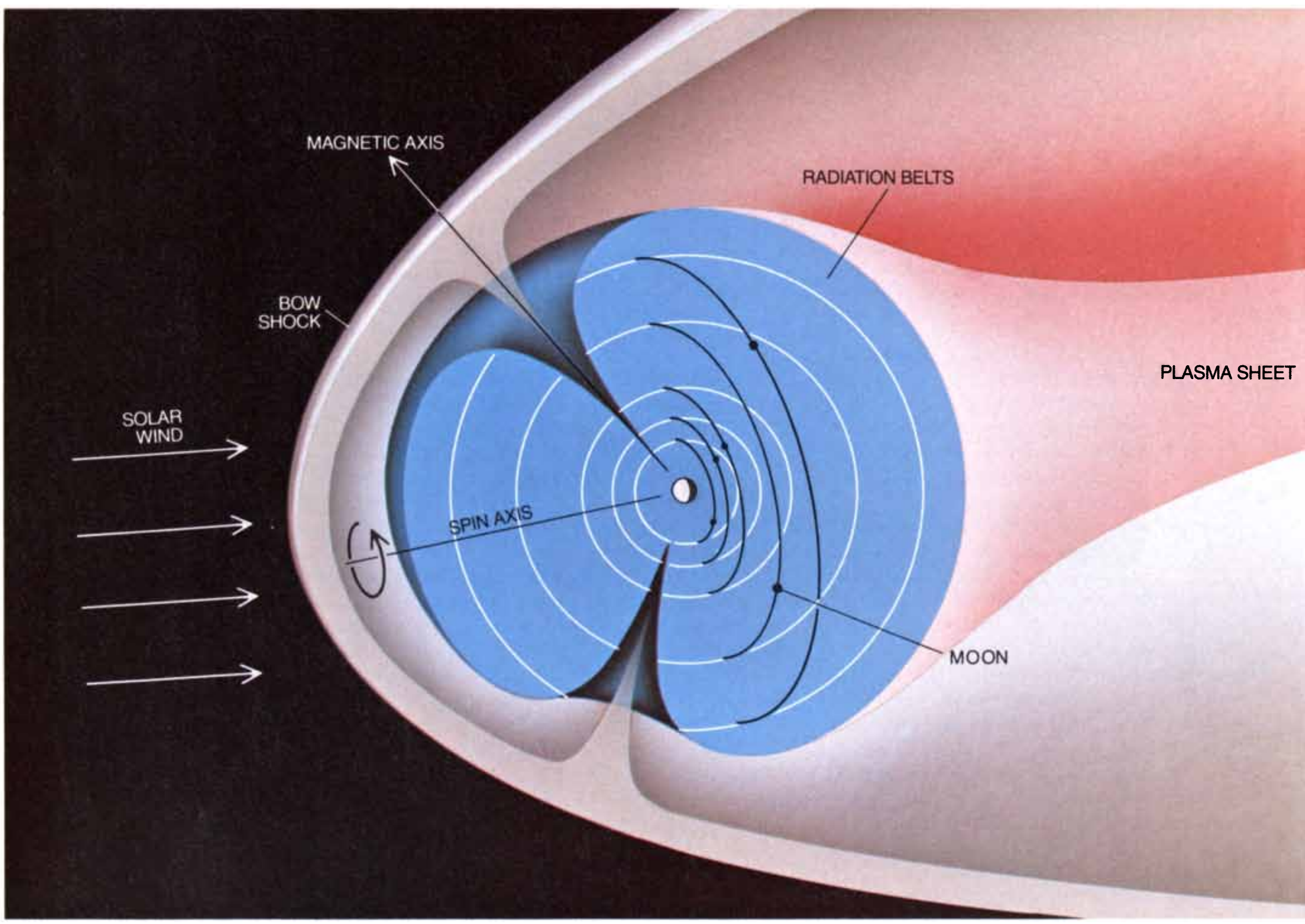
The radiation in the Uranian belts is so intense that within a few million years it can cause significant damage to exposed surfaces. This effect may explain the dark color of the rings and the dark patches on the moons. The moons and the rings orbit within the radiation belts, and as a result they sweep up high-energy particles. If, as is generally thought, their surfaces consist in part of methane ice, then radiation-belt protons may break down the methane and convert it into complex hydrocarbons that have a dull, dark color.

Voyager 2 spent more than two Uranian days in the magnetosphere. Since the magnetic field is fixed to the planet, the planet's rotation rate could be determined from the periodic fluctuations of the field intensity. It could also be calculated from the radio emissions that had been the first sign of a magnetic field on Uranus. The emissions come from near the magnetic poles, and so they too fluctuate periodically as the magnetic axis precesses about the rotation axis. The planetary radio-astronomy team headed by James W. Warwick of Radiophysics, Inc., in Boulder, Colo., observed more than 10 cycles of the radio emissions. According to the workers' calculations, which agree with those based on the magnetic field intensity, Uranus rotates once every 17.24 hours. The best guess before the flyby had predicted a somewhat faster rotation.

The internal rotation rate of a giant planet, in conjunction with the size of its equatorial bulge, is a sensitive probe of its internal structure. The faster the rotation is, the stronger is the centrifugal force on the planet, the more mass is shifted toward the equator and the larger is the equatorial bulge. The size of the bulge also depends on the distribution of mass inside the planet: if two planets have the same mass, radius and rate of rotation, the one that has more of its mass concentrated near the center will have a smaller bulge. The bulge can be measured visually, or its size can be inferred from its gravitational effect on the orbits of the planet's moons and rings. In the case of Uranus the equatorial diameter of the planet is about 2.4 percent larger than the polar diameter.

Even before the Voyager encounter William B. Hubbard and Joseph J. MacFarlane of the University of Arizona had used this figure and the best estimate of the rotation rate to evaluate different models of Uranus' internal structure. The models differ in the relative proportions and degree of mixing of three principal components: rock (metals and metallic oxides), ice (water, methane and ammonia) and gas (hydrogen, helium and neon). One popular model had the three components entirely separate; a gaseous atmosphere overlay a deep "ocean" of ices melted by the high temperatures inside the planet, and the ocean surrounded a rocky core. Hubbard and MacFarlane found that this mass distribution was too centrally concentrated to produce Uranus' rather large equatorial bulge.

When *Voyager 2*'s lower value for the rotation rate is introduced into the calculation, the bulge produced by the three-layer model becomes even smaller and the discrepancy with the observations becomes even greater. Instead the spacecraft data favor a two-layer model, in which the ices and gases are mixed in a dense atmosphere



MAGNETOSPHERE of Uranus is produced by the interaction of its magnetic field with the solar wind. A bow shock analogous to the shock wave preceding a supersonic aircraft forms "upstream" from the magnetic field. The magnetosphere, which begins slightly inside the shock front, is filled with a plasma of protons and electrons, some of which probably come from the solar wind and some of which come from hydrogen in the planet's atmosphere. The charged particles are trapped in the magnetic field; those with high energy oscillate back and forth between the magnetic poles, forming torus-shaped radiation belts. Low-energy particles are

extending from the core to the visible layers of the planet. The bulk of the atmosphere is probably water. Toward the top of the atmosphere, where the temperature declines rapidly to a minimum of 52 degrees K., the water, ammonia and methane condense (in that order) to form thick, icy cloud layers. The top layer, the methane, is visible in the Voyager images. Above it lies a thin upper atmosphere, a gaseous mixture that consists primarily of hydrogen with a little helium and neon.

The internal rotation rate deduced from the Uranian magnetic field came as a surprise to meteorologists, because the 17.24-hour rotation period is longer than the periods of the cloud features seen in the Voyager images. In other words, the atmosphere at the level of the cloud tops is rotating faster than the interior of Uranus, at least in the 25-to-70-degree latitude band where features were observed. The difference is greatest at high latitudes, where the features circle the poles in 14 hours, and it gets progressively smaller toward the equator. Near the equator the relation seems to be reversed, so that the atmosphere rotates slower than the interior.

most abundant in the plasma sheet, which separates the north magnetic hemisphere from the south. As the moons orbit the planet in the plane of the equator, they sweep out particle-free lanes (*white lines*).

The distribution of wind intensities is surprising for the same reason that the dominance of east-west winds on Uranus is surprising: because it closely resembles the pattern on the earth. For example, in the earth's mid-latitudes the circulation is dominated by high-altitude eastward jet streams. The pattern is a direct consequence of the fact that the Equator is hotter than the poles: the latitudinal temperature gradient creates a pressure gradient aloft that is balanced by the Coriolis force, resulting in eastward winds. Since Uranus lies on its side, one would expect its poles to be hotter than its equator rather than colder. Nevertheless, the rapid rotation of cloud features at high latitudes indicates that Uranus has winds like the earth's jet streams.

There are two ways to explain the apparent contradiction. First, the Uranian circulation may be driven not by temperature gradients but by density gradients associated with condensation and precipitation. When water vapor condenses in part of the earth's atmosphere, the density of the atmosphere in that area changes by less than 2 percent. The dense Uranian atmosphere, however, might be as much as 50 percent water. If for some reason a substantial part of the water near the Uranian equator were to condense, the resulting density gradient might mimic the earth's temperature gradient and drive a poleward flow. (A terrestrial analogue would be ocean currents that are driven by salinity differences rather than by temperature gradients.)

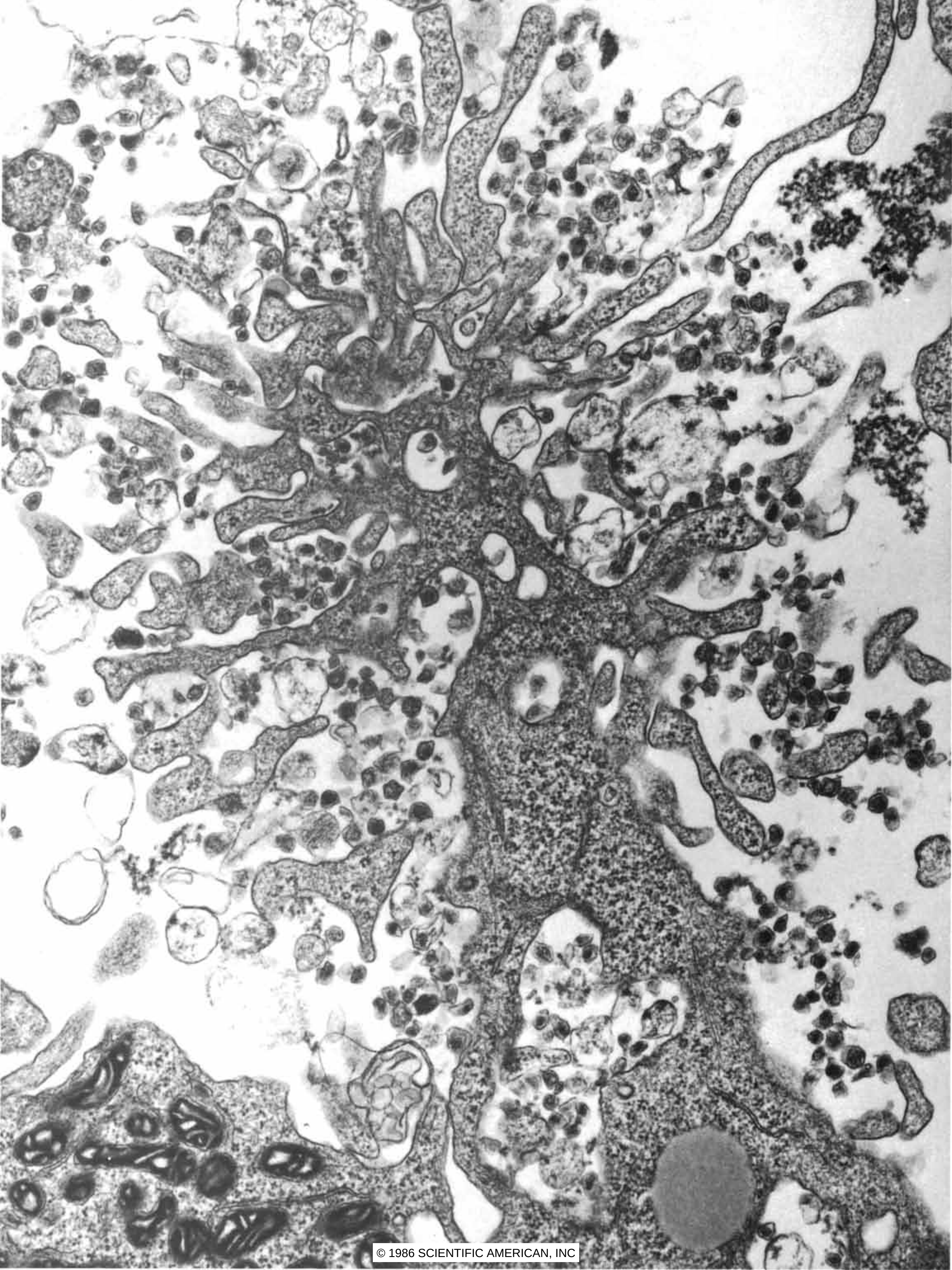
Alternatively, the poles of Uranus may not really be hotter than the equator even though they receive more sunlight. The *Voyager 2* infrared spectrometer team, led by Rudolph A. Hanel of the National Aeronautics and Space Administration's Goddard Space Flight Center, measured the temperature from one pole to the other just above the cloud tops, at a constant pressure level of about .6 earth atmosphere. (The pressure at the cloud tops is roughly one earth atmosphere.) The group found the same temperature, 64 degrees K., at both poles and at the equator; in the mid-latitudes of both hemispheres the temperature was between one and two degrees lower.

Theoretical models had actually predicted that the two poles would have nearly the same temperature. Sunlight is so weak at Uranus that seasonal temperature swings should be no more than about two degrees K. Moreover, James Friedson of Caltech and I have calculated that winds may limit the seasonal swings by transferring heat from one hemisphere to the other; the cooling of the dark pole may also be balanced by enhanced convection of internal heat. Still, none of the models can explain why the equator should be as warm as the poles. Apparently heat circulates in the Uranian atmosphere in a more complex way than the models envision.

The atmosphere of Uranus does not stop at the planet's visible surface. Above the clouds there is a tenuous upper atmosphere consisting primarily of hydrogen molecules. (The breakdown of this hydrogen by sunlight and by charged particles may be the major source of the protons and electrons that make up the radiation belts.) The temperature in the upper atmosphere rises to a warm 750 degrees K., causing it to balloon 6,000 kilometers above the cloud tops. Sunlight alone cannot account for the high temperature; another source of energy must be involved.

Whatever the energy source is, it probably also underlies the curious emissions observed by the ultraviolet-spectrometer team under the leadership of Lyle Broadfoot of the University of Arizona. The emissions were detected only on the day side of Uranus, indicating that sunlight is required as a stimulus. A similar phenomenon was observed at Jupiter and Saturn. It has been named electroglow on the theory that electrons may be exciting hydrogen molecules in the upper atmosphere of all three planets. How the electrons get their energy is not known.

A similar degree of uncertainty persists concerning the answer to the biggest question about Uranus: Why is it spinning on its side? Although *Voyager 2* did not find a smoking gun at Uranus, the spacecraft has over the years found dramatic evidence of violent collisions in the early solar system. The moons of Jupiter, Saturn and Uranus all bear the scars of impacts large enough to have nearly disrupted them. As the debris orbiting the newly formed sun aggregated into planet-size bodies, the final collisions would have been the largest; at least one body the size of the earth probably collided with what is now Uranus. A large, off-center impact could have knocked the planet onto its side. This is the hypothesis most investigators now accept. It remains to be confirmed or disproved by future spacecraft missions.



The AIDS Virus

Part II of a two-part article on the human retroviruses. In 1984 the cause of AIDS was shown to be the third human retrovirus. Although knowledge of the virus was rapidly obtained, its toll will be heavy

by Robert C. Gallo

It is a modern plague: the first great pandemic of the second half of the 20th century. The flat, clinical-sounding name given to the disease by epidemiologists—acquired immune deficiency syndrome—has been shortened to the chilling acronym AIDS. First described in 1981, AIDS is probably the result of a new infection of human beings that began in central Africa, perhaps as recently as the 1950's. From there it probably spread to the Caribbean and then to the U.S. and Europe. By now as many as two million people in the U.S. may be infected. In the endemic areas of Africa and the Caribbean the situation is much worse. Indeed, in some areas it may be too late to prevent a disturbingly high number of people from dying.

In sharp contrast to the bleak epidemiological picture of AIDS, the accumulation of knowledge about its cause has been remarkably quick. Only three years after the disease was described its cause was conclusively shown to be the third human retrovirus: human *T*-lymphotropic virus III (HTLV-III), which is also called human immunodeficiency virus (HIV). Like other retroviruses, HTLV-III has RNA as its genetic material. When the virus enters its host cell, a viral enzyme called reverse transcriptase exploits the viral RNA as a template to assemble a corresponding molecule of DNA. The DNA travels to the cell nucleus and inserts itself among the host's chromosomes, where it provides the basis for viral replication.

In the case of HTLV-III the host cell is often a *T*4 lymphocyte, a white blood cell that has a central role in regulating the immune system. Once it is inside a *T*4 cell, the virus may remain latent until the lymphocyte is immunologically stimulated by a secondary infection. Then the virus bursts into action, reproducing itself so furiously that the new virus particles escaping from the cell riddle the cellular membrane with holes and the lymphocyte dies. The resulting depletion of *T*4 cells—the hallmark of AIDS—leaves the patient vulnerable to "opportunistic" infections by agents that would not harm a healthy person.

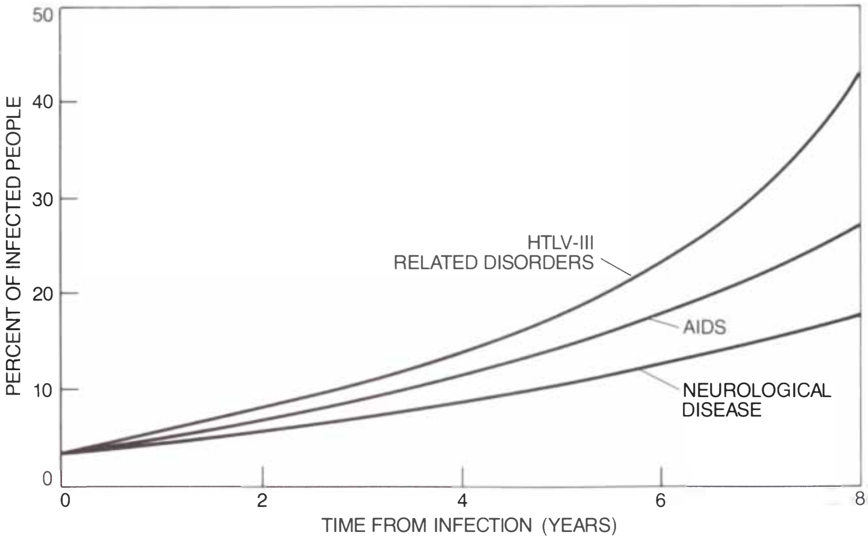
How HTLV-III manages to replicate in a single burst after lying low, sometimes for years, is one of the most fundamental questions confronting AIDS researchers. Another important question is the full spectrum of diseases with which the virus is associated. Although most of the attention given to the virus has gone to AIDS, HTLV-III is also associated with brain disease and several types of cancer. In spite of such lingering questions, more is known about the AIDS virus than is known about any other retrovirus. The rapidity of that scientific advance was made possible partly by the discovery in 1978 of the first human retrovirus, HTLV-I, which causes leukemia. In its turn the new knowledge is making possible the measures that are desperately needed to treat AIDS and prevent its spread.

DEGENERATING *T*4 CELL gives rise to a mass of newly made particles of the virus that causes AIDS: human *T*-lymphotropic virus III (HTLV-III), also called human immunodeficiency virus (HIV). The cell (magnified some 15,000 diameters) is the irregular treelike shape; the virus particles are the small, black specks. HTLV-III is a retrovirus: when it enters a cell, the RNA it carries as its genetic material forms the template for assembling a set of DNA genes. The DNA inserts itself among the cell's chromosomes, where it can remain latent until it is activated to make new virus particles. In the case of HTLV-III many particles are made at once, and the burst of replication may kill the host cell. Because the virus's chief host—the *T*4 lymphocyte—is a white blood cell that regulates the immune response, HTLV-III infection can cause profound immune deficiency.

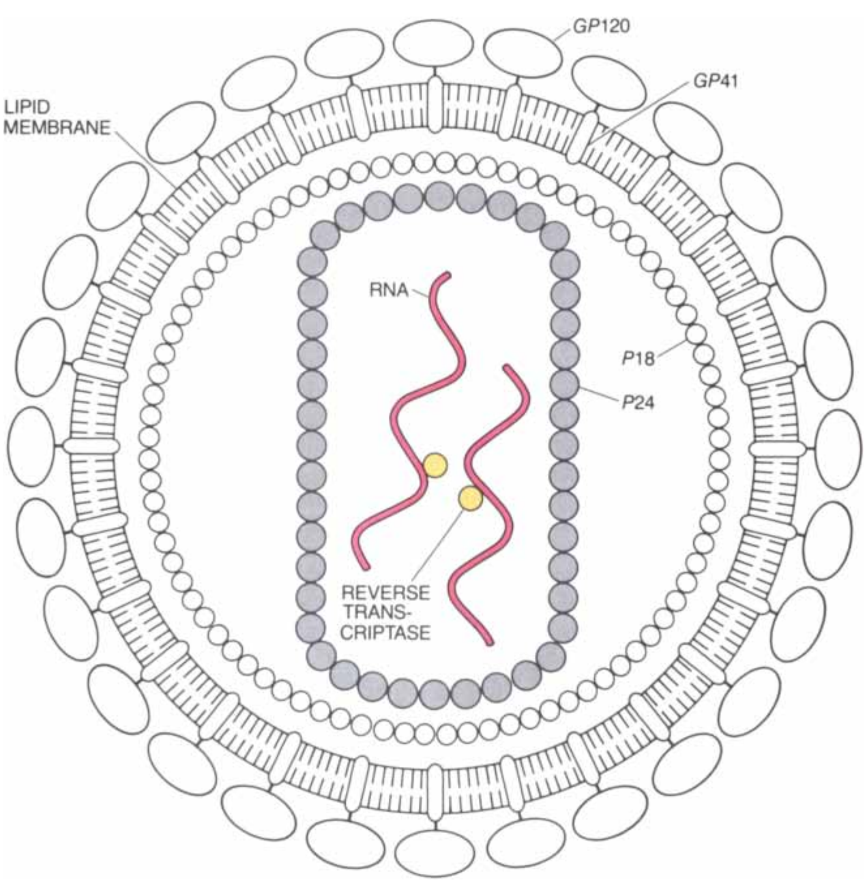
The first sign that a new disease was afoot was the appearance of a rare cancer called Kaposi's sarcoma among the "wrong" patients. Kaposi's sarcoma is a tumor of blood-vessel tissue in the skin or internal organs that had been known mainly among older Italian and Jewish men and in Africa. In the late 1970's, however, a more aggressive form of the same cancer began to appear among young white middle-class males, a group in which it had been extremely rare. Many of the new Kaposi's sarcoma patients turned out to have a history of homosexuality, and these young men provided the basis for the first reports of a new syndrome, which came in 1981 from Michael S. Gottlieb of the University of California at Los Angeles School of Medicine, Frederick P. Siegal of the Mount Sinai Medical Center and Henry Masur of New York Hospital.

Seen mainly among young homosexual men, the new syndrome included opportunistic infections and a depletion of *T*4 cells as well as, in some cases, Kaposi's sarcoma. Soon epidemiologists at the U.S. Centers for Disease Control (CDC) noted a dramatic increase in pneumonia caused by *Pneumocystis carinii*, a widespread but generally harmless protozoan. It seemed clear that an infectious form of immune deficiency was on the rise, and the name AIDS was coined to describe it. AIDS was quickly found to be spreading among users of intravenous drugs, recipients of frequent blood transfusions and Haitians. A mysterious and fatal illness, apparently associated with life-style, had appeared.

Hypotheses about the cause of AIDS proliferated rapidly. It was suggested that the disease resulted from exposure to sperm or to amyl nitrate, a



SPECTRUM OF DISEASES due to HTLV-III includes neurological disorders and cancer as well as immune deficiency. The curves show a hypothetical model for the proportions of infected people who might ultimately develop such illnesses. The neurological syndromes appear to be caused directly by HTLV-III infection of the brain, independent of secondary infections due to immune deficiency. The cancers are among a range of HTLV-III-related illnesses that may or may not be dependent on the immune deficiency.



HTLV-III VIRION, or virus particle, is a sphere that is roughly 1,000 angstrom units (one ten-thousandth of a millimeter) across. The particle is covered by a membrane, made up of two layers of lipid (fatty) material, that is derived from the outer membrane of the host cell. Studding the membrane are glycoproteins (proteins with sugar chains attached). Each glycoprotein has two components: *gp*41 spans the membrane and *gp*120 extends beyond it. The membrane-and-protein envelope covers a core made up of proteins designated *p*24 and *p*18. The viral RNA is carried in the core, along with several copies of the enzyme reverse transcriptase, which catalyzes the assembly of the viral DNA.

stimulant used by some homosexuals. It was even proposed that AIDS had no specific etiologic agent: the patients' immune systems had simply broken down under chronic overexposure to foreign proteins carried by other people's white blood cells or by infectious agents. Yet it seemed more plausible to think of a single cause, and several workers suggested known viruses such as Epstein-Barr virus or cytomegalovirus, which are members of the herpes virus family. Both were long-established viruses, however, whereas AIDS seemed to be a new disease. Moreover, neither virus has an affinity for *T* cells.

James W. Curran of the CDC and his colleagues, who had been following the nascent epidemic, clearly favored the notion of a new infectious agent. In late 1981, as I listened to Curran outline what was known about the epidemiology of AIDS, I was already in agreement with him. A clue as to what the new agent might be came from the fact that some hemophiliacs had developed AIDS after receiving infusions of a preparation called Factor VIII, prepared from the plasma of many blood donors. In preparing Factor VIII the plasma is passed through filters fine enough to remove fungi and bacteria—but not viruses.

That observation supported those who had argued in favor of a virus. Yet if one could not look to established viruses as the cause, how could the culprit be identified? Any virus that was a candidate would have to fit what was known about the agent, which included the following. It was present in whole blood, plasma and semen as well as in Factor VIII. The epidemiological pattern showed that it could be transmitted by sexual contact, blood and congenital infection. Infection led, directly or indirectly, to the loss of *T*4 cells.

As it happened, that pattern was familiar to me and my co-workers, because HTLV-I had been isolated in my laboratory in 1978. (Its story is told in the first part of this article: "The First Human Retrovirus," in last month's *Scientific American*.) HTLV-I can be transmitted by blood, intimate contact and congenital infection; it has a strong affinity for *T* cells. Furthermore, although the chief effect of HTLV-I is leukemia, the virus can also cause a mild immune deficiency in some patients. Accordingly, in the spring of 1982 I proposed that the cause of AIDS was likely to be a new human retrovirus.

To refine and test the retrovirus hypothesis I assembled a small working group of scientists, each chosen for a specific expertise. Along with clinicians, epidemiologists, immunologists and molecular biologists were investigators experienced in animal retrovirology. One of the retrovirologists, Myron Essex of the Harvard Medical School, had published results lending support to the idea that a human retrovirus might cause AIDS. Essex had shown that a retrovirus called feline leukemia virus (FeLV) can cause either leukemia or immune deficiency in cats. A minor variation in the virus's outer envelope, it was later shown, determines whether infection leads to immune suppression or to cancer.

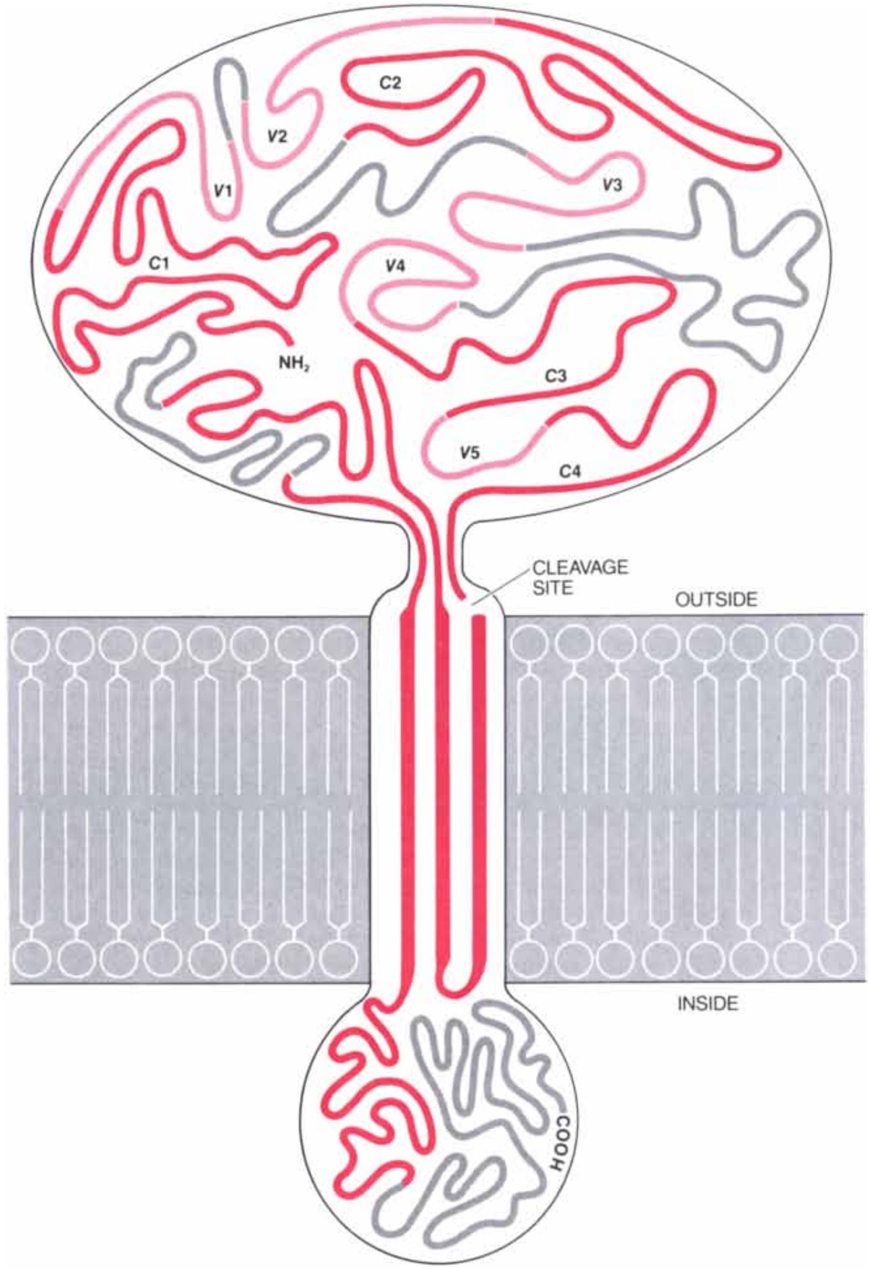
These suggestive results made it seem even more plausible that a variant of HTLV-I (or its near relative HTLV-II, isolated in 1982) might be the AIDS agent. Essex's group and my own quickly began searching for such a virus. Soon we were joined by a third group, led by Luc Montagnier of the Pasteur Institute, who had been stimulated by the retrovirus hypothesis. All three groups employed the methods that my colleagues and I had developed for isolating HTLV-I: the virus was cultured in *T* cells stimulated by the growth factor called IL-2 and its presence was detected by sensitive assays for the viral reverse transcriptase.

Those methods quickly produced results. Beginning in late 1982 and continuing throughout 1983 my coworkers and I found preliminary evidence of retroviruses different from HTLV-I or II in tissues from people with AIDS or pre-AIDS conditions. Then in May of 1983 Montagnier and his colleagues Françoise Barré-Sinoussi and Jean-Claude Chermann published the first report of a new retrovirus from a patient with the lymphadenopathy ("swollen glands") typical of some pre-AIDS cases. The French investigators later gave their find the name lymphadenopathy-associated virus (LAV).

The initial report of LAV was intriguing, but it was hardly a conclusive identification of the cause of AIDS. The reason is that the methods then available (reverse-transcriptase assays accompanied by electron microscopy) can show that a retrovirus is present in a tissue sample but cannot specify the precise type of virus. Unique identification is possible only if reagents (such as antibodies) are available that react with the proteins of that virus and no other. Making such reagents requires large quantities of purified viral proteins; to obtain them the virus must be grown in the laboratory.

The new virus (or viruses), however, resisted the early attempts at laboratory culture: when they were put in *T* cells, the cells died. Hence no specific reagents to the new isolates could be made. We had previously learned how to culture HTLV-I and II, and reagents to those viruses were available. As a result it was possible to show that the viruses present in AIDS patients were not HTLV-I or HTLV-II, but throughout most of 1983 it was not possible to make a positive identification because specific reagents were lacking. Moreover, in the absence of reagents one could not say that any two of the new isolates were the same, which was clearly a requirement for showing that AIDS has a single cause.

The answer to such difficulties was



ENVELOPE GLYCOPROTEIN has an important role in HTLV-III's entry to its host and also in the death of the host cell. The protein includes some regions that are constant from one viral strain to another (*dark color*), some that are highly variable (*light color*) and others that are intermediate (*gray*). Entry to the cell seems to depend on an interaction between one or more of the constant regions and molecules in the cell membrane. The envelope protein is also involved in the budding of a new virus particle from the cell, which may leave a hole in the cell's surface. The model shown (one of several possible models of the protein's structure) was developed by Hans Wolf and his colleagues at the Max von Pettenkoffer Institute in Munich in collaboration with the author's group.

to find a way to grow the virus. In the fall of 1983 my colleague Mika Popovic identified several cell lines that could be infected with the virus but resisted being killed. To obtain them the blood cells of a person with leukemia were separated and allowed to proliferate into clones of genetically identical cells. Many clones were screened, and several were found to have the right combination of qualities; the most productive of them was the clone designated *H*9. All the resistant lines are made up of leukemic *T*4 cells that are immortal in culture and therefore an endless source of virus.

Why certain *T*4 cell lines should resist the cytopathic effects of the virus is a significant question that has not been answered. In the winter of 1983–84, however, my colleagues and I had little time for that puzzle because we were concentrating on growing the virus. By December substantial quantities were being grown, and soon afterward reagent production was under way. With reagents in hand, we could go back and identify the many stored viral isolates. Initial testing showed that 48 isolates from AIDS patients or members of risk groups were of the same type. In contrast, the virus so identified was not found in any members of a control group of 124 healthy heterosexuals.

Continuous production of the virus also yielded enough viral proteins to provide the basis of a blood test. (Although there are several methods of testing blood for the AIDS agent, all of them rely on the reaction between viral proteins and antibodies in the infected person's blood.) The first blood testing was done by my colleague M. G. Sarngadharan, working on serum identified only by a code. By means of such "blind" testing Sarngadharan found virus in the serum of from 88 to 100 percent of AIDS patients (depending on the study), in a high but varying proportion of people in risk groups and in almost no healthy individuals outside the risk groups. The cause of AIDS had been established.

My colleagues and I reported these results in a series of publications in May, 1984. The retrovirus we had identified showed an affinity for *T*4 cells and also killed those cells. In accord with the prevailing conventions of virus nomenclature, the isolates were given the generic name HTLV-III and individual strains were distinguished by the initials of the patient from which they had come. Later it was shown that LAV is a different strain of the same virus. Later still, the name HIV was coined by a committee set up to resolve the problems caused by the existence of multiple names for the same biologic object.

Demonstrating the cause of AIDS was a fundamental step. Perhaps equally important from the viewpoint of public health was the fact that growing the virus had provided the basis for a practical blood test. The infected *H*9 line was given to several biotechnology companies, which used it as a source of viral proteins for a commercial blood test. The commercial test, marketed in 1985, virtually eliminated the risk of contracting AIDS through blood tranfusion.

Although only three years have elapsed since the cause of AIDS was identified, much has been learned about how the virus gives rise to disease. When a person is first infected, his (or her) immune system does respond by making antibodies. That response is clearly not adequate, however, and the virus takes hold. In many cases lymphocytes then begin to proliferate abnormally in the lymph nodes. Thereafter the node's intricate structure collapses, and a decline in the number of lymphocytes in the node follows. Soon the number of lymphocytes in the blood also decreases, leaving the patient open to opportunistic infections [see "The Immune System in AIDS," by Jeffrey Laurence; *Scientific American,* December, 1985].

What events at the cellular level underlie this clinical catastrophe? It seems infection may be initiated by free virus or by virus carried in infected cells. Once the virus is inside the body its target consists of cells bearing the *T*4 molecule in their outer membrane. That molecule defines the category of *T*4 lymphocytes, but it is also found on cells called monocytes and macrophages, and it appears that *T*4-carrying monocytes and macrophages are among the first targets of infection by the AIDS virus.

Monocytes and macrophages arise from the same bone-marrow precursors as lymphocytes, but they have different roles in the immune response. Among the roles of the macrophage are interactions with *T*4 lymphocytes that stimulate the *T*4 cells to undertake their tasks. Some of the interactions occur in the lymph node, and observations by Peter Biberfeld of the Karolinska Institute in Stockholm and Claudio Baroni of the University of Rome suggest that many *T*4 cells are infected in the lymph node during contact with a macrophage. After a variable latency the infected lymphocyte may be killed by viral replication.

Clearly the *T*4 population is reduced by the death of infected cells. The effect is compounded by the fact that the killing halts the normal proliferation of the lymphocytes that accompanies their immune functions. In the interaction with a macrophage the *T*4 cell not only is primed to respond to a particular protein but also is activated. Growth factors secreted by the macrophage cause it to begin a process of cell division that ultimately yields a clone of perhaps 1,000 descendants, all programmed to respond to the same antigen (protein). The descendants circulate in the blood and, on encountering the antigen they are programmed for, they induce the maturation of cells called *B* lymphocytes and *T*8 cytotoxic cells that attack pathogens directly. In this way the "memory clone" provides part of the basis of lasting immunity.

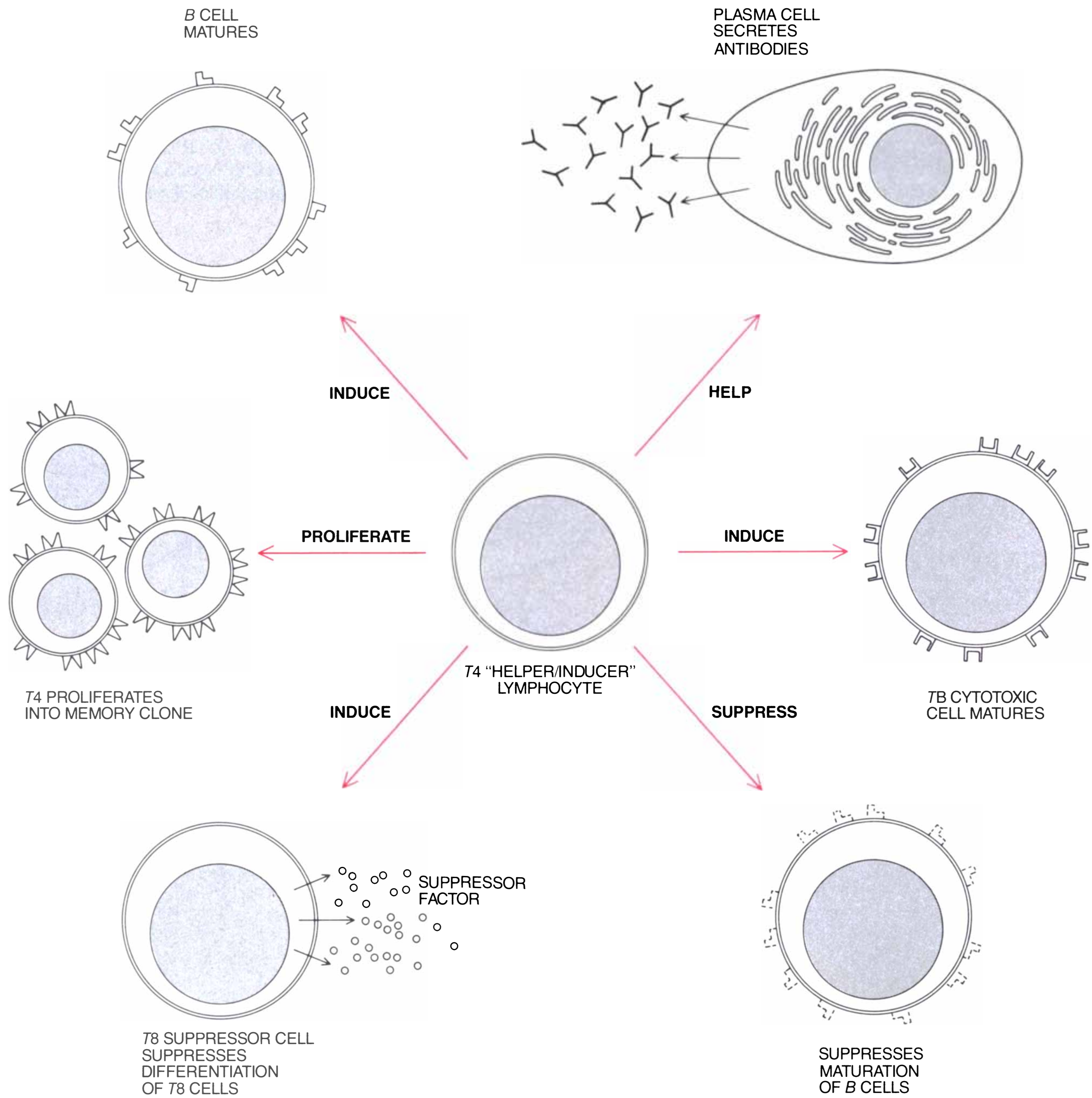
When a *T*4 cell infected with the AIDS virus is activated, however, the result is quite different, as Daniel Zagury of the University of Paris has shown in collaboration with me. Instead of yielding 1,000 progeny, the infected *T* cell proliferates into a stunted clone with perhaps as few as 10 members. When those 10 reach the bloodstream and are stimulated by antigen, they begin producing virus and die. Other suggestions have been made, but I think the direct killing of infected lymphocytes and the abortive expansion of the memory clones are largely responsible for the profound depletion of *T*4 cells observed in AIDS.

And what underlies these cellular events at the level of molecules? One of the most significant molecules in HTLV-III infection is *T*4. Indeed, by interacting with the outer envelope of the virus, *T*4 may provide entry to the cell. The viral envelope consists of a membrane studded with glycoprotein molecules (proteins with attached sugar chains). Each glycoprotein has two subunits, called *gp*41 and *gp*120. When HTLV-III makes contact with a cell, *gp*120 appears to interact with a *T*4 molecule in the cell's outer membrane. Thereafter the cell's membrane may form a vesicle that draws the virus into the cell. (This process, which is known as receptor-mediated endocytosis, provides entry to the cell for a variety of molecules needed for normal metabolism.)

The by now overwhelming evidence that *T*4 is involved in infection was gathered in steps. The first step was the clinical observation that the infected cells are *T*4 lymphocytes. Next, Robin A. Weiss of the Chester Beatty Laboratory in London, Angus Dalgleish of University College Hospital Medical

School and David Klatzmann of the Hôpital Salpêtrière found that antibodies to the *T*4 molecule, which cover part of its structure, block HTLV-III infection. Then my colleagues and I found infected *T*4-carrying monocytes and macrophages. Some of the most telling evidence, however, came from Weiss and Richard Axel of the Columbia University College of Physicians and Surgeons, who inserted the *T*4 gene into cells that do not ordinarily carry the marker and are not infected. Expression of the gene, entailing synthesis of the marker and its insertion into the cell membrane, is sufficient for infection of any human cell.

A similar, if less formidable, combination of evidence implicates *T*4 in cell death. There too the initial observation was a clinical one: that the *T*4-lymphocyte population was depleted. In contrast to infection, however, the presence of *T*4 alone is not sufficient for cell killing. Although monocytes and macrophages can be infected with HTLV-III, they are not easily killed, and the reason may be that they have few *T*4 molecules on their surface. It seems that although a low level of *T*4



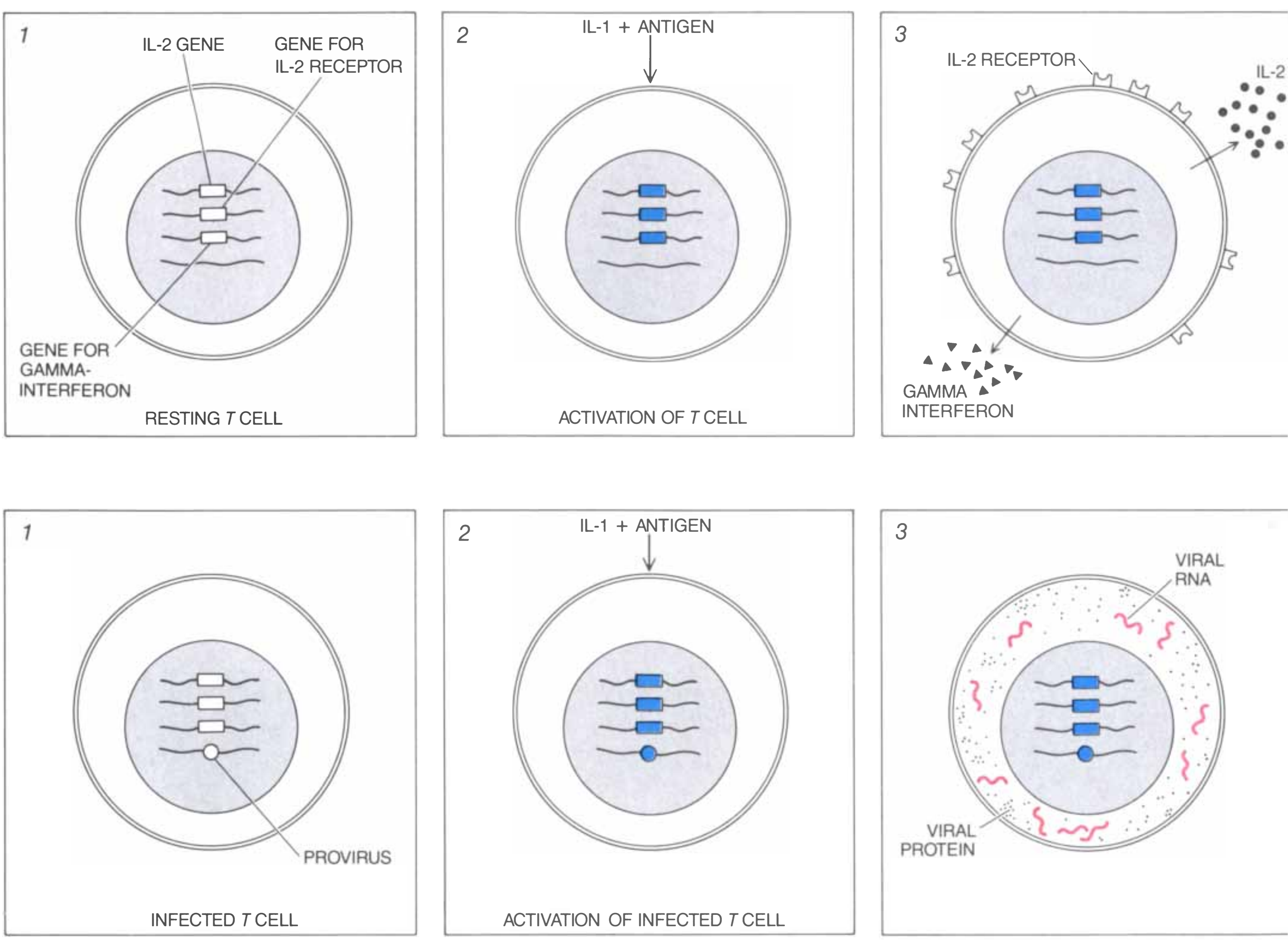
ROLES OF THE *T*4 CELL in the immune response include interactions that prepare other cells to attack invading organisms. The *T*4 cell produces substances that stimulate the maturation of the other main class of lymphocytes: the *B* cells. When a *B* cell matures, it differentiates into a plasma cell that secretes antibodies; *T*4 cells may help them in that task. Other signals given off by the *T*4 cell trigger the maturation of a second subset of *T* cells, the *T*8 cells, that attack and kill cells infected by pathogens. When an infection has been brought under control, the *T*4 cell has a role in suppressing further maturation of *B* and *T*8 cells. As a final safeguard the *T*4 cell proliferates into a clone of memory cells that circulate in the blood, ready to recognize a specific pathogen and carry out their multiple roles. As a result of these functions the *T*4 cell is often referred to as the "helper/inducer" lymphocyte.

is sufficient for entry, a higher level may be required for the cytopathic effect. Indeed, as William Haseltine of the Dana Farber Cancer Institute has suggested, the rate of cell killing may be proportional to the concentration of *T*4 in the surface membrane of the infected cell.

Although no one knows why the death of the *T*4 cell should depend on the molecule that defines it, some suggestive findings make it possible to formulate a hypothesis. The killing depends not only on the *T*4 molecule but also on the viral envelope. My colleagues Flossie Wong-Staal and Amanda G. Fisher showed that mutant viruses lacking a piece of the inner end of *gp*41 have a drastically reduced cytopathic effect. Thus, like entry to the cell, its death may depend on an interaction between the viral envelope and the cell membrane. Perhaps that interaction (which takes place as the virus particle buds from the cell) punches a hole in the membrane. Because the virus buds in a mass of particles, the cell cannot repair the holes as fast as they are made; its contents leak out and it dies.

If that model (for the moment only a model) is correct, attention naturally falls on the question of how the virus is able to raise its rate of replication very rapidly from zero to a level high enough to kill the host cell. That question in turn focuses attention on the viral genome (the full complement of genetic information). The genome's DNA form, transcribed from the viral RNA and integrated into the cell's chromosomes, is called the provirus. The provirus includes the genes for the components of the virus particle, and in order for the virus to replicate, those genes must be expressed. How is gene expression controlled?

The answer (still very much under investigation) appears to lie in a group of regulatory genes whose presence renders the HTLV-III genome more complex than that of any other known retrovirus. The genetic complement of many retroviruses consists chiefly of the three genes that encode the components of the virus particle: *env* (which codes for the envelope proteins), *gag* (for the RNA-containing core) and *pol* (for the reverse transcriptase). Those three genes are flanked by stretches of DNA called the long terminal redundancies, or LTR's. The LTR's include DNA sequences that have a role in controlling the expression of the viral genes.



STUNTED MEMORY CLONE results from HTLV-III infection. The five upper panels depict the formation of a normal clone of memory *T*4 cells. Prior to infection the *T* cell is in a resting state (*1*). During infection a cell known as a macrophage secretes a protein called IL-1 and presents an antigen (a protein from the invading organism) to the *T*4 cell. The *T* cell is thereby immunologically activated. Several of its genes are turned on, including those for the growth factor IL-2 and its receptor (*2*). The activated cell secretes IL-2, and receptors for the protein appear on its surface (*3*). The binding of IL-2 to the receptors (*4*) initiates a process of proliferation that culminates in a memory clone with perhaps 1,000 members, each primed to react to the antigen with which

The genome of HTLV-III, however, includes at least four other genes, called *tat, trs, sor* and 3'*orf*. Rather than encoding viral components, the four additional genes encode small proteins that help to regulate gene expression. The *tat* gene (discovered by Haseltine and his colleague Joseph Sodroski, and independently by Wong-Staal and Suresh Arya in my laboratory) has a dual function. Like its analogues in HTLV-I and II, *tat* appears to regulate the transcription of messenger RNA (mRNA) from the viral genes. In addition the *tat* protein affects events after transcription, perhaps the translation of the viral mRNA into proteins. The *trs* gene (discovered by Haseltine) appears to control the balance among the various forms of viral mRNA. The functions of *sor* and 3'*orf* are unknown.

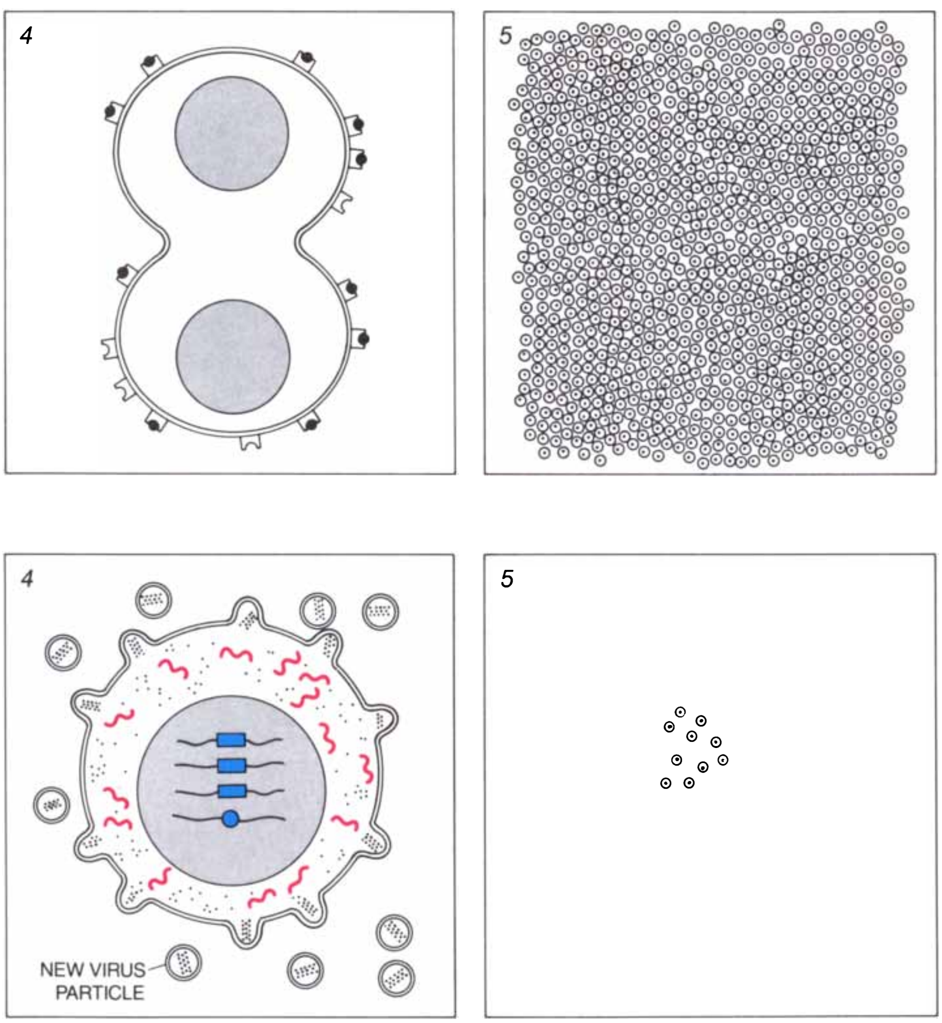
There are many other unknowns in this complex system, and it is too soon to say confidently how it works. It is not too soon, however, to hazard a general hypothesis, taking as a premise the fact that the virus does not replicate until the *T* cell is immunologically activated. The LTR's of the AIDS virus share some DNA sequences with the cellular genes that are turned on during immune activation. I think the chemical signals that activate the *T*4 cell simultaneously activate the viral LTR's. Somehow the small regulatory proteins interact with the provirus to boost synthesis of the viral components very quickly. The components self-assemble and bud from the cell in a pulse that may kill the host.

In summary form that is what is known about how HTLV-III cripples the immune system. Although most of the attention given to the virus has been devoted to that process, it has become increasingly clear that immune deficiency is only one effect of the AIDS agent. The other main type of disease caused by HTLV-III is seen in the central nervous system. HTLV-III was first detected in brain and spinal-cord tissues from AIDS patients by my colleagues George M. Shaw, Beatrice Hahn, Wong-Staal and me in 1984. The infected cells appear to have some of the properties of monocytes and macrophages. Those cells may be able to cross the blood-brain barrier, which separates the central nervous system from the blood supply; perhaps macrophages become infected in the blood and transport the virus from there to the brain.

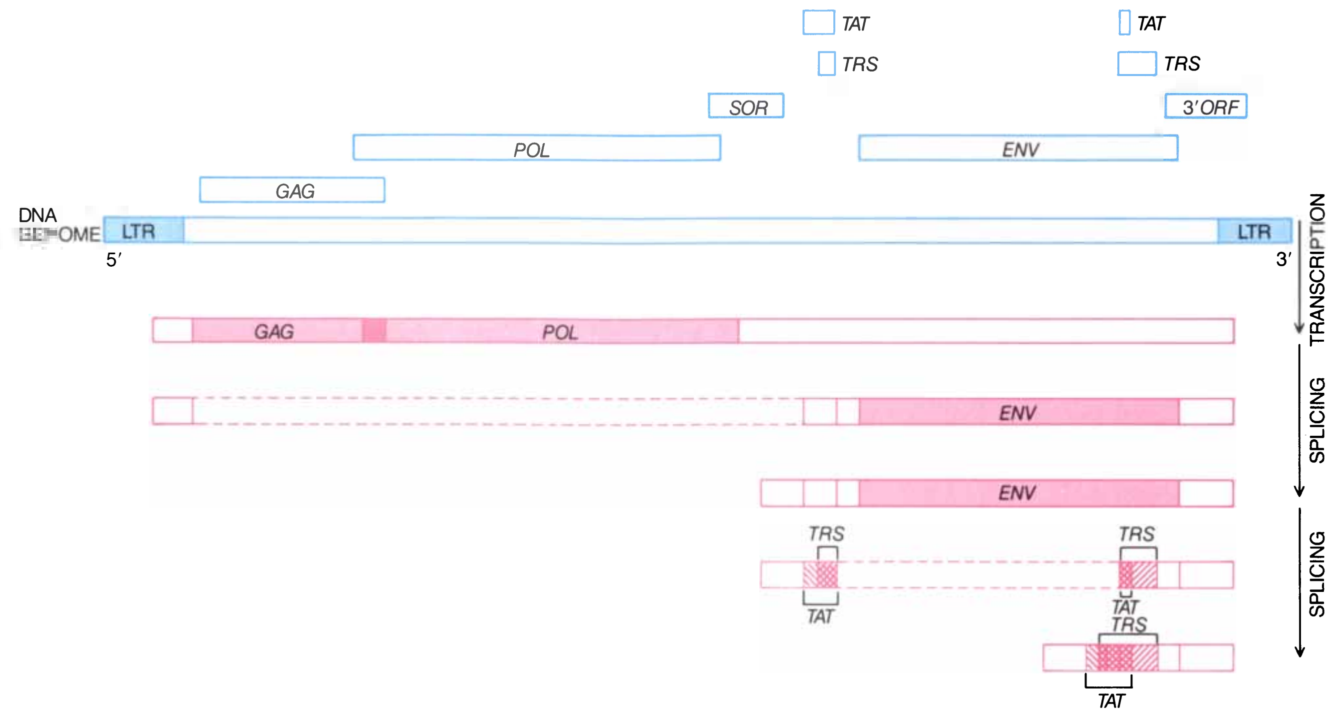
In the brain and spinal cord the virus appears to have a direct pathogenic effect that is not dependent on the immune deficiency. The chief pathologies observed in the brain are an abnormal proliferation of the glial cells that surround the neurons and lesions resulting from loss of white matter (which is, along with gray matter, one of the two main types of brain tissue). How the virus causes such anatomical changes is not understood. Nor is it known how the relatively limited range of structural aberrations due to the virus gives rise to a wide range of symptoms including dementia and mimicry of other neurological syndromes such as multiple sclerosis.

Whereas the neurological effects of HTLV-III are distinct from immune deficiency, the third main type of pathology—cancer—has a more ambiguous relation to the crippling of the immune system. People infected with the virus have an increased risk of at least three types of human tumor: Kaposi's sarcoma, carcinomas (including skin cancers often seen in the mouth or rectum of infected homosexuals) and *B*-cell lymphomas, which are tumors originating in *B* lymphocytes.

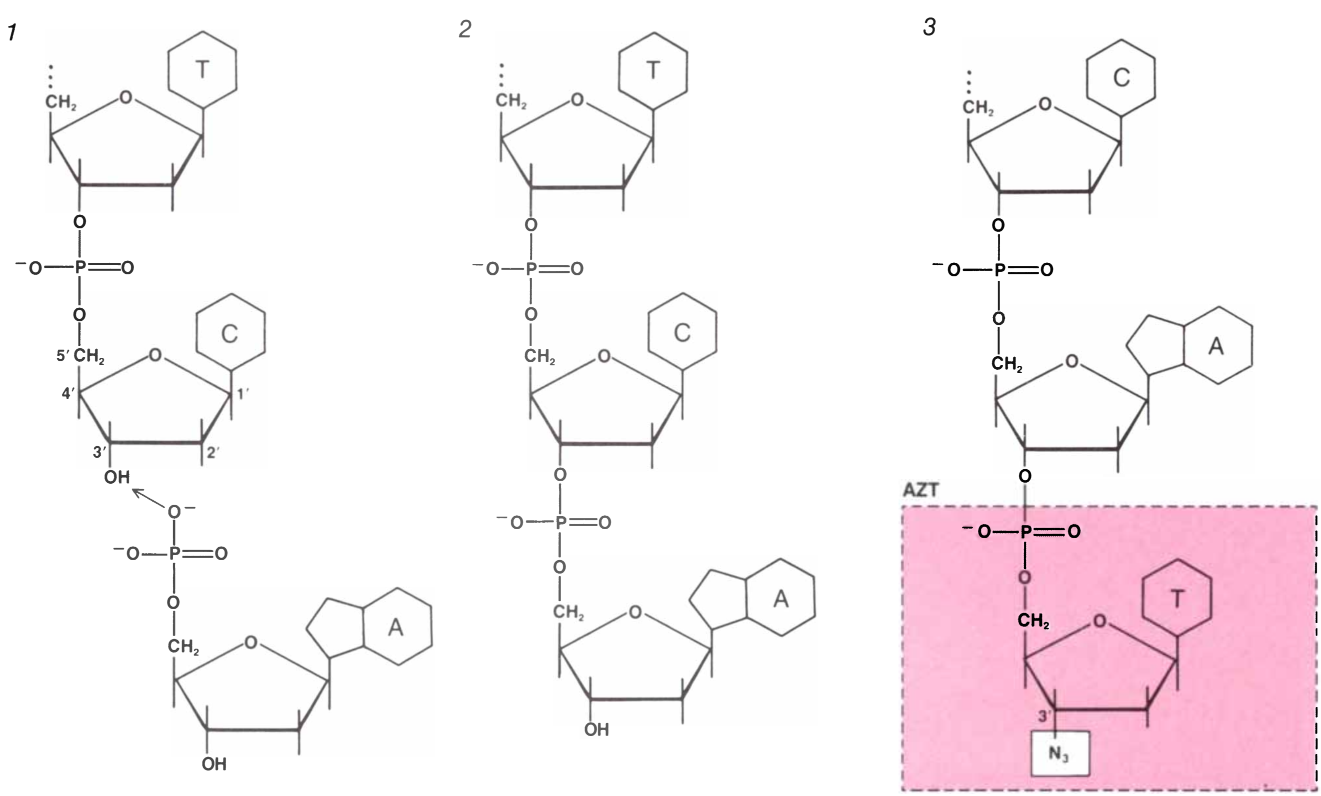
In some instances the tumors appear to be independent of immune deficits, as is suggested by the fact that homosexuals may have an increased risk of developing Kaposi's sarcoma even if they are not infected with the AIDS virus. Pathogens other than HTLV-III—perhaps sexually transmitted agents—are likely to be involved in these tumors. Yet infection with HTLV-III greatly increases the risk that Kaposi's sarcoma will develop. Therefore it seems likely that depression of the immune response enables secondary tumor-causing agents to infect and replicate freely. What they are is not known, but one may be human *B*-lymphotropic virus (HBLV), a new DNA-



the process began (*5*). The lower panels show what happens when an infected cell is activated. The viral DNA in the cell's chromosome is called the provirus (*1*). In the interaction with the macrophage (*2*) the provirus is activated along with the cellular genes. Viral RNA and proteins are made (*3*). They self-assemble into particles that leave the cell, often killing it (*4*). Hence the memory clone may have only 10 cells (*5*). This model may explain the loss of *T*-cell-based immunity in people infected with HTLV-III.



GENOME OF HTLV-III (its full complement of genetic information) is more complex than that of any other known retrovirus. At each end of the provirus (*blue*) are DNA sequences called long terminal redundancies (LTR's), which have a regulatory role. Between the LTR's are at least seven genes. Three code for viral components: *gag* for the core proteins, *pol* for reverse transcriptase and *env* for the envelope proteins. Four unusual genes—*tat, trs, sor* and *3′orf*—encode proteins that help to control the expression of viral genes. The RNA made from the provirus (*red*) is spliced to yield the array of messenger RNA's (mRNA's) from which viral proteins are assembled. The core proteins and reverse transcriptase are made from an mRNA corresponding to the entire genome. One splice yields the envelope-protein mRNA, a second the small mRNA from which the *tat* and *trs* proteins are made.



DNA CHAIN IS TERMINATED by azidothymidine (AZT), the first drug to treat the symptoms of AIDS effectively. DNA consists of subunits called nucleotides, each of which includes a five-carbon sugar molecule. Normally a hydroxl group (OH) is present on the carbon designated 3′ (*1*). The hydroxl group provides the attachment point for the next nucleotide by taking part in the formation of a linkage called a phosphodiester bond (*2*). AZT is an analogue of the usual nucleotides. The viral reverse transcriptase will incorporate it into a DNA chain (*3*). Because AZT lacks the 3′ hydroxyl group, the chain is terminated, yielding an inactive provirus.

containing member of the herpes virus family recently isolated by my colleagues Zaki Salahuddin, Dharam Ablashi and Biberfeld and me.

The welter of pathologies caused by HTLV-III seems daunting, but the knowledge already gained about the virus has begun to lay the groundwork for treatment and prevention. The most promising therapies currently under investigation are based on interrupting the reverse transcriptase as it assembles the viral DNA destined to become the provirus. The drugs used for this purpose are chemical analogues of the nucleotides that form the subunits of DNA. When the analogue is supplied to an infected cell, reverse transcriptase will incorporate it into a growing DNA chain. Because the analogue lacks the correct attachment point for the next subunit, however, the chain is terminated. The truncated DNA cannot integrate itself into the chromosomes or provide the basis for viral replication, and so the spread of infection is halted.

Recent tests of azidothymidine, or AZT, have shown that this strategy can reduce mortality among AIDS and pre-AIDS patients as well as moderating their symptoms. AZT was formulated some 20 years ago as an anticancer drug. Although a failure in that role, it was resurrected in 1984 as a possible means of treating AIDS. After initial studies of the interaction of AZT and the viral reverse transcriptase, the drug was brought to the threshold of clinical use by Samuel Broder and Robert Yarchoan of the National Cancer Institute. Recently a multicenter trial was interrupted to begin wide distribution of AZT as a result of the dramatic benefits observed in the tests. It is not known, however, how toxic AZT may prove to be when it is used for a long period.

AFRICAN GREEN MONKEY (*Cercopithecus aethiops*) may have harbored the ancestor of the AIDS virus: simian *T*-lymphotropic virus III (STLV-III). STLV-III does not usually cause disease in its simian host, but it might have infected humans and given rise to HTLV-III after many genetic changes. The monkey virus was isolated in 1985 by Myron Essex and Phyllis J. Kanki of the Harvard Medical School. The photograph is by Kanki.

Perhaps the most important work now being done in the effort to curb AIDS is the development of a vaccine. For a vaccine to be effective it must safely evoke two different types of immunologic response. The *B* cells must be stimulated to produce neutralizing antibodies, which bind to the virus's envelope and prevent it from entering cells. In addition the cellular system anchored by the *T* cells must be capable of attacking and destroying cells already infected with the virus. Although, as I have mentioned, people infected with HTLV-III do make antibodies to the virus, the amount of effective, neutralizing antibody is worrisomely low, and of course cellular immunity is subverted by the death of the *T*4 cells. A successful vaccine must boost both reactions greatly.

That task is made considerably more complex by the virus's great genetic variability. Unlike many viruses, which have only a few strains, HTLV-III comprises a great many variants that form a continuum of related strains. Some pairs of variants differ by as few as 80 nucleotides of the 9,500 making up the viral genome; others differ by more than 1,000 nucleotides. Since the nucleotide sequence of the genome constitutes the genetic code for the viral proteins, such differences translate into variations in protein composition. Differences in proteins may in turn account for variations in biological activity seen among strains of HTLV-III, including preferences for infecting either *T*4 cells or macrophages.

Intriguingly, Wade P. Parks of the University of Miami (collaborating with Shaw and Hahn in my group) showed that an individual infected with HTLV-III may harbor several strains of the virus, all closely related in their genetic makeup. The fact that all the coexisting strains are closely related suggests that somehow their presence may "vaccinate" the infected person against reinfection by more distantly related strains. This pattern offers hope that a synthetic vaccine may be able to do the same. As yet, however, no manmade vaccine has been able to cope with the profusion of strains. My group and others are working on many approaches to a vaccine, some of which have yielded neutralizing antibodies. Yet so far the vaccines have been type-specific, neutralizing many but not all HTLV-III variants.

The progress made in only three years—identification of the cause of AIDS, formulation of a blood test, the first effective therapy and the beginning of vaccine development—is striking, particularly in view of the fact that AIDS is a viral illness, a type that has generally resisted effective therapy. Yet even if therapy and vaccine are brought into being on the fastest possible schedule, HTLV-III's toll will be heavy: many of the millions already infected will become ill before treatment is available.

The proportion of infected people who do go on to become ill may be considerably higher than was once thought. Along with Mark H. Kaplan of North Shore University Hospital on Long Island, Robert R. Redfield of the Walter Reed Army Institute of Research has led the way in developing clinical categories that go beyond classical AIDS to consider HTLV-III infection in its full context. Redfield has developed a six-stage system of classification beginning with a positive blood test and ending with full-blown AIDS. Recently he used that system to follow a group of patients for as long as 36 months and found that about 90 percent of them progressed from the stage in which they began the study to a subsequent stage. Such results suggest that, contrary to what has been suggested, there may not be a large group of infected people who remain without symptoms.

It is difficult to say what the final toll will be. Regardless of its size, however, much of it will be felt in Africa. In some African countries epidemiological results show that a sizable fraction of people in the sexually active age groups are already infected. The high prevalence of infection in Africa is due partly to the fact that universal testing of the blood supply is beyond the economic reach of most African countries. As a result the virus is still being transmitted by contaminated blood. In addition, it appears that the virus has had more time to spread in Africa than it has had in any other part of the world.

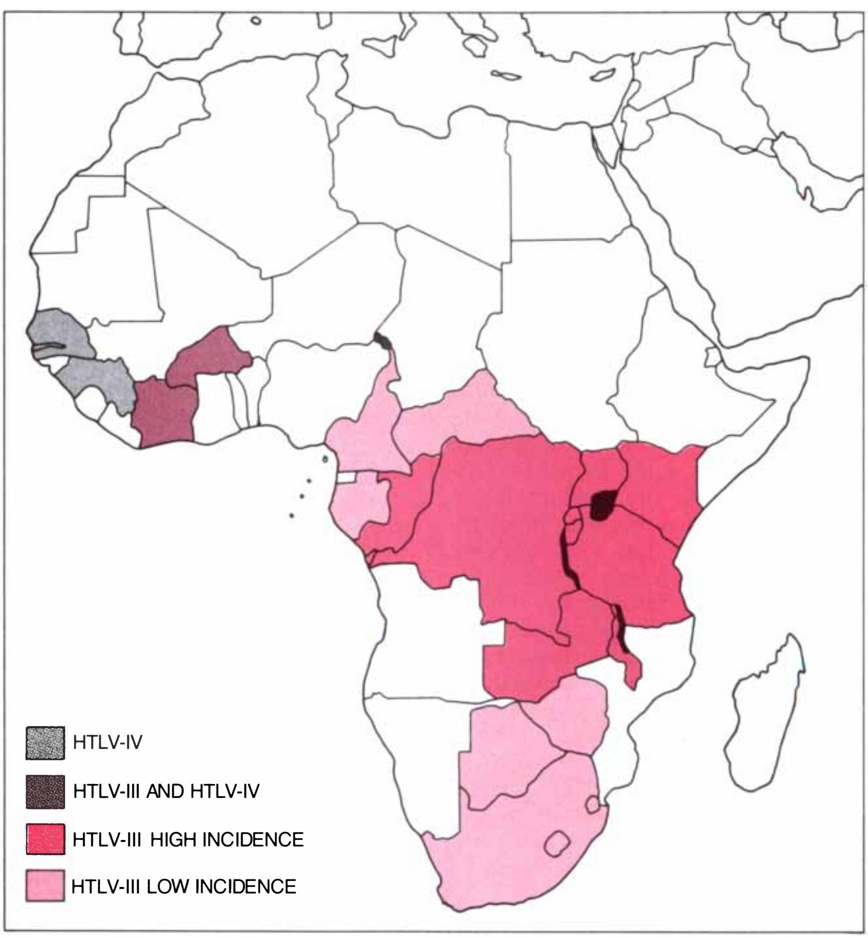
Recent results have begun to provide a picture of how the AIDS virus may have come to be. In 1985 Essex and his colleague Phyllis J. Kanki isolated a virus related to HTLV-III in African green monkeys, whose range includes much of equatorial Africa. The monkey virus, which is called simian *T*-lymphotropic virus III (STLV-III), may well be an ancestor of the AIDS agent. Yet although STLV-III is a closer relative of HTLV-III than any other animal retrovirus is, the relation between them is still not particularly close. Nor is the monkey virus pathogenic in its usual host.

Recently, however, the gap between the simian virus and the human one has begun to be filled by the discovery of a group of intermediate viruses. The first of these, called HTLV-IV, is closely related to STLV-III and is nonpathogenic, but it infects humans. It was isolated by Essex and Kanki in West Africa in 1985. More recently two viruses that are closely related to HTLV-IV but do cause immune deficiency have been discovered in the same region. Called LAV-2 and the SBL virus, they were isolated by the Pasteur group and by a Swedish group respectively. A plausible hypothesis is that STLV-III somehow entered human beings, initiating a series of mutations that yielded the intermediate viruses before terminating in the fierce pathology of HTLV-III.

That those fierce effects are of recent origin is shown by tests done on stored blood serum from many parts of the world. Tests on sera from the 1960's and 1970's detect no antibodies to HTLV-III anywhere except in a small region of central Africa, where the earliest signs of infection have been found in serum samples taken in the 1950's. It appears that after remaining localized for some time, the virus began spreading to the rest of central Africa during the early 1970's. Later in that decade it reached Haiti and may have reached Europe and the Americas from there.

Analysis of the origin and spread of HTLV-III leads to a conclusion that cannot be sufficiently emphasized: AIDS is not a disease of homosexuals or drug addicts or indeed of any particular risk group. The virus is spread by intimate contact, and the form of contact seems to be less important than the contact itself. Rapid spread of the virus depends on the accumulation of a pool of infected people that is large enough for a few exposures to result in infection. The pool need not consist of homosexuals or drug addicts. In Africa the pool is made up of heterosexuals, and Redfield, Kaplan and others have demonstrated heterosexual transmission in the U.S. Until a reliable vaccine is developed, intelligent caution and an understanding of the virus are the best weapons against its spread.

Does this terrible tale have a moral? Yes. In the past two decades one of the fondest boasts of medical science has been the conquest of infectious disease, at least in the wealthy countries of the industrialized world. The advent of retroviruses with the capacity to cause extraordinarily complex and devastating disease has exposed that claim for what it was: hubris. Nature is never truly conquered. The human retroviruses and their intricate interrelation with the human cell are but one example of that fact. Indeed, perhaps conquest is the wrong metaphor to describe our relation to nature, which not only surrounds but in the deepest sense also constitutes our being.



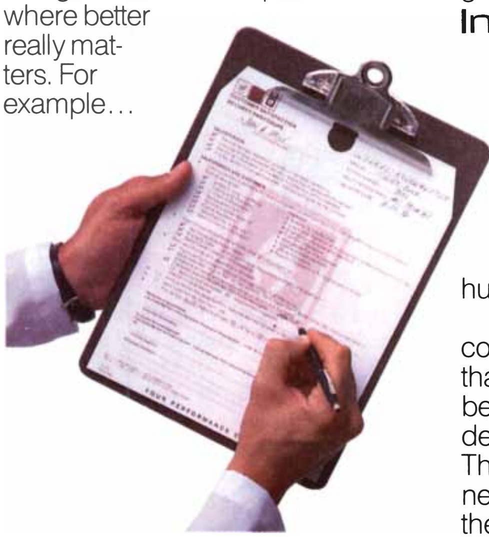
EPIDEMIOLOGICAL RESULTS indicate that HTLV-III and its relatives spread first in Africa. The map shows results of blood testing done in 1985 and early 1986. HTLV-IV, a nonpathogenic human virus closely related to STLV-III, is one of a group of viruses that may be intermediate between STLV-III and HTLV-III. The AIDS virus, still most prevalent in central Africa, has spread to the rest of Africa, Europe and both Americas.

We confidently invite you to give your brand new Buick its 201st inspection.

Buick Century

The first 113 quality inspections happen while your Buick is being built. But that's just a good beginning. Once it reaches your dealer, 87 additional items must be checked before it can be subjected to the most important inspection of all: yours.

And that's just one of the important things we pay attention to. Things that make us a place where better really matters. For example…



Underneath

Unfortunately, the road of life has its share of potholes, washboards and bumps. Fortunately, it also has an occasional sweeping curve, long empty stretch or twisting, tree-lined lane. Century is prepared to deal with it, either way.

Its independent front suspension is designed for comfort plus control. Computer-selected springs and shock absorber valving deliver the kind of smooth, effortless ride for which Buick has long been admired, while taking very good care of business when the going gets brisk and sporting.

Inside

When you sit behind the wheel of a Century, you find that the controls are where you want them for ease of operation. That's because of our dedication to ergonomics — the science of matching machine design to human design.

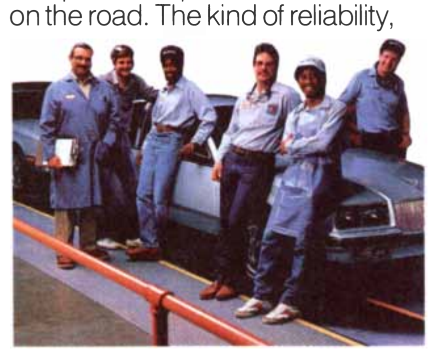
Other reasons for Century's comfort include things like seats that provide real support without being hard or harsh. Aerodynamic design to cut down on wind noise. The fact is that smooth ride, quietness and a sense of control result in the kind of enjoyable driving experience Buick is known for.

Satisfaction

Century is the best-selling Buick in history. The reasons are many. But one of them surely is the way this automobile manages to blend today's technology with traditional grace and comfort.

It is a front-wheel-drive car, with an engine controlled by microchips. It was designed with the aid of computers and wind tunnels for superb aerodynamics. It represents what can be done with the application of the latest technology.

But all this electronic wizardry is there to fulfill one simple, single function: to provide satisfaction. A car that starts when you want it to, has the kind of power you expect and provides quiet, solid comfort on the road. The kind of reliability,



quality and drivability that Buick and its dealers are committed to offering with every automobile they sell.

If that sounds like the automobile you've been looking for, buckle up and see your Buick dealer.

For more information, a brochure or a test drive, call 1-800-87-BUICK (1-800-872-8425).

GM

Where better really matters.

B U I C K

The AIDS Pandemic

In a moral sense any disease that affects some people affects all people. Yet it has been possible for mainstream members of industrial societies to feel safely (even smugly) distant from the groups that have been the main victims of AIDS: male homosexuals, intravenous-drug abusers, Haitians, some black Africans and hemophiliacs. That security blanket has now been stripped away. Three major health authorities assert that AIDS is everyone's problem. Late last year the U.S. Surgeon General urged that the public be given specific information about AIDS and how to avoid contracting it. Soon after, the Institute of Medicine and the National Academy of Sciences (IM/NAS) published a report that criticized U.S. Government efforts to combat AIDS as "woefully inadequate" and recommended an eightfold increase in funding for research and education over the next four years. The report was the work of a panel chaired by Nobel laureate David Baltimore of the Whitehead Institute for Biomedical Research and the Massachusetts Institute of Technology, and Sheldon M. Wolff of the Tufts University School of Medicine. Finally the United Nations' World Health Organization raised the disease to the level of an international crisis. Halfdan Mahler, head of the WHO, dubbed AIDS "a health disaster of pandemic proportions" and called for a massive international program to combat it.

Members of the AIDS research community have almost unanimously welcomed the call to arms. Indeed, some investigators suggest that the NAS report could have taken an even stronger stand on the need for more Federal funding. "We took that position two years ago," says Myron Essex of the Harvard Medical School, "and there's reason to be more concerned now."

Certainly the threat of AIDS should not be underestimated, but neither, other leaders in health care believe, should the public underestimate what has been accomplished so far. "Polio, smallpox, tuberculosis—look back on all the scourges of life," says Robert E. Windom, assistant secretary for health in the U.S. Department of Health and Human Services. "None of them has been handled as quickly and efficiently as this one. I don't think people have any appreciation of that."

Windom has a point. Ironically, AIDS may be a disease that has arrived at the right time. Ten years ago biologists had not even confirmed the existence of human retroviruses, and the physiology of the immune system was far from clear. Investigators would have been able to do little more than track the progress of the disease. Yet as Harold W. Jaffe of the U.S. Centers for Disease Control (CDC) notes, in the six years since AIDS was first described workers have isolated a suspect retrovirus, cultured it in the laboratory and identified it as the cause of the disease; they have devised ways to detect AIDS-invoked antibodies, thereby drastically reducing the likelihood that blood transfusions will cause infection; they have found some potential treatments, such as azidothymidine (AZT), and they have begun the formidable task of developing a vaccine.

The social as well as the medical value of this progress is immeasurable. Without such reassuring achievements the AIDS outbreak might conceivably have triggered a severe degree of publicly sanctioned discrimination and widespread physical violence. Even in the relatively rational atmosphere that currently prevails, a proposal that could have led to a quarantine of AIDS victims found its way onto the ballot in last November's elections in California. (The proposal failed to pass.)

Given the success of investigators so far, then, should the Federal budget for AIDS research grow to $1 billion a year by 1991, as the IM/NAS panel suggested? "A lot would be wasted," suggests Robert C. Gallo, who established HTLV-III as the cause of AIDS (see page 46). "I think we could do with more money," Gallo says. "But if the money is properly streamlined and properly organized we'll have less duplication of effort and fewer trivial subjects being pursued."

Asserting that to date "most of the results have come from a few people," Gallo suggests that the National Institutes of Health set aside a building at its headquarters in Bethesda, Md., as a temporary AIDS institute where people can easily share their results.

Others disagree that centralization is the answer. "There are a lot of negatives involved with a centralized authority," says Martin S. Hirsch, who is conducting trials of the promising drug AZT at the Harvard Medical School. "When Nixon declared a war on cancer, directed by a small coterie at NIH, a lot of imagination and creativity was stifled."

Hirsch agrees with Gallo that progress and large outlays of funds are not ipso facto synonymous. The U.S. Food and Drug Administration, for example, merely by accelerating its approval process, has ensured that trials of new treatments such as AZT can proceed rapidly. Moreover, Hirsch praises Surgeon General C. Everett Koop's recommendations on education as "most encouraging." Because Koop is generally seen as being politically conservative, Hirsch suggests, his recommendations may ultimately have greater impact than the more widely publicized NAS report.

Yet the coverage already devoted to AIDS by the media raises the question of how much more publicly funded education is needed. Research progress, epidemiology, community-action programs, death and morbidity rates—all have been front-page grist since 1981. The story has been covered not only in New York, San Francisco and Los Angeles, where the disease has been most prevalent, but also in heartland areas. Officials worry, however, that the information is not being absorbed by those who need it most: drug abusers and young people who are becoming sexually active.

In fact, for all its latent danger, AIDS is not yet as great a killer as some other infectious diseases. By the end of 1986 about 15,000 Americans had died of the disease; in comparison, influenza took 18,500 lives in 1957 alone. Wolff, cochair of the IM/NAS report, acknowledges that AIDS is potentially more easily avoided than the flu. "If people use condoms and don't shoot up drugs," he says, "we'll all be O.K." But Wolff insists that there haven't been any epidemics as potentially serious as AIDS in this century.

Part of the answer to the question of how great the threat of AIDS will be years from now lies embedded in the virus's RNA genome. The protean nature of the AIDS virus is probably the greatest source of concern to the research community. "The envelope of the virus seems to be changing," says Jeffrey Laurence, an immunologist at the Cornell University Medical Center. This mutability, manifested not just in the population as a whole but in individual patients, makes the development of a vaccine extremely difficult. Still another danger, says Laurence, is that the "new isolates may slip through the U.S. screening process" and infect blood supplies now considered safe. It is also possible that a more virulent strain could emerge.

There has been concern that the situation in Africa, where men and women are equally infected, may foreshadow what is to come in the U.S. Experts point out, however, that such factors as the prevalence of other venereal diseases that cause genital sores, the use of unsterilized needles in clinics and the lack of blood-screening tests may explain the different epidemiology of AIDS in Africa.

Yet the ability of the AIDS virus to lie dormant in cells for years makes it difficult to predict the national or global course of the epidemic, notes Anthony Fauci of the National Institute for Allergy and Infectious Diseases. For example, it is estimated that of the 10 million people who carry the virus worldwide (1.5 million of them in the U.S.), 25 to 30 percent will have the disease within five years. Fauci notes that "we don't know whether that figure will increase over time." And while acknowledging that breakthroughs in AIDS research have been reached "in record time," Fauci warns, "you never know when you will reach a brick wall."

Can SDI Survive House Arrest?

The Soviet Union has made it an unalterable condition for pursuing the dramatic arms reductions flirted with at Reykjavík that the U.S. agree to confine to the laboratory for 10 years the development and testing of space-based systems for the Strategic Defense Initiative (SDI). What would this limitation prevent and what would it permit? Could the U.S. keep the SDI under house arrest for a decade without stifling it?

The answer depends on what work the Soviets intend to limit, a point they have not clarified. Herbert F. York, former director of defense research and engineering in the Department of Defense, seeks an answer by posing the converse question: What anti-ballistic-missile (ABM) devices are the Soviets not likely to exclude from testing?

"Certainly the Soviets are not asking for a cessation of testing conventional, terminal ABM components because they themselves do this," says York. "For years we and they have also done experiments with various types of sensors in space. Surely the Soviets do not mean to limit those either." York concludes: "What the Soviets probably mean to exclude are experiments that employ destructive devices deployed on satellites."

If York is correct, then only a relatively small fraction of the SDI R&D effort is actually at stake. What the Soviets want to see restrained, says York, are "the 'sexier,' futuristic technologies, some of which wouldn't be out of the laboratory in less than 10 years anyway." An agreement to cut offensive weapons provided the SDI is scaled back more or less along the lines York surmises is a "perfectly acceptable limitation for the U.S."

Gerold Yonas, former chief scientist in the SDI Organization, disagrees. He finds the Soviet proposal, if its intent is literally to confine SDI development and testing within the walls of a laboratory building, "totally asymmetrical and inequitable." Yonas asserts: "It would kill the SDI program. Certain SDI experiments must be done in space. It is very hard to do them under a roof." An agreement that would allow testing of only those components in space that are not destructive weapons would be unverifiable, Yonas believes. "Finally," Yonas argues, "if the Soviets allow SDI to continue within the ABM Treaty, that's precisely what President Reagan has offered them."

Well, not precisely: some equivocation seems to blur the Administration's position. Last year Abraham D. Sofaer, chief legal counsel of the Department of State, announced that according to his reading of the 1972 ABM Treaty, development and testing are prohibited only for ABM systems and components that were "current" at the time the treaty was written. The Reagan Administration has since indicated it would abide by a more restrictive interpretation.

Even so, the Reagan Administration has exploited definitional ambiguities in order to legitimize several planned SDI space experiments. A study by the Aspen Strategy Group, cochaired by William J. Perry, undersecretary of defense for research and development in the Carter Administration, and Brent Scowcroft, national security adviser to President Ford, states that the U.S. has conducted SDI experiments "whose consistency with [the ABM Treaty] is measured by criteria that we would probably reject if the Soviets used them to justify their programs."

The Aspen study concludes that "progress towards a new [strategic arms] agreement should be possible" if the U.S. modifies its position that virtually all development and testing of modern space-based defensive systems are allowed, in return for the Soviet Union's dropping its demand that SDI development and testing be done exclusively in laboratories.

For the Record

The U.S. National Archives and Records Administration tends more than three billion documents, ranging from the Declaration of Independence to ledgers recording the sale of slaves, at 13 centers across the U.S. How can the documents themselves or, even more important, the information they contain best be preserved?

Two years ago, spurred by estimates that more than 500 million of its papers were deteriorating, the archives asked the National Research Council to study that question. The council assembled a Committee on Preservation of Historical Records that included experts in paper, film and electronic-storage media. The electronics experts came from the IBM San Jose Research Laboratory, the University of California at San Diego's Center for Magnetic

NAVAL PENSIONER'S RECORD is one of many frequently handled documents that the U.S. National Archives hopes to save from further wear by copying onto optical disks.

Recording Research and the U.S. National Security Agency, whose collection of intelligence at Fort Meade, Md., may be the largest and most advanced data base in existence.

The committee's recommendations, issued recently in a 108-page report titled *Preservation of Historical Records,* nonetheless were not slanted toward advanced technology. They called for the use of paper of a better quality by the Federal Government and for research into how the microenvironment within a container in which a document is stored can best be controlled to ensure long life. If documents are in danger of crumbling, the report advised, they should be photocopied or microfilmed. The report rejected such advanced storage methods as optical disks and magnetic tape as "inappropriate at the present time."

Although he says he is in general pleased with the report, the acting administrator of the National Archives, Frank G. Burke, is "a little disappointed" by the committee's brusque treatment of electronic storage techniques. The archives, he points out, recently launched a two-year, $1-million pilot project in which 1.5 million pages of 19th-century military pension records (one of the largest and most frequently examined categories of papers) are to be stored on optical disks. Investigators will then be able to call up the records on terminals and obtain hard copy from laser printers, Burke says.

The chairman of the preservation committee, Peter Z. Adelstein, who recently retired from the materials science and engineering division of the Eastman Kodak Company, says he was never told of the archives' pilot project. He acknowledges that, for "rapid access" to relatively small numbers of frequently used documents, a retrieval system based on optical disks would be valuable. But he maintains that the longevity of disks has yet to be determined (the most optimistic estimates are 20 years); moreover, the technology is changing so fast that systems procured now will soon be obsolete. Another disadvantage of optical disks or magnetic tape, Adelstein notes, is that without the appropriate software and hardware they are worthless.

Emphasizing preservation instead of ease of access seems to make sense when one considers that a typical file stored at the archives is viewed, on the average, less than once every 100 years. Burke suggests, however, that a more accessible system might be used more frequently. The consequences of a deeper delving into the archives could be profound. This year the archives yielded documents that expose the World War II records of both Ferdinand E. Marcos, former president of the Philippines, and Kurt Waldheim, now president of Austria.

The Unhappy Years

Is the well-being of the American adolescent declining? Peter Uhlenberg and David Eggebeen of the University of North Carolina write in *The Public Interest* that improvements in some traditional barometers of teen-age health mask a "uniform and serious decline" in adolescent well-being between 1960 and 1980.

In a study of accepted indices of white teen-agers' well-being, Uhlenberg and Eggebeen found that the proportion of young people who lived in homes without poverty, who did not have large families and who did not have poorly educated parents doubled (from 32 to 63 percent) over the two decades. Per-pupil expenditures on schooling also doubled. Balanced against those figures, however, are some disturbing trends: an 11 percent decrease in SAT verbal scores and more than doubled rates of teen-age delinquency, illegitimate birth, drug usage and suicide.

Uhlenberg and Eggebeen assign the blame for this deterioration primarily to modern parents, who the investigators say make fewer sacrifices for their children and spend less "quality time" with them—trends flagged by the rising divorce rate and the growing female work force.

Other investigators are less eager to proclaim a direct causal link between family structure—or indeed any other element of slow societal flux—and the general level of adolescent well-being. "It is reasonable to suppose," says Leon Eisenberg of the Harvard Medical School, "that living in the nuclear era, rates of unemployment, corruption on the public scene and so on are connected with rising indices of social disruption, but there is no methodology to prove it."

Deborah K. Walker of the Harvard School of Public Health finds that different "at-risk behaviors" share common roots. Dropping out of school and drug abuse are not discrete problems but rather, according to Michael S. Jellinek of the Massachusetts General Hospital, "final common pathways" resulting from a wide range of pressures on young people.

Identifying the stresses most often associated with a final common pathway pinpoints groups of teen-agers who are at risk and thereby provides a focus for government, community and family intervention. This is one goal of the year-old Task Force on Youth Suicide of the Department of Health and Human Services. In a meeting late last year the task force undertook to rank as signals for intervention some 100 factors thought to precipitate suicide, which is the second most frequent cause of death (after accidents) for males from 15 through 19 years old.

According to James A. Mercy of the U.S. Centers for Disease Control, social factors often associated with the rise in the teen-age suicide rate, which has tripled in the past 30 years, include the cohort effect (children of the baby boom suffer from competition for limited resources); family instability, exemplified by a severalfold increase in reports of child neglect; drug and substance abuse, and the increased availability of firearms. Areas of high gun ownership are also areas of high suicide rates.

These factors are associated with suicides catalyzed both by clinical depression and by impulsive responses of troubled teen-agers to "windows of vulnerability" in their lives, according to Madelyn S. Gould of the Columbia University College of Physicians and Surgeons.

Another social factor often cited is the effect of the popular media. Two studies reported in the *New England Journal of Medicine* associate television with a type of at-risk behavior now receiving much attention: imitation of suicides described in news stories or depicted in television films.

David P. Phillips and Lundie L. Carstensen of the University of California at San Diego analyzed the number of teen-age suicides reported in the week following news or feature stories about suicide. In a study of 38 stories that were broadcast nationally from 1973 through 1979, they found on the average about three more teen-age suicides occurred in the first week after media coverage than would have been predicted by other factors.

Gould and David Shaffer of Columbia analyzed the effect of four fictional television programs about suicide that were broadcast in the New York City area. The rates of both successful and attempted suicides rose in the two weeks following the telecasts.

BIOLOGY

Lighting Up

What do you get when you insert a gene from a firefly or a luminescent microbe into a tobacco plant? Two groups of geneticists have independently found the answer: plants that glow in the dark—and a powerful tool for tracing the activity of genes.

A group at the University of California at San Diego transplanted the firefly gene; investigators from the Boyce Thompson Institute for Plant Research at Cornell University and from the Max Planck Institute for Plant Breeding in Cologne transplanted two genes from a light-emitting bacterium that lives in seawater.

The six researchers at San Diego, who report their results in *Science*, isolated the firefly gene that encodes the enzyme luciferase. A firefly glows when luciferase mediates the reaction of the small organic molecule luciferin with the cellular fuel adenosine triphosphate (ATP) and oxygen. The investigators decided to insert the gene into tobacco plants because "they're the laboratory rats of the agricultural scientists," says Donald R. Helinski of San Diego. "They're easy to put a gene into and they're easy to regenerate."

The workers spliced the luciferase gene to a fragment of DNA, drawn from a plant virus, that promotes the expression of the gene. They then introduced the recombinant gene into *Agrobacterium tumefaciens*, a bacterium that transfers part of its own DNA to plant cells it infects. From tobacco cells infected by the altered bacteria the investigators were able to regenerate whole plants. The plants glow when they are watered with a solution spiked with luciferin.

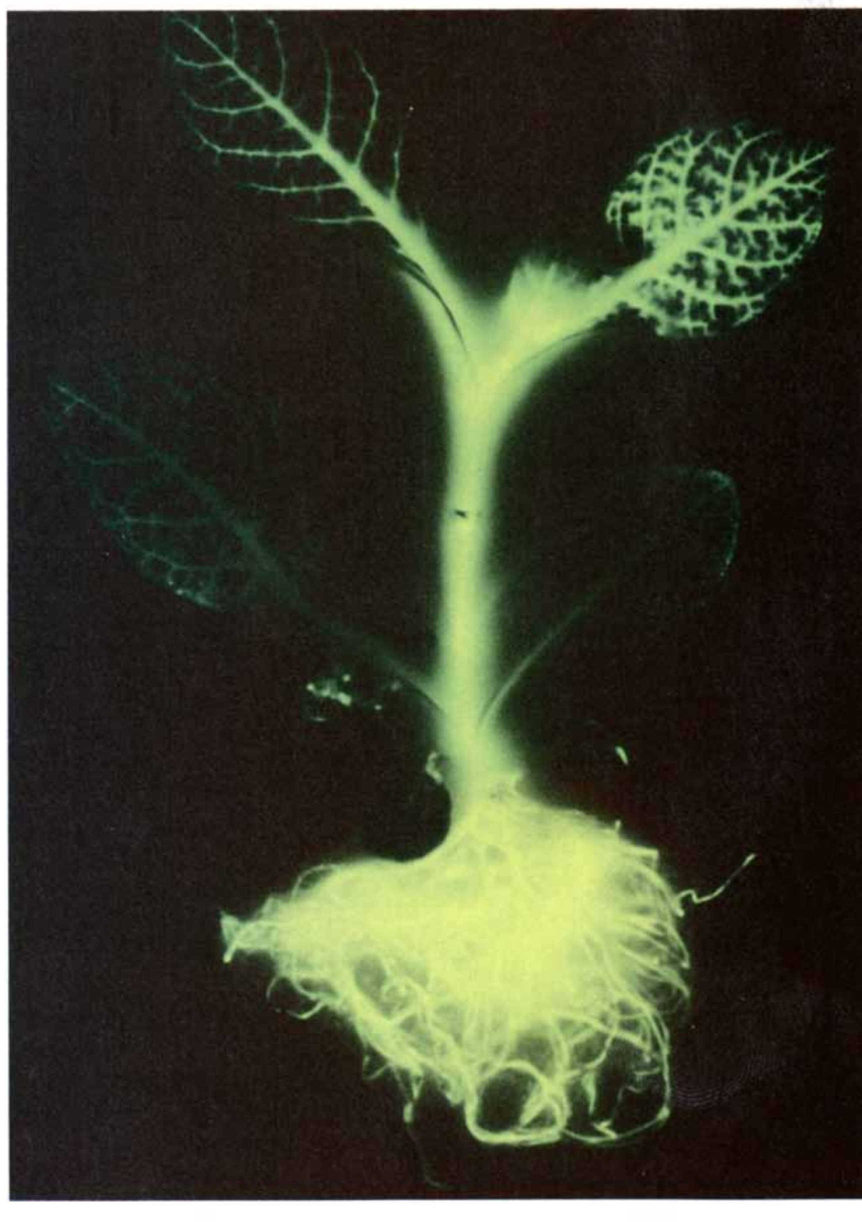
The Boyce Thompson–Max Planck workers, headed by Aladar A. Szalay, report in *Proceedings of the National Academy of Sciences* that they extracted and combined two luciferase genes from *Vibrio harveyi*, bacteria that cluster on some deep-sea fishes and cause them to glow. To produce light in tobacco plants, these prokaryotic genes (in contrast to the eukaryotic firefly gene) required not luciferin but the organic chemicals decanal and flavin mononucleotide. Pieces of the plant must be soaked in these chemicals; the group has not yet induced whole plants to absorb the chemicals through their roots.

Both groups believe luciferase genes will make ideal "reporters" on the activity of other genes. Workers would attach the reporter gene to the promoter fragment from a specific gene—extracted from a specific organism—that interests them. (In the tobacco-plant experiments, the promoter fragments were chosen for their ability to introduce and activate the luciferase gene throughout the plant.) The reporter gene would then be introduced into a cell of the organism. An organism regenerated or grown from the cell would glow only when and where the gene under study is activated.

Thus with simple light detectors geneticists could presumably monitor a broad array of genetic events. Szalay and investigators at Boyce Thompson, Texas A&M University and the University of Georgia have achieved some first results of this kind. They attached their luciferase genes to a promoter for the genes that encode the enzyme nitrogenase in the bacterium *Bradyrhizobium japonicum*. The enzyme catalyzes the fixation of atmospheric nitrogen in the root nodules of legumes harboring *B. japonicum*. The promoter-gene fusion was introduced into the bacterial chromosome. Root nodules of soybean plants inoculated with the bacteria glowed when a signal from the plants called for nitrogen.

In the future it might be possible to observe how a gene that promotes disease resistance in corn manifests itself in successive generations, or to determine the moment at which a particular gene is activated in a developing embryo. The system can potentially trace "any gene you are interested in," says Helinski, including those of human beings. A photodetection scheme would be from 100 to 1,000 times more sensitive and far less expensive than the assay for gene tracking that is most widely used now, says Marlene DeLuca of San Diego.

Both the Boyce Thompson Institute and the University of California at San Diego have filed for patents covering the production of luciferase genes. The groups are now introducing the luciferase gene into other organisms—fungal, fruit-fly, rabbit and bovine



TOBACCO PLANT containing firefly genes glows in the dark when it is watered with the organic chemical luciferin. Keith V. Wood of the University of California at San Diego exposed film to the glowing tobacco plant for 24 hours to obtain this photograph.

cells at Boyce Thompson and yeast and monkey cells at San Diego.

Networking

Neurobiologists seeking to study how electrical impulses travel through networks of nerve cells encounter a problem familiar to physicists studying quantum events: the attempt to observe disrupts the phenomenon. Unlike the physicists, for whom the quandary is a condition of existence, the neurobiologists may have a solution to their problem.

The technique, now being developed at the AT&T Bell Laboratories in Murray Hill, N.J., the California Institute of Technology and North Texas State University, involves culturing neurons on arrays of microelectrodes. These arrays, built with techniques like those used to manufacture semiconductor chips, can both monitor and initiate electrical activity within the network of interlaced cells.

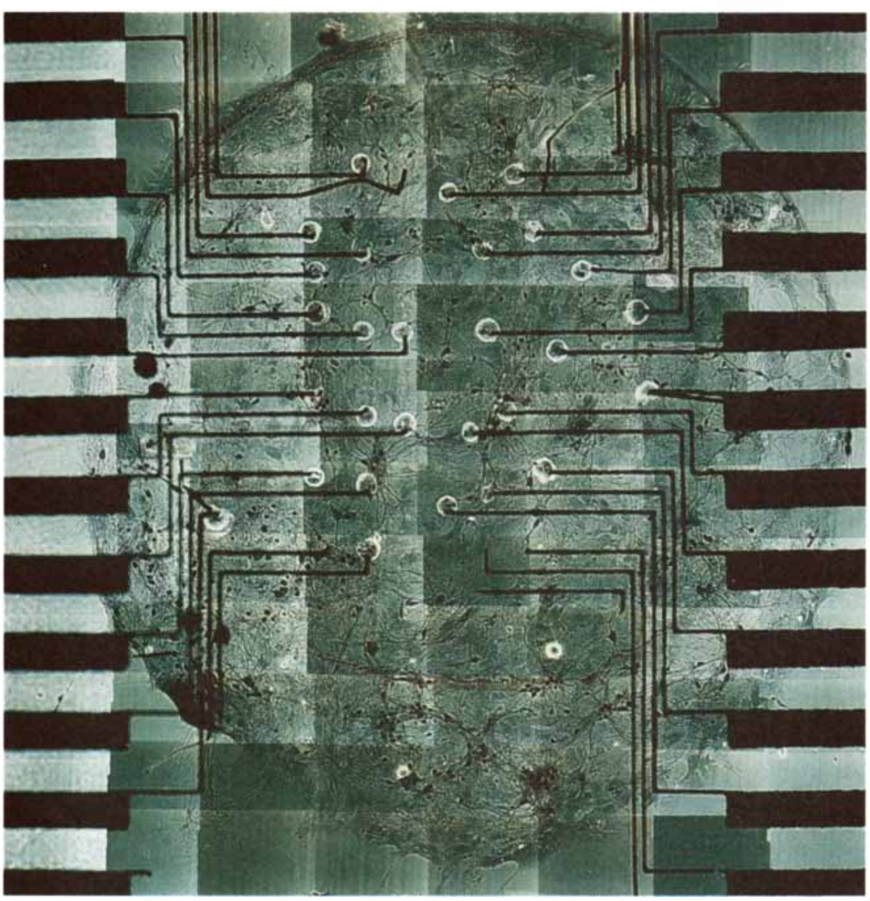
The North Texas State investigators have reported the first results from long-term monitoring of nerve cultures with microelectrode arrays. In *The Journal of Neuroscience* they write of having recorded electrical impulses—often occurring in rhythmic patterns—in cultures of some 300 to 400 neurons extracted from the spinal cord of a mouse embryo. The head of the project, Guenter W. Gross, suggests that the rhythmic firings may reflect a basic form of cell-to-cell recognition and communication, but he acknowledges that the patterns of activity may turn out to be specific to cultured networks rather than to networks in a living organism.

Gross and his colleagues hope to learn more with more advanced technology. Recently they began replacing chips that have 36 opaque gold leads with chips whose leads are transparent. The new leads, made of indium–tin oxide, allow an unobstructed view of the entire culture. Gross estimates that chips with such leads could contain as many as 400 electrodes in each square millimeter and still remain transparent.

The group is also working on methods for "conditioning," or manipulating, nerve cultures by exposing them to electrical impulses, to chemicals that affect neurotransmission and to laser microbeams, which sever connections between cells. Gross says he hopes to develop a technique whereby a laser can temporarily—for perhaps five minutes—shut down a cell's ability to fire. This capability would allow time to observe how a simplified network operates without sacrificing any circuitry.

Phillip G. Nelson, chief of the laboratory of developmental neurobiology at the National Institute of Child Health and Human Development, suggests that Gross's early findings of rhythmic activity are less important than the methods he has devised to observe and manipulate cultures of nerves. With these techniques, Nelson thinks, Gross and others may soon elucidate how neural networks process and store information.

Interest in neural networks ranges across many disciplines, Gross notes. A meeting he convened last October drew experts in artificial intelligence, parallel processing and prosthetic devices as well as neurobiology.



NERVE CELLS grow in a monolayer on a chip of microelectrodes in this composite micrograph made at North Texas State University. The chip is 1.8 millimeters wide.

Dinosaurs: Decline or Fall?

By now it is almost commonplace to ascribe the extinction of the dinosaurs to the impact of an asteroid, one effect of which might have been a prototypical nuclear winter. Without denying that an asteroid did strike, some paleontologists maintain that dinosaurs were in decline long before the catastrophe, which marked the end of the Cretaceous period some 65 million years ago. New results suggest they may even have survived the impact.

Robert E. Sloan of the University of Minnesota, J. Keith Rigby, Jr., of the University of Notre Dame, Leigh M. Van Valen of the University of Chicago and Diane Gabriel of the Milwaukee Public Museum have described work supporting the gradualist view in *Science*. The work focused on the Hell Creek Formation, a fossil-rich badlands of sedimentary rock in Montana laid down during the final two million years of the Cretaceous, when the area was a swampy delta.

Sloan and his colleagues compared the number of dinosaur species catalogued at Hell Creek with data from older sites in Wyoming and Alberta. The combined record showed that over the last eight million years of the Cretaceous the number of dinosaur genera in the region declined from 30 to 12. Within the Hell Creek deposit, the workers report, the number of dinosaur specimens that have been recovered per square kilometer of exposed stratum falls steeply in the most recent Cretaceous sediments.

The findings are at odds with some other claims. For example, Dale A. Russell of the Canadian National Museum of Natural Sciences has argued that there is no clear evidence for a loss of dinosaur species until they disappear at the end of the Cretaceous, which at Hell Creek and elsewhere is demarcated by a thin stratum, rich in the rare element iridium, thought to represent fallout from the asteroid impact. Russell, who believes the record argues for a catastrophic extinction, is skeptical of the current results. He points out that most of the fossils at Hell Creek are fragmentary, making it hard to assess species diversity, and that the deposits have undergone erosion and redeposition, which may have blurred the chronology.

Another finding by Sloan and his coworkers is sparking new debate. In sediments lying above the iridium layer at Hell Creek the group discovered teeth from seven species of dinosaurs together with fossils from mammals and plants characteristic of the Paleocene, the first epoch after the Cretaceous. The group estimates that the strata date from the 40,000 years following the end of the Cretaceous and the asteroid impact.

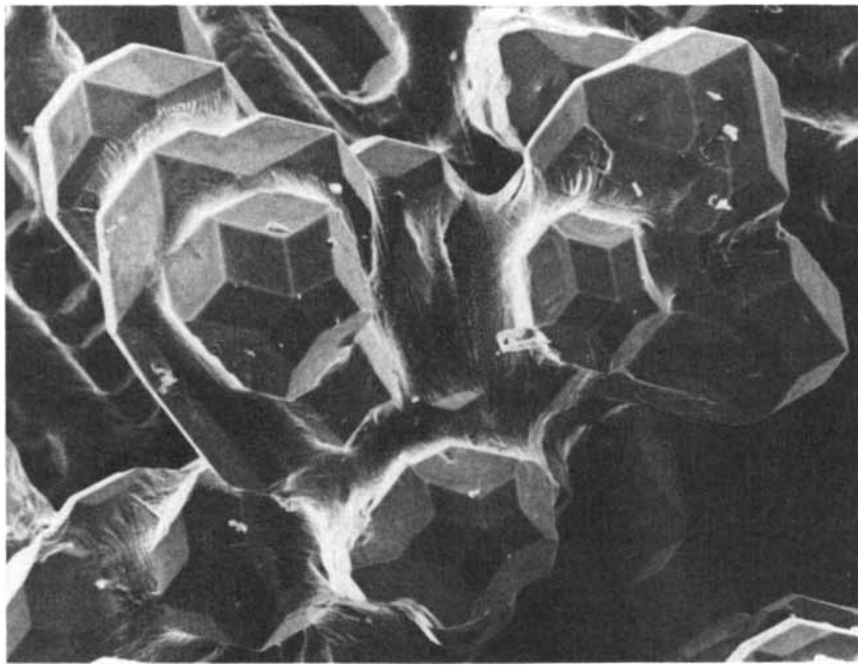
Other workers raise the possibility that the dinosaur fossils were eroded from Cretaceous strata and redeposited with the younger plant and mammal fossils long after the dinosaurs themselves had become extinct. If so, Sloan and his colleagues argue, one would expect to see considerable abrasion of the teeth and to find them mixed with other anomalously old fossils that are commoner than dinosaur teeth in the Cretaceous strata. The investigators found neither.

Southern China may have harbored other dinosaur survivors. A study of fossils from the area, done in collaboration with H. K. Erben of the University of Bonn and Ting Su Yin and Zhao Zi Kuei of the Institute for Vertebrate Paleontology and Paleoanthropology in Peking, enabled Sloan to announce a second discovery of post-Cretaceous dinosaurs at a recent meeting of the Society for Vertebrate Paleontology.

The highest dinosaur-bearing strata, which contained eggshells, tracks and teeth, lay just below strata preserving species of mammals known to have entered China from North America, where they had evolved a million years or more after the end of the Cretaceous. The mammal-bearing strata in China, then, must have been laid down at least that late, and the dinosaur-bearing sediments just slightly earlier. Hence the last dinosaurs may have died out in southern China even later than they did in Montana.

What caused the long decline? In both Montana and southern China it seems to have coincided with a diversification of mammals. As the number of dinosaur genera at Hell Creek shrank from 12 in the late Cretaceous to about seven in the early Paleocene, fossil ungulates (ancestors of today's hoofed mammals) increased from one species to eight.

Sloan and his colleagues believe both trends reflect climatic and ecological change. Fossil leaves and pollen give some evidence of climatic cooling at the end of the Cretaceous; the geologic record testifies to a fall in sea level, which would have sharpened seasonal temperature contrasts over continental areas. The extinction of the dinosaurs is best explained, the workers believe, as a process of gradual ecological succession: while the climate grew harsher a tropical or subtropical ecosystem in which dinosaurs were the dominant fauna gave way to a temperate ecosystem characterized by mammals.



QUASICRYSTAL made by Frank W. Gayle of the Reynolds Metals Company has faceted dendrites, or branchings. The fivefold symmetry of the faceting (five facets meet at a point) reflects the fivefold symmetry of the quasicrystal's underlying microscopic structure. The long diagonal of the largest facet seen in this image is about 100 micrometers.

PHYSICS

Quasiprogress

Investigators in several laboratories, in the U.S. and abroad, have now made large samples of the surprising materials known as quasicrystals. The samples will make it possible to focus powerful analytical tools on the central question raised by quasicrystals: How are the atoms in this newly discovered phase of matter arranged?

The term quasicrystal was coined to describe certain properties first found in an alloy of aluminum and manganese. Diffraction patterns of some samples of the alloy have fivefold rotational symmetry; that is, a pattern rotated by one-fifth of a full circle looks the same as an unrotated pattern. It is impossible to devise a crystalline structure—a periodic, or regularly repeating, arrangement of "unit cells"—that has fivefold symmetry. It was therefore proposed that the new alloy, and others with similar diffraction patterns, have a quasi-periodic structure—not precisely periodic but not completely random either.

Linus Pauling's contention that the observations do not represent a phase of matter but are merely an instance of a crystallographic phenomenon called twinning has not played much of a role in the discussion. Still, some investigators maintain that "quasicrystal" is an inappropriate name. They propose that the materials are indeed a new phase of matter but that their structure is not quasi-periodic at all: instead it is what might be called an icosahedral glass. In an icosahedral glass, some atoms would be bound in groups shaped like icosahedrons (polyhedrons that have 20 triangular sides; an icosahedron has fivefold symmetry because five faces meet at each vertex). The icosahedrons would be oriented in exactly the same direction, but they would be distributed at random loca-

Hitachi's advances in transportation include the joint development with Japan National Railways of a linear motor train with a potential speed of 500km/h, a smoother, quieter linear motor train, large-capacity thyristor and monorail car.

tions throughout a material that is otherwise amorphous, or glassy.

Until now investigators have been unable to check such proposals because they have been able to produce only micrometer-size samples of quasicrystals. The new samples, made from an alloy of aluminum, lithium and copper, are much larger. Some, such as those made by Earle R. Ryba of Pennsylvania State University, are shaped like toothpicks; they are .1 to .2 millimeter in diameter and more than two millimeters long. Others, such as those made by B. Dubost and his colleagues at the Pechiney Research Center in France and by Frank W. Gayle of the Reynolds Metals Company, are chunks of material sprouting beautifully faceted branchings. The facets on each branch are perfectly aligned with those on all other branches, and the faceted structure covers as much as a square centimeter of material, indicating that each sample represents, in some sense, a single "crystal."

Accordingly investigators have begun to probe the samples with diffraction techniques that are normally applicable only to single crystals. An early analysis done by Paul M. Horn of the IBM Thomas J. Watson Research Center, Paul A. Heiney of the University of Pennsylvania and Heiney's student Peter A. Bancel suggests that the icosahedral-glass model may indeed be more suitable than the quasicrystal model for describing phases having fivefold symmetry.

In addition to answering questions, the faceted quasicrystals pose a few. Gayle notes that neither the quasicrystal model nor the icosahedral-glass model includes uniform, flat planes of atoms. How, then, do the samples grow to have such clean facets?

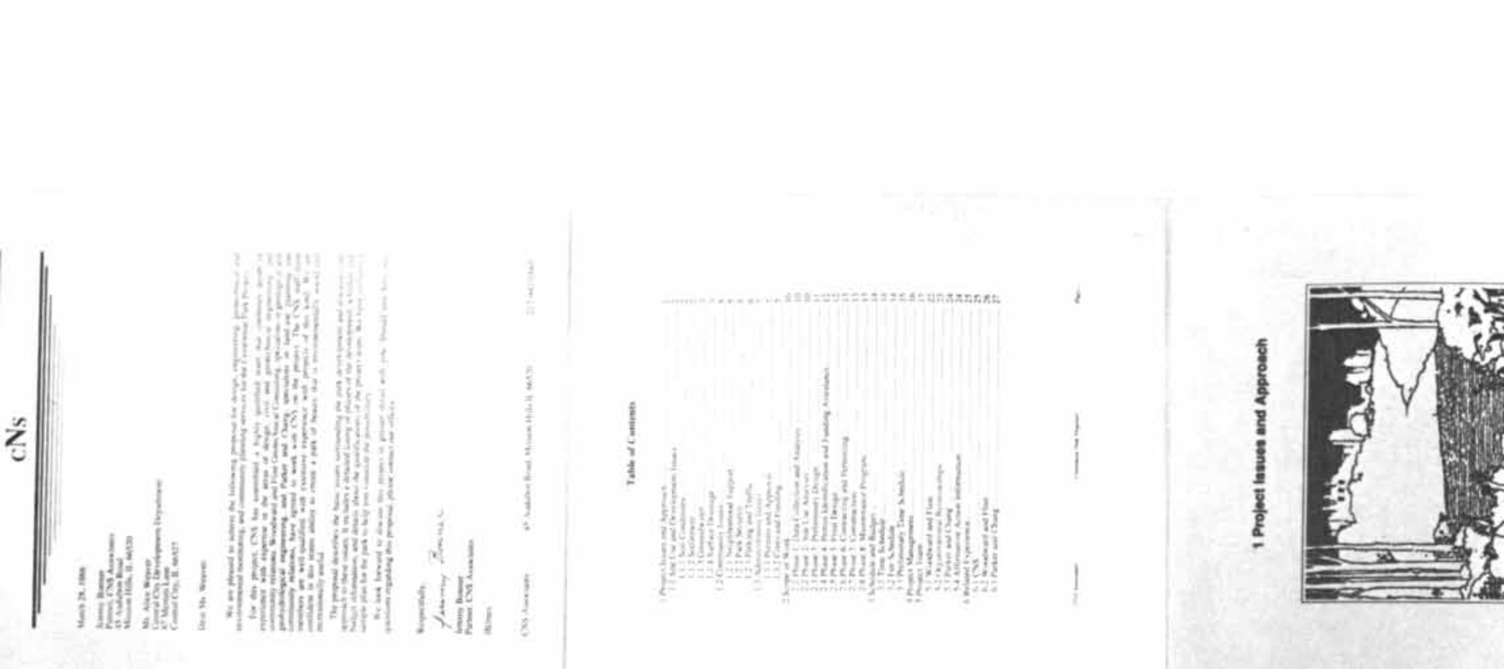
Gone with the Wind?

One year ago a group of physicists announced they had discovered a new fundamental force. The discovery would bring to five the number of forces that determine the nature of matter; the others are gravity, electromagnetism and the weak nuclear force (these two are now recognized as a single entity, the electroweak force), and the strong nuclear force. Now two other physicists contend that the case for the fifth force rests on nothing but puffs of air.

The fifth-force theorists, headed by Ephraim Fischbach of Purdue University, made their discovery by poring over records of experiments carried out in Budapest early in this century by Baron Roland von Eötvös. Eötvös' measurements of the tension exerted on a torsion balance by weights made of different materials had been considered proof of the notion, first advanced by Galileo, that gravity exerts force on objects without regard to their composition.

There the matter rested until Fischbach and his colleagues, writing in *Physical Review Letters,* reported that tiny, seemingly insignificant discrepancies in Eötvös' data appeared to vary according to the number of baryons (protons and neutrons) per unit of mass in the weights. To the Fischbach group this suggested the existence of a baryon-sensitive force that causes objects to repulse one another weakly over distances of about 1,000 feet. Such a force would explain some open questions, the Fischbach group wrote, such as why the gravitational constant seems to be larger underground than it is at the earth's surface.

A decidedly banal explanation for the variances in the baron's data has now been proposed by Shu-Yuam Chu of the University of California at Riverside and R. H. Dicke of Princeton University. They argue in *Physical Review Letters* that subtle variations in the temperature of the air in Eötvös' laboratory created gentle breezes that exerted a pressure on the weights proportional to their size. Dicke says he encountered such convection currents when he tried to reproduce the Eötvös



experiments in the 1950's. There is still a chance Fischbach could be correct, Dicke believes, although he adds, "I don't think it's very likely."

The Chu-Dicke hypothesis, Fischbach says, is "very clever and sufficiently promising to warrant more detailed study." In his view, however, it is not without flaws of its own. Fischbach notes that the discrepancies in Eötvös' data are consistent for the 4,000 hours during which the baron conducted his experiments. "Would you expect the breezes in a room to be the same day in and day out and year in and year out?" asks Fischbach. "I have a mobile in my room and it changes direction all the time."

Both Dicke and Fischbach agree that the matter will be resolved soon by experimenters. Several are testing, for example, whether large, vertical landmasses exert a repulsive force on nearby objects that varies according to baryon content. Peter Thieberger of the Brookhaven National Laboratory has done experiments on the Palisades, the cliffs overlooking the Hudson River in New Jersey. Val L. Fitch of Princeton took his experimental apparatus to Montana. "I wanted a mountain," he explains.

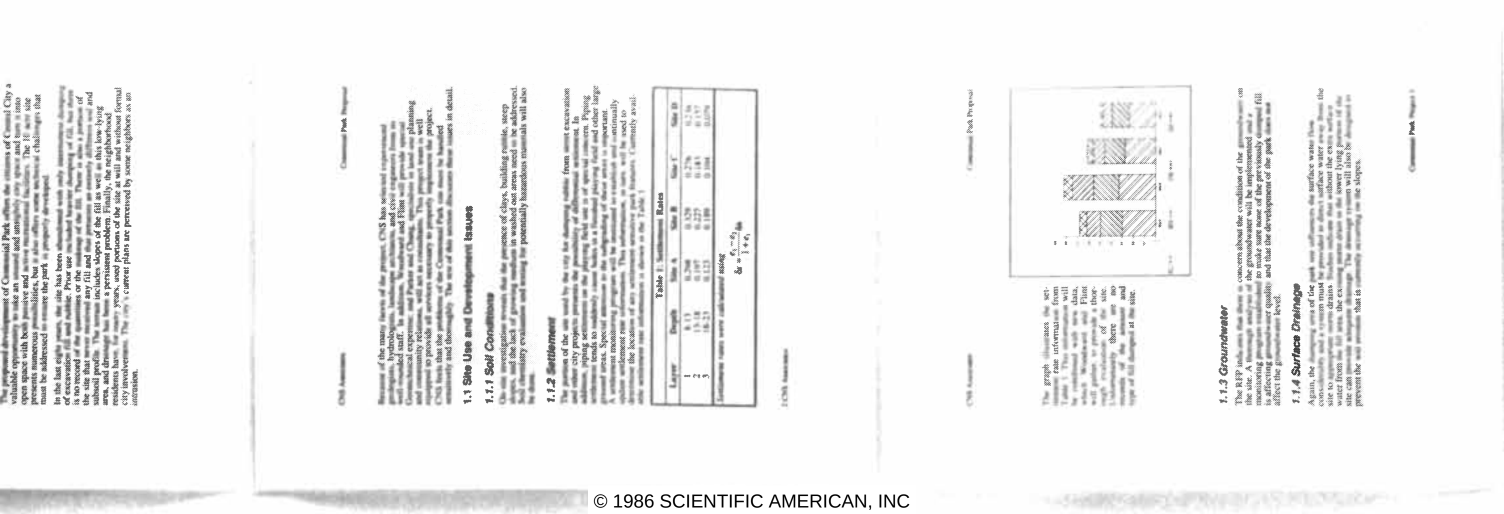
If the fifth force is proved to exist, Fitch says, theoretical physicists seeking a unified theory of fundamental forces will not have to discard their work. "The theorists have a remarkable knack for landing on their feet when confronted with some new experimental fact," he notes.

Ozone

Two observations have been made lately concerning the ozone layer that protects the biosphere from much of the sun's ultraviolet radiation. First, satellite measurements seem to confirm the theory that the ozone layer is gradually being depleted. Since 1979, the measurements suggest, ozone concentrations have declined by several percent; in the mid-latitudes of the Southern Hemisphere the drop has been as much as 18 percent. Second, over roughly the same period a more dramatic ozone depletion—it has come to be known as the ozone "hole"—has been observed over Antarctica every spring, in September and October. Although the hole is filled during the summer, it has returned each year in deeper form: the mean ozone concentration over the Antarctic in October of 1985 was some 40 percent less than it had been in October, 1979. A number of recent articles, including those in a special issue of *Geophysical Research Letters,* demonstrate that no consensus exists on the cause of the Antarctic trend. Nor is there agreement on the relation of the Antarctic hole to global ozone depletion.

Some investigators regard the hole as a warning of imminent global disaster: an extreme example of ozone destruction caused by manmade chlorofluorocarbons. When these compounds are broken down by ultraviolet light in the stratosphere, they release chlorine, which catalyzes the breakdown of ozone (O_3) into molecular oxygen (O_2). The U.S. banned the use of chlorofluorocarbons as propellants in aerosol sprays in 1978, but they continue to serve as refrigerants and as foaming agents in this country, and in most of the world as propellants too. Hence their concentration in the atmosphere is still rising—by more than 5 percent per year over the Antarctic.

A second theory, put forward by Linwood B. Callis of the National Aeronautics and Space Administration's Langley Research Center and by Murali Natarajan of SASC Technologies, Inc., of Hampton, Va., argues that both the Antarctic hole and the global ozone depletion are caused by naturally occurring nitrogen oxides. According to Callis and Natarajan, the levels of nitrogen oxides in the lower stratosphere rise and fall with the 11-year cycle of solar activity. Near a solar maximum (the last of which was in



1979), large amounts of nitrogen compounds produced above the stratosphere are drawn into the lower stratosphere by the south-polar vortex, a circular wind pattern that lodges over the Antarctic during the long polar winter. Some of the compounds are photolyzed into ozone-destroying NO and NO_2 by the spring sun. From the Antarctic the nitrogen oxides spread through the stratosphere, destroying ozone over other parts of the globe. Although this theory by no means exonerates chlorofluorocarbons, it does say that most of the ozone depletion observed so far is cyclical and has a natural cause.

Still other workers contend that the Antarctic hole and the global ozone thinning are essentially two distinct phenomena, and that the hole is produced not by chemicals but by winds. Writing in *GRL,* Richard S. Stolarski and Mark R. Schoeberl of NASA's Goddard Space Flight Center report that the total amount of atmospheric ozone poleward of 44 degrees south latitude remains almost unchanged as the polar hole forms each spring. A possible explanation is that upwelling air currents push ozone-rich air in the lower stratosphere away from the pole and replace it with ozone-poor air from lower altitudes; the ozone-rich air collects in a ring at subpolar latitudes, where high concentrations of ozone have in fact been observed. In this view the hole has been deepening from year to year because the upwelling, for reasons that are not yet clear, has been intensifying.

Although workers on a recent expedition to the Antarctic reported preliminary evidence against both the solar-cycle theory and the wind theory, the controversy will probably take several years to resolve. A continued deepening of the Antarctic hole would tend to incriminate chlorofluorocarbons; conversely, a recovery of ozone levels would support one of the other two theories. Satellite measurements indicate that the hole was less intense in 1986 than it was in 1985, when ozone levels over the Antarctic hit an all-time low, but it is too early to say whether the recent recovery marks a new trend.

Meanwhile the U.S. Environmental Protection Agency has added its voice to the cacophony. A draft report released by agency officials in November projected that ozone depletion could cause 40 million cases of skin cancer (of which 800,000 would end in death) in the U.S. by the year 2074. The projection assumes that chlorofluorocarbons are depleting the ozone layer and that emissions will continue to grow at the rate of 2.5 percent per year. In the view of some investigators the report, which is still undergoing scientific review, is at least premature; John R. Wiesenfeld of Cornell University called it "irresponsible and misleading."

Other observers, however, note that if chlorofluorocarbons are in fact thinning the ozone layer, it may be imprudent for policymakers to await conclusive proof before attempting to curtail emissions.

TECHNOLOGY

Bottling the Bomb

Weapons designers at the Los Alamos and Lawrence Livermore national laboratories have long debated whether to build an underground chamber that could contain repeated low-yield nuclear blasts. Recently both laboratories decided against building reusable chambers. Although Livermore's decision is apparently final, Los Alamos may issue a contract for a reusable facility within a year.

For several years Livermore physicist Ray E. Kidder has argued that as "a hedge against a test ban" the U.S. Department of Energy (DOE), which funds research on nuclear weapons, should consider building an underground facility called the high-energy-density facility (HEDF).

The HEDF might help the laboratories to bear the loss of high-yield tests, says Kidder, by providing an economical way to conduct many subkiloton tests of the kind a treaty is likely to allow. The high cost of drilling shafts and tunnels now contributes to limiting tests at the Nevada Test Site to about 15 per year. Kidder says the HEDF could accommodate up to 40 blasts per year, each with a yield of 300 tons.

Although 300-ton blasts are far too small to test warheads at their normal yield, they would be sufficient for much weapons-related work, Kidder says, such as modeling the physics of nuclear blasts, observing the effects of radiation on military hardware and staying abreast of what the Soviets might learn with a similar facility. The HEDF, he argues, could thus allay the complaints of U.S. defense officials that a test ban would cause weapons researchers to become bored and leave for other fields. Kidder charges that Livermore and DOE officials decided against the facility because they do not want to support any program that might seem to encourage negotiations for a low-threshold test ban.

John D. Immele of Livermore's nuclear-design division angrily denies the charge, saying, "We are on record as supporting facilities that would provide a hedge against a test ban." Immele contends the laboratory's money would be better spent on a massive laser that produces controlled fusion reactions with yields of about a quarter of a ton.

Whereas Livermore has rejected the HEDF, Los Alamos has quietly drawn up plans for its own version of a reusable test chamber, the large pressure vessel. Last spring the laboratory asked contractors to submit bids for a vessel that could contain nuclear blasts with yields of about 20 tons. The specifications called for a spherical steel vessel, 55 feet across with walls from eight to 12 inches thick, in which the vacuum of space could be simulated, according to George T. Zirps of Babcock & Wilcox, one bidder.

Six contractors responded with bids averaging about $50 million, according to Peter B. Lyons of Los Alamos. Although laboratory officials decided in September that their budget was too small to award a contract for 1987, Lyons says he is "sure it will be discussed" when the 1988 budget is drawn up this summer.

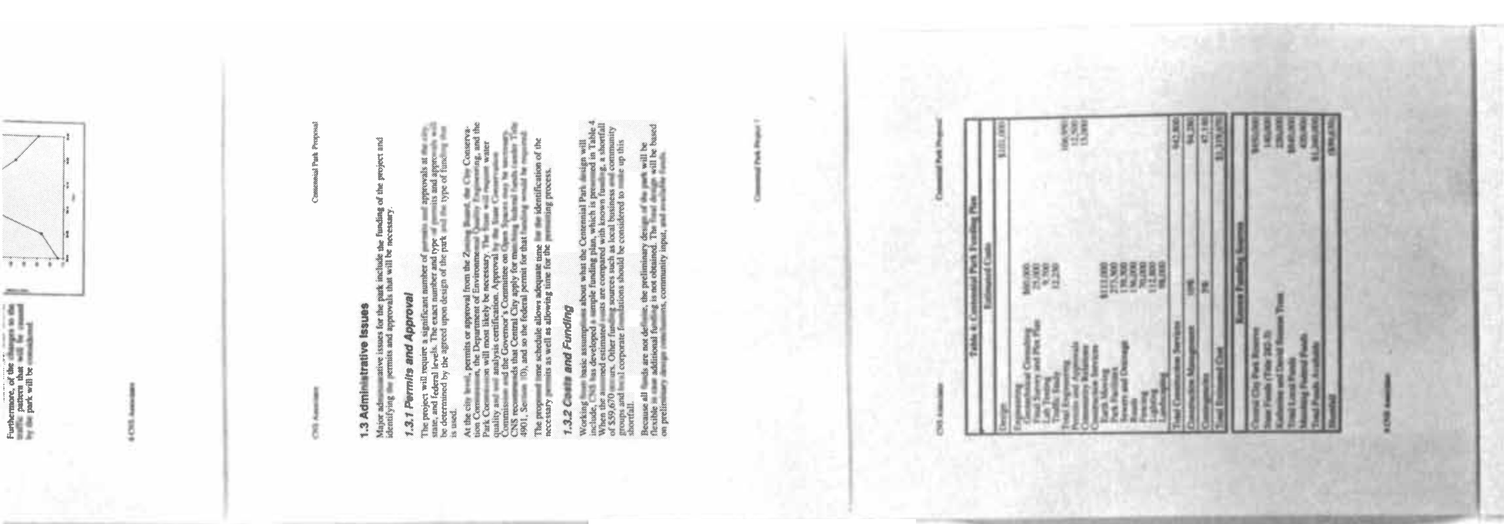
Unlike the HEDF, the pressure vessel is not intended to be a hedge against a test ban, according to Zirps and Lyons. Zirps says that data generated from tests of the 55-foot sphere would be extrapolated to design larger spheres for larger nuclear tests. The 55-foot sphere would also be used for a specific set of experiments Lyons would only describe as "weapons-related."

John Pike, a space-technology expert for the Federation of American Scientists, speculates that the sphere might be used for testing nuclear-pumped directed-energy weapons, a number of which are being considered for the Strategic Defense Initiative. Two likely candidates, says Pike, are weapons that channel a subkiloton nuclear blast in space to form a beam of lethal microwaves or a stream of so-called hypervelocity pellets.

Vital Signs

With funds from the U.S. Navy, engineers at the Georgia Institute of Technology and at Michigan State University in Lansing have developed forms of radar that can detect heartbeat and breathing at a distance. A variant of the system can detect individuals behind concrete.

The technology, under development for the past four years, is based on the fact that microwave radiation reflected off a living person is modulated by the motion of the heart and chest. A signal processor can extract this low-



power, rhythmic signal from the rest of the reflected beam. The investigators report that with a radiation frequency of 10 gigahertz (in the case of Michigan State) or 35 gigahertz (at Georgia Tech) in a very still environment they have monitored the heartbeat and respiration of subjects up to 300 feet away.

The technology is still crude, and it may prove valueless under field conditions. Even the nodding of blades of grass creates electronic noise that can drown out the desired signal. Joseph Seals, project leader at Georgia Tech, says he intends to experiment with a narrower beam and more advanced signal-processing techniques to reduce noise and extend the radar's range.

Elliot Postow of the Naval Medical Research and Development Command, who supervises the Michigan State and Georgia Tech projects, says the Navy has a specific application in mind for the technology: helping medical personnel working under extreme battle conditions, such as the aftermath of a nerve-gas attack, to rapidly distinguish the living from the dead so that rescue efforts can proceed more efficiently.

A low-frequency system tested at Michigan State may have other applications. Using a two-gigahertz beam, the engineers monitored individuals shielded by concrete or brick walls as much as three feet thick, according to Kun-Mu Chen, head of the project. This capability could help to find survivors buried under rubble after an earthquake or explosion, Chen says. It could also reveal the presence of someone hidden in a building or other structure.

Both Seals and Chen suggest that the technology could also provide an alternative method for monitoring patients in a hospital. Postow says Navy officials will decide whether to continue funding after field tests planned for this spring.

MEDICINE

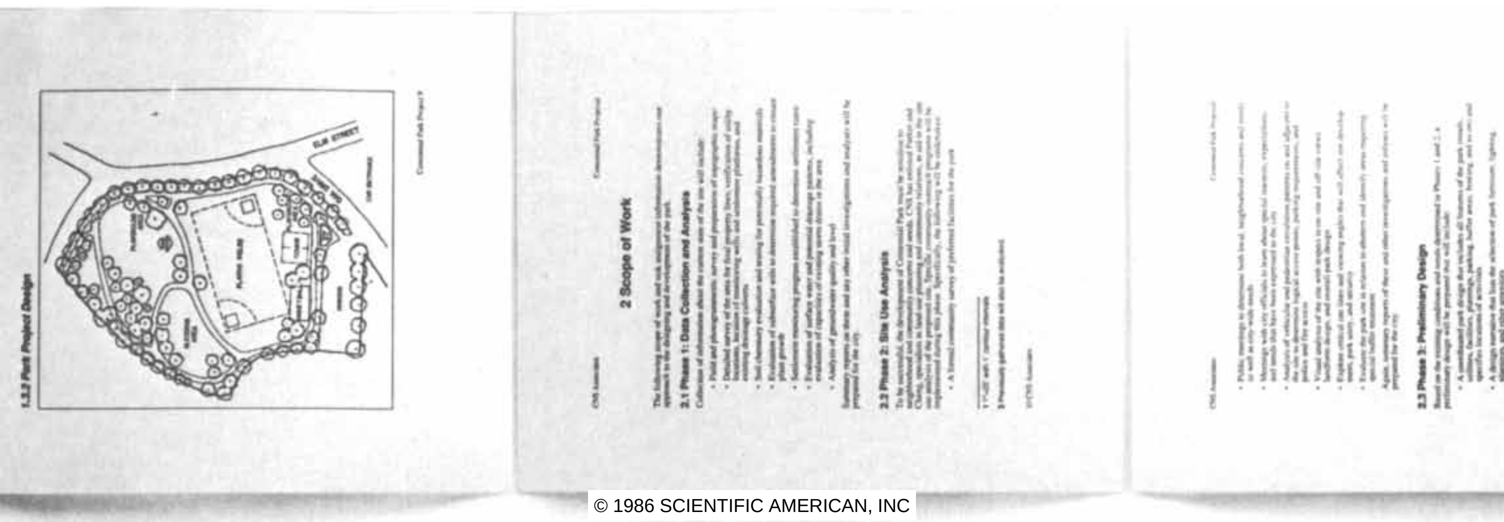
Drugs on the Mind

The proliferation of pharmaceutical drugs has given physicians an unprecedentedly powerful armamentarium for protecting their patients from illness. At the same time, however, reports of psychological side effects, ranging from euphoria to psychosis, have been increasing. *The Medical Letter on Drugs and Therapeutics,* a newsletter published in New Rochelle, N.Y., recently listed 103 drugs that reportedly had caused such reactions, albeit rarely in some cases. Not surprisingly, the list included such powerful psychoactive drugs as amphetamines and barbiturates. But it also included anti-inflammatories, antibiotics, contraceptives, antihypertensives and even nasal sprays.

The *Medical Letter* published a similar list in 1984 that cited only 81 drugs. Does the growing list indicate a growing problem? No, according to Harold C. Neu, professor of medicine and pharmacology at the Columbia University College of Physicians and Surgeons. "It's a reporting artifact," says Neu: doctors have recently grown "more attuned" to the problem and so are more likely to diagnose reactions correctly and report them.

Yet psychiatric symptoms caused by drugs are still misdiagnosed "more often than many of us would like to think," says Steven E. Hyman, a fellow in psychiatry and molecular biology at the Massachusetts General Hospital, "especially in the elderly." The aged are at greater risk because their liver and kidneys metabolize and filter drugs less rapidly and, "for reasons we don't understand," says Hyman, their brains are in many cases more sensitive to drugs.

Moreover, both elderly patients and their doctors often ascribe relatively mild symptoms, such as lethargy, mental dullness and forgetfulness, to aging



instead of to drugs. "If the patient says there are little green men hiding in his radio, the doctor will think of drugs," observes Richard W. Besdine, head of the department of gerontology at the University of Connecticut School of Medicine, "but he won't if the patient forgets his keys or doesn't remember a grandchild." Almost any drug, not just the 103 listed in the *Medical Letter,* can trigger subtle mental side effects in the aged, Besdine says.

Evidence linking some drugs and mental symptoms is largely anecdotal. Several of the *Medical Letter* citations, for example, particularly those describing acute reactions, are based on the experience of one or two patients. But a study published in *Journal of the American Medical Association* last January statistically confirmed a link between depression and beta-blockers, a family of drugs given to treat hypertension. Medicaid records of 143,253 Middle Westerners revealed that patients taking beta-blockers were about 50 percent more likely than patients on other antihypertensives to require antidepressants.

Prescribing drugs to treat the undiagnosed side effects of other drugs may well make matters worse, notes Jerome L. Avorn of the Harvard Medical School, who headed the beta-blocker study. As a first step toward avoiding such eventualities, patients "should take a more active role in asking whether the way they've been feeling lately could be caused by a drug." Avorn acknowledges that in certain cases a particular drug may be so crucial to a patient's health that he or she must simply endure the side effects. The enormous variety of drugs does, however, give doctors the option of prescribing an alternative medication. The medication can also be reduced or eliminated entirely, Avorn points out, adding that doctors might consider these options more often.

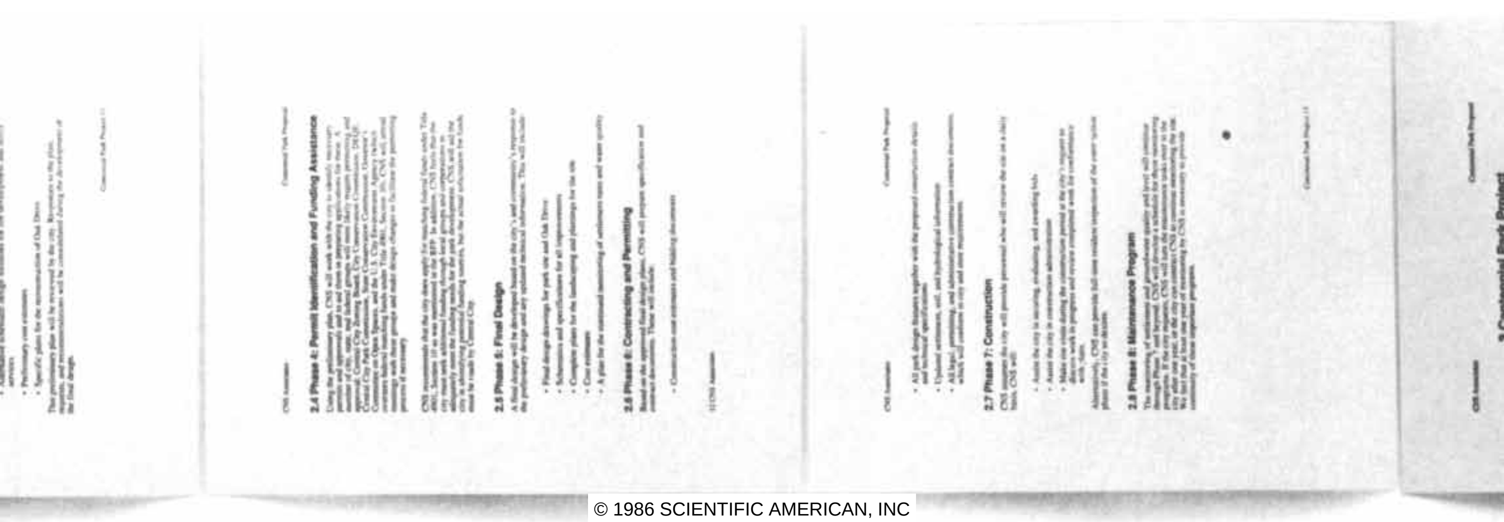
Heart Beat

Good news and a hint of bad news emerged at the annual scientific meeting of the American Heart Association. The good news was that "Type A" people, whose aggressive and impatient behavior is thought to predispose them to heart attacks, are at least no more likely than other people to have a second attack. The possible bad news is that the body's natural endorphins, which give rise to the runner's euphoric "high," may also mask painful symptoms of heart disease.

The work on Type A people was reported by Charles A. Dennis of the Stanford University School of Medicine. He and his colleagues studied 303 men about 50 years old who were recovering from heart attacks. On the basis of personality tests 61 percent of them were classified as Type A; the others had the more relaxed and easygoing approach to life that is called Type B. The investigators found no significant differences between those two groups in the frequency or type of coronary events during a six-month follow-up period.

The study of endorphins was made by David S. Sheps of the University of North Carolina School of Medicine and several colleagues. They had a group of 25 people with confirmed ischemia (a decreased flow of blood to the muscles of the heart) do a pedaling exercise that continued until the pedaler reported angina (a heart pain often experienced in ischemia) or was too tired to continue. Afterward the investigators measured the level of beta endorphins in the blood and found they were from 35 to 40 percent higher in the 10 people who felt no pain than they were in the others. Since exercise causes the endorphin level to rise, the finding could mean that some people with ischemia might not be made aware of it by pain felt during or after exercise.

In Sheps's opinion, however, the finding should be viewed with caution because of the smallness of the sample



and the fact that the habitual physical activity and the fitness level of the participants were not taken into account. Sheps and his colleagues plan broader studies to refine the results.

Considerably more certainty underlies results concerning the relation between blood cholesterol levels and mortality that appeared in the *Journal of the American Medical Association* (JAMA). Jeremiah Stamler and his colleagues at Northwestern University argue that 80 percent of middle-aged men are at risk of premature death because of high levels of cholesterol (more than 180 milligrams per 100 milliliters of blood). Stamler's group based this conclusion on data from a study by the National Heart, Lung, and Blood Institute of 361,662 men aged 35 to 57. Cholesterol is a family of fatty substances that can narrow the arteries and give rise to coronary heart disease.

A hope of reducing cholesterol in some people is held out by an investigational drug, lovastatin, according to another JAMA report. The drug should be available in about two years.

Palliative

Neurochemical research has won what may be a significant skirmish—although by no means the war—against Alzheimer's disease. Preliminary tests of a drug that inhibits the breakdown of the neurotransmitter acetylcholine suggest that the agent can markedly alleviate some of the symptoms of the degenerative disease.

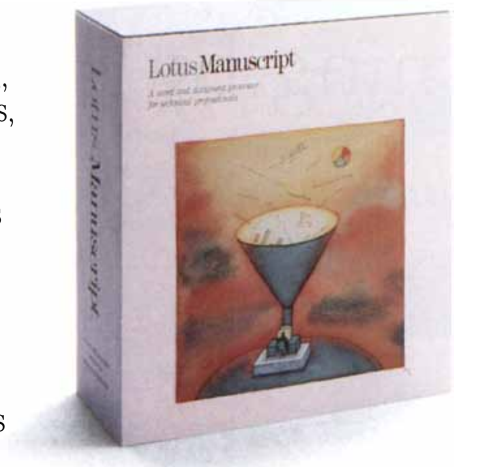
In the *New England Journal of Medicine* William Koopmans Summers of the University of California at Los Angeles and four co-workers describe a trial in which oral tetrahydroaminoacridine (THA) was given to patients afflicted by moderate to severe Alzheimer's disease. THA hobbles the activity of the enzyme cholinesterase. The enzyme breaks down acetylcholine, an important neurotransmitter that is deficient in patients. And, according to Summers, THA also "soups up" the cholinergic (acetylcholine) system, making nerve cells more receptive to transmission by blocking potassium channels in the cells.

By administering four tests of cognitive functioning and general orientation, the team demonstrated that 14 patients who were alternately given oral THA or a placebo improved significantly during periods when they were on the drug. Twelve of these patients have now been taking THA for an average of a year, and all are doing well. Such results, Summers and his colleagues suggest, are "unusual in an illness that is considered to be progressive, with only rare plateaus at best."

In an editorial accompanying the report, Kenneth L. Davis and Richard C. Mohs of the Mount Sinai Medical Center lauded the methodological advances of the study, particularly the exploitation of nuclear-magnetic-resonance (NMR) scanning as a diagnostic tool. The only definitive diagnosis for Alzheimer's disease is tissue diagnosis at autopsy, but with NMR the team was able to screen out subjects who clearly did not have the disease and would therefore have been less responsive to the therapy.

Is blocking the activity of cholinesterase an effective treatment strategy for Alzheimer's disease? According to Davis and Mohs, keeping acetylcholine levels up cannot cushion the inevitable loss of acetylcholine-producing neurons. Nor does it cover deficits in other neurotransmitters and neuropeptides that may contribute to disease symptoms.

Summers describes this line of argument as "the party line at the moment." He is pleased with THA's palliation of significant symptoms. He does concede that even though THA "covers symptoms," there is no way to know whether the disease process itself "is progressing half as fast, at the same rate or twice as fast" until hundreds of cases have been studied.



Yes, Lotus Manuscript™ is a very different kind of word processor. Because when you're processing long, technical documents such as proposals, specifications and reports, you have very different kinds of needs.

Needs like mixing text and graphics on the same page. Or, creating complex equations. Or, working with columns and tables. Or, making wholesale revisions.

Very often technical documents require the mixing of text and graphics on the same page. Graphics from Lotus® 1-2-3,® Symphony® and Freelance® Plus. With Manuscript you can import spreadsheets and charts, as well as diagrams and scanned images.

Another feature designed specifically for the writer of technical documents is Manuscript's integrated Outliner. When you collapse the document you're able to navigate the outline much easier than wading through the entire piece. You can move a page, or even an entire chapter with just a few keystrokes.

Also, because Manuscript is structured, you can globally format an entire document or format by individual sections. Imagine changing all headlines from 14 point plain text to 16 point bold, by a simple menu selection.

How about the age-old dilemma of proof reading revisions? Manuscript's Document Compare feature highlights changes between revisions. You might call it fool-proof-reading.

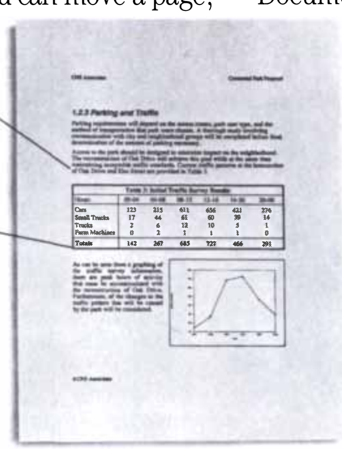
Also, Manuscript automatically sizes and generates math equations as well as upper and lower case Greek symbols, diacritical marks and brackets. No longer do you need to literally cut and paste to create equations.

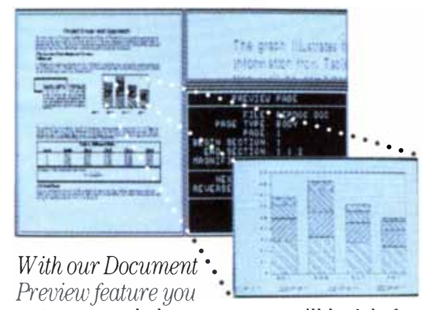
With our powerful Print Formatter you have complete control over the look of your document. You can position and size graphics anywhere on the page. You can choose fonts and point sizes from a built in typeface list. You can even save format information as a template for regular use.

As for the quality of output, Manuscript takes full advantage of today's printing technology, from dot-matrix to laser, including PostScript® printers.

Manuscript allows you to mix text and graphics on the same page.

You can throw away your scissors and glue, cut and paste are a thing of the past.

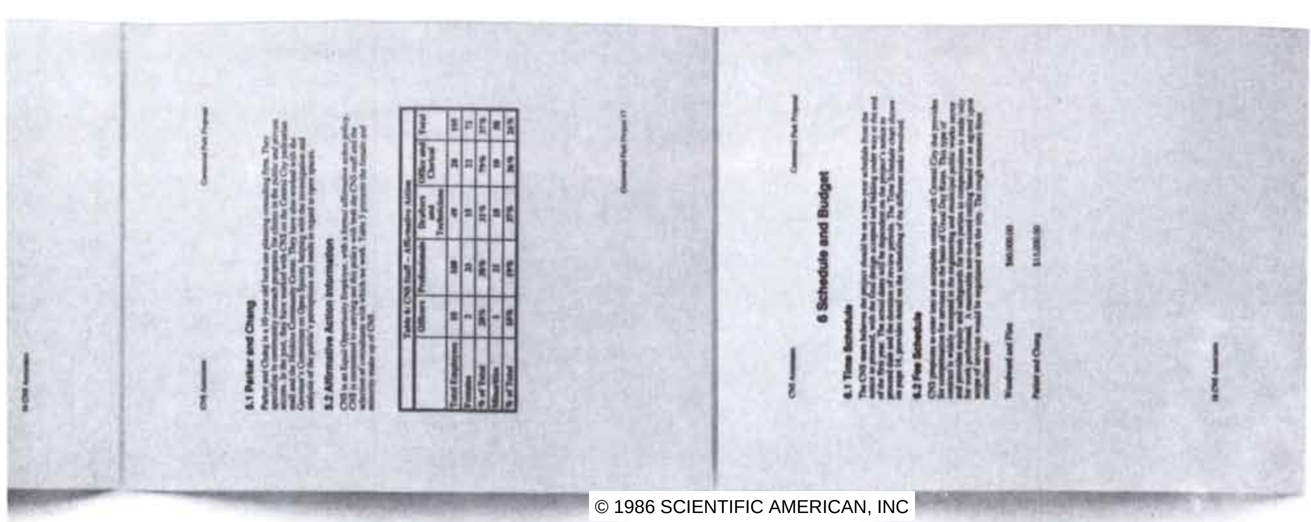




With our Document Preview feature you can see exactly how any page will look before it's printed. Zoom capability lets you take a closer look at graphics and equations.

Manuscript is designed to work on most IBM PCs and compatibles. Its familiar 1-2-3 interface makes it easy to use. And our Manuscript evaluation kit makes it easy to try. For $10.00, you'll get a presentation disk, working software, and a tutorial manual. To get your evaluation kit, call 1-800-345-1043, ask for lot #BD-1450. Or, for more information, see your authorized Lotus dealer, or write Lotus Development Corp., 55 Cambridge Parkway, Cambridge, MA 02142.

May we suggest Lotus Manuscript.



Learning by Instinct

Usually seen as diametric opposites, learning and instinct are partners: the process of learning, in creatures at all levels of mental complexity, is often initiated and controlled by instinct

by James L. Gould and Peter Marler

Learning is often thought of as the alternative to instinct, which is the information passed genetically from one generation to the next. Most of us think the ability to learn is the hallmark of intelligence. The difference between learning and instinct is said to distinguish human beings from "lower" animals such as insects. Introspection, that deceptively convincing authority, leads one to conclude that learning, unlike instinct, usually involves conscious decisions concerning when and what to learn.

Work done in the past few decades has shown that such a sharp distinction between instinct and learning—and between the guiding forces underlying human and animal behavior—cannot be made. For example, it has been found that many insects are prodigious learners. Conversely, we now know that the process of learning in higher animals, as well as in insects, is often innately guided, that is, guided by information inherent in the genetic makeup of the animal. In other words, the process of learning itself is often controlled by instinct.

It now seems that many, if not most, animals are "preprogrammed" to learn particular things and to learn them in particular ways. In evolutionary terms innately guided learning makes sense: very often it is easy to specify in advance the general characteristics of the things an animal should be able to learn, even when the details cannot be specified. For example, bees should be inherently suited to learning the shapes of various flowers, but it would be impossible to equip each bee at birth with a field guide to all the flowers it might visit.

Innately guided learning—learning by instinct—is found at all levels of mental complexity in the animal kingdom. In this article our examples will be drawn primarily from the behavior of bees and birds, our respective fields of particular expertise, but the results can be generalized to the primates, even to man. There is strong evidence, for example, that the process of learning human speech is largely guided by innate abilities and tendencies.

Two Theoretical Frameworks

The distinction often made between learning and instinct is exemplified by two theoretical approaches to the study of behavior: ethology and behaviorist psychology. Ethology is usually thought of as the study of instinct. In the ethological world view most animal behavior is governed by four basic factors: sign stimuli (instinctively recognized cues), motor programs (innate responses to cues), drive (controlling motivational impulses) and imprinting (a restricted and seemingly aberrant form of learning).

Three of these factors are found in the egg-rolling response of geese, a behavior studied by Konrad Z. Lorenz and Nikolaas Tinbergen, who together with Karl von Frisch were the founders of ethology. Geese incubate their eggs in mound-shaped nests built on the ground, and it sometimes happens that the incubating goose inadvertently knocks an egg out of the nest. Such an event leads to a remarkable behavior. After settling down again on its nest, the goose eventually notices the errant egg. The animal then extends its neck to fix its eyes on the egg, rises and rolls the egg back into the nest gently with its bill. At first glance this might seem to be a thoughtful solution to a problem. As it happens, however, the behavior is highly stereotyped and innate. Any convex object, regardless of color and almost regardless of size, triggers the response; beer bottles are particularly effective.

In this example the convex features that trigger the behavior are the ethologists' sign stimuli. The egg-rolling response itself is the motor program. The entire behavior is controlled by a drive that appears about two weeks before the geese lay eggs and persists until about two weeks after the eggs hatch. Geese also exhibit imprinting: during a sensitive period soon after hatching, goslings will follow almost any receding object that emits an innately recognized "kum-kum" call and thereafter treat the object as a parent.

Classical behaviorist psychologists see the world quite differently from ethologists. Behaviorists are primarily interested in the study of learning under strictly controlled conditions and have traditionally treated instinct as irrelevant to learning. Behaviorists believe nearly all the responses of higher animals can be divided into two kinds of learning called classical conditioning and operant conditioning.

Classical conditioning was discovered in dogs by the Russian physiologist Ivan P. Pavlov. In his classic experiment he showed that if a bell is rung consistently just before food is offered to a dog, eventually the dog

INSTINCTIVELY GUIDED LEARNING enables the cuckoo (*large bird at left*) to parasitize other species of birds (in this case a hedge sparrow). Cuckoos lay their eggs in the nests of other birds. When the cuckoo hatches, its appearance and the begging call it emits trick those birds into accepting it as their own young. As the photograph shows, the foster parents continue to feed the cuckoo even after it has grown to be much larger than they are. While it is still in the nest, the cuckoo (if it is female) must somehow learn to recognize the species of its foster parents so that it can lay its own eggs in the nests of that species. In this learning process the cuckoo is guided by instinct to ignore a world full of distracting information in order to focus on the details that must be memorized.

will learn to salivate at the sound of the bell. The important factors in classical conditioning are the unconditioned stimulus (the innately recognized cue, equivalent to the ethological sign stimulus, which in this case is food), the unconditioned response (the innately triggered behavioral act, equivalent to the ethological motor program, which in this case is salivation) and the conditioned stimulus (the stimulus the animal is conditioned to respond to, which in this case is the bell). Early behaviorists believed any stimulus an animal was physically capable of sensing could be linked, as a conditioned stimulus, to any unconditioned response.

In operant conditioning, the other major category of learning recognized by most behaviorists, animals learn a behavior pattern as the result of trial-and-error experimentation they undertake in order to obtain a reward or avoid a punishment. In the classic example a rat is trained to press a lever to obtain food. The experimenter shapes the behavior by rewarding the rat at first for even partial performance of the desired response. For example, at the outset the rat might be rewarded simply for facing the end of the cage in which the lever sits. Later the experimenter requires increasingly precise behavior, until the response is perfected. Early behaviorists thought any behavior an animal was physically capable of performing could be taught, by means of operant conditioning, as a response to any cue or situation.

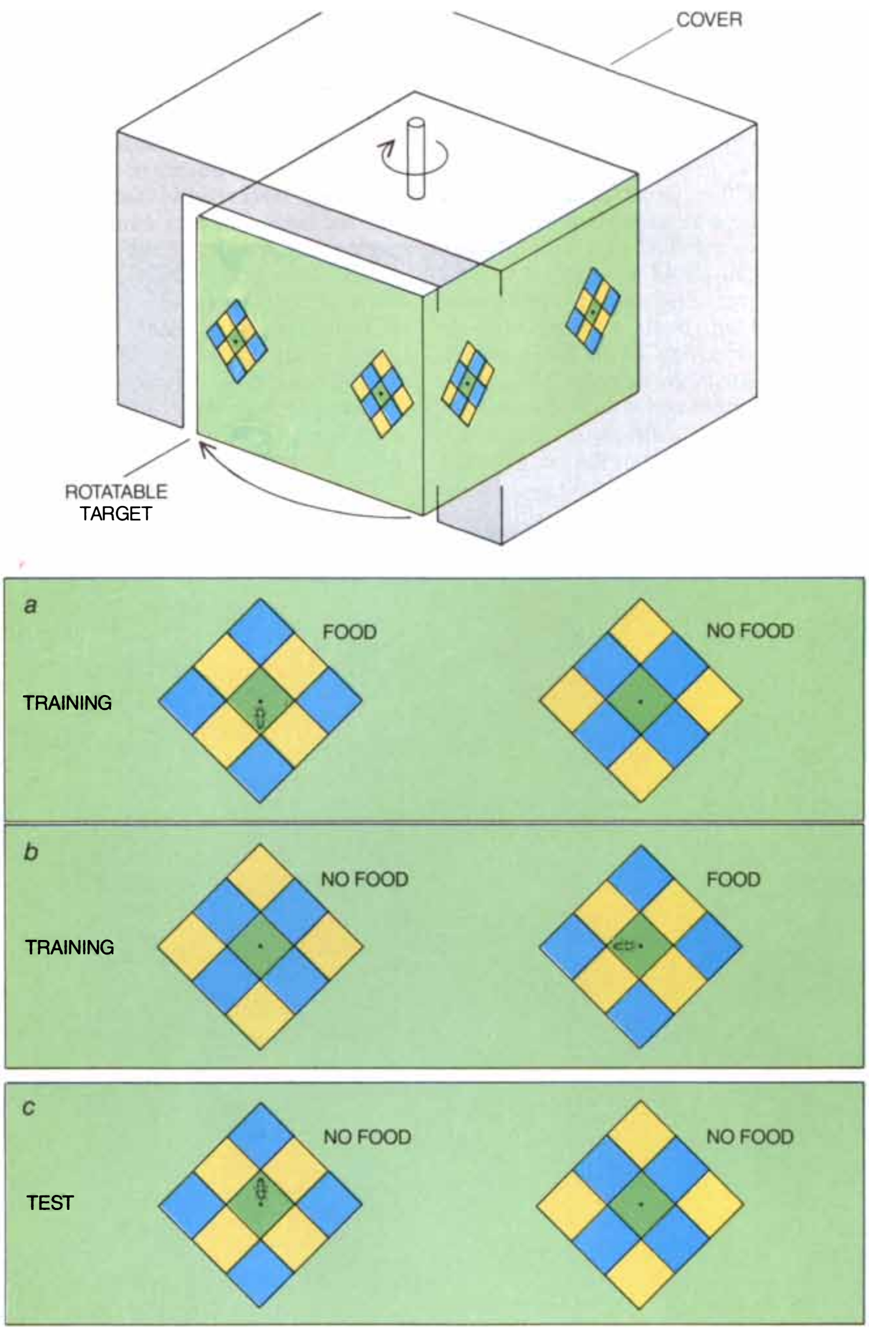
Challenges to Behaviorism

By 1970 several disturbing challenges to the behavioristic world view had appeared. The idea that any perceptible cue could be taught, by classical conditioning, as a conditioned stimulus was dealt a severe blow by John Garcia, now at the University of California at Los Angeles. He showed that rats could not associate visual and auditory cues with food that made them ill, even though they could associate olfactory cues with such food. On the other hand, he found that quail could associate not auditory or olfactory cues but visual ones—colors—with dangerous foods. Later work by other investigators extended these results, showing, for example, that pigeons readily learn to associate sounds but not colors with danger and colors but not sounds with food. The obvious conclusion was that these animals are predisposed to make certain associations more easily in some situations than in others.

The same kind of pattern was discovered in experiments in operant conditioning. Rats readily learn to press a bar for food, but they cannot learn to press a bar in order to avoid an electric shock. Conversely, they can learn to jump in order to avoid a shock but not in order to obtain food. Similarly, pigeons easily learn to peck at a spot for a food reward but have great difficulty learning to hop on a treadle for food; they learn to avoid shock by hopping on a treadle but not by pecking. Once

again it seems that in certain behavioral situations animals are innately prepared to learn some things more readily than others.

The associations that are most easily learned have an adaptive logic. In the natural world odor is a more reliable indicator than color for rats (which are notoriously nocturnal) trying to identify dangerous food; the color of a seed is a more useful thing for a pigeon to remember than any sounds the seed makes. Similarly, a pigeon is more likely to learn how to eat novel seeds if it experiments on food with its beak rather than with its feet. Animals that have innate biases concerning which cues they rely on and which procedures they attempt are more likely to ignore spurious cues, and they will learn faster than animals without inherent biases. The idea that animals are innately programmed to attend to specific cues in specific behavioral contexts and to experiment in particular ways in other contexts suggests a mutually reinforcing relation between learning and instinct. This relation helps to explain the once anomalous phenomenon of imprinting and to reconcile the approaches of behaviorists and ethologists.



TRAINING AND TESTING APPARATUS teaches bees to land on particular targets and checks their ability to remember the targets. Pairs of targets are arranged on each side of a rotatable box (*top*). The box is covered so that only one pair is visible at a time. Each target has a feeder at the center, which can be filled with sugar solution. To train the bee, the investigator supplies food to one target but not the other and lets the bee feed there (*bottom*, a). To keep the bee from simply memorizing the location of the target bearing food, the box is rotated so that the bee can also be trained on a second pair of targets that is a mirror image of the first (*b*). After the bee has been fed about 10 times, it is tested: the box is rotated to expose a pair of targets that both contain no food (*c*), and the bee is watched to see which it chooses to land on first. The apparatus can test the bee's ability to remember such characteristics as a target's color, pattern, shape and odor.

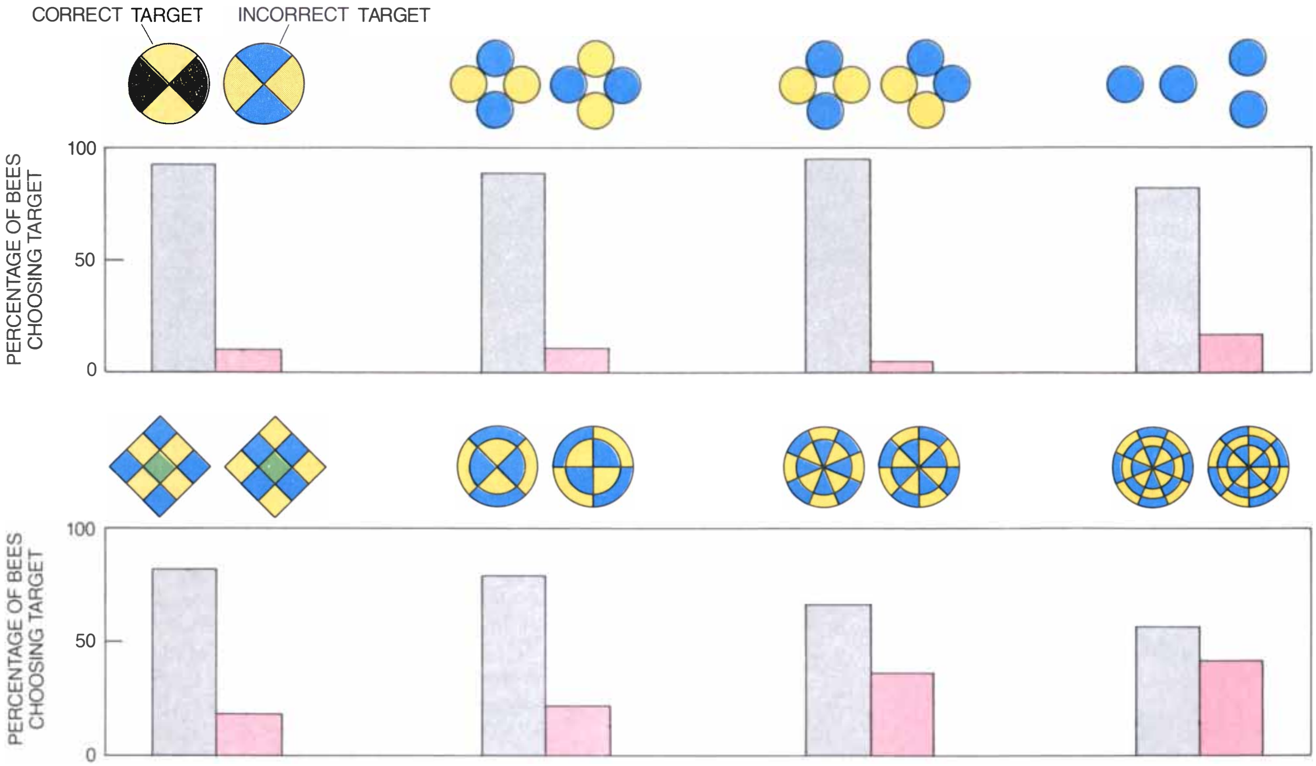
Instinctive Learning in Bees

The convergence of the two perspectives is illustrated by the ways honey bees learn about flowers. Bees inherently learn certain characteristics of a flower more easily than they learn others. Perhaps even more significant, once bees have acquired knowledge about a flower the ways in which they organize and refer to that knowledge are entirely instinctive.

Bees make their living by collecting nectar and pollen. Both of these essential foods are found in flowers, which offer them as a bribe to attract pollinating insects. Bees recognize flowerlike objects instinctively: they land spontaneously on small, brightly colored objects that have a high spatial frequency, or ratio of edges to unbroken areas (the spatial frequency of an object is high, for example, if the object has petals), and centers that (like the center of a flower) absorb ultraviolet light and so appear dark to bees.

Although bees recognize flowerlike objects innately, they have to learn which of those objects are likely to hold food. The initial flowerlike characteristics constitute an unconditioned stimulus: a set of sign stimuli. They trigger the unconditioned responses of landing and probing with the proboscis, behaviors that represent two innate motor programs. If a flowerlike object rewards a bee with food, the flower's specific characteristics may be learned—imprinted—as conditioned stimuli.

The first thing honey bees learn about a flower is its odor. Von Frisch showed early in his career that after a bee has been trained by being fed at a feeder that has a particular odor, it selects flowers of like odor from among hundreds of alternatives. Randolf Menzel of the Free University of Berlin showed that even one training visit is enough to teach the bee to choose the same odor 90 percent of the time in later visits; after only three training visits the rate of success is higher than



TARGETS provided evidence that bees remember a flower as a picture, not as a list of characteristics. The targets in each pair differ from each other only in the arrangement of their components (often one is merely a rotation of the other), and so they would be represented by identical lists of characteristics. Bees would be able to distinguish the targets in a pair only if they remember a target as a picture. Bees remembered the target they had been trained on except when it was very intricate; their memory may not have fine enough resolution to distinguish such targets. The targets shown are a few of many used in experiments.

98 percent [see "Learning and Memory in Bees," by Randolf Menzel and Jochen Erber; SCIENTIFIC AMERICAN, July, 1978]. Bees do not learn all odors with equal ease. Nonfloral odors take longer to learn, although it is unclear whether the bias results from insensitivity to inappropriate odors or from some problem in remembering them.

The next thing honey bees learn about a flower is its color. Menzel has shown that roughly three training visits to flowers of the same color are necessary before bees select that color over an alternative color 90 percent of the time. After about 10 training trips the bees choose the correct color more than 95 percent of the time. As with odors, bees do not learn all colors equally quickly, but enough is known about the vision of bees to rule out the possibility that this bias is based on an unevenness in the bees' ability to see different colors.

Honey bees also learn the shapes and color patterns of flowers, but they need more training visits in order to reach the level of 90 percent accuracy in remembering shape; about five or six visits suffice to enable them to distinguish a square "flower" (a plastic square containing a feeder) from a triangular one.

As with odor and color, bees inherently prefer some shapes over others. They particularly prefer busy patterns to simple ones. Until recently most investigators had thought bees do not remember a pattern as a picture (unlike human beings, and vertebrates in general) but rather as a list of defining characteristics, in much the same way as advertisements for real estate often depend not on photographs but on verbal lists: "Red Cape Cod, three bedrooms, two baths, detached garage." Such a list might enable bees to distinguish among species of flowers, and it would not require as large and complex a central nervous system as a picture memory would. Recent experiments by one of us (Gould) indicate, however, that bees do store low-resolution pictures of flowers.

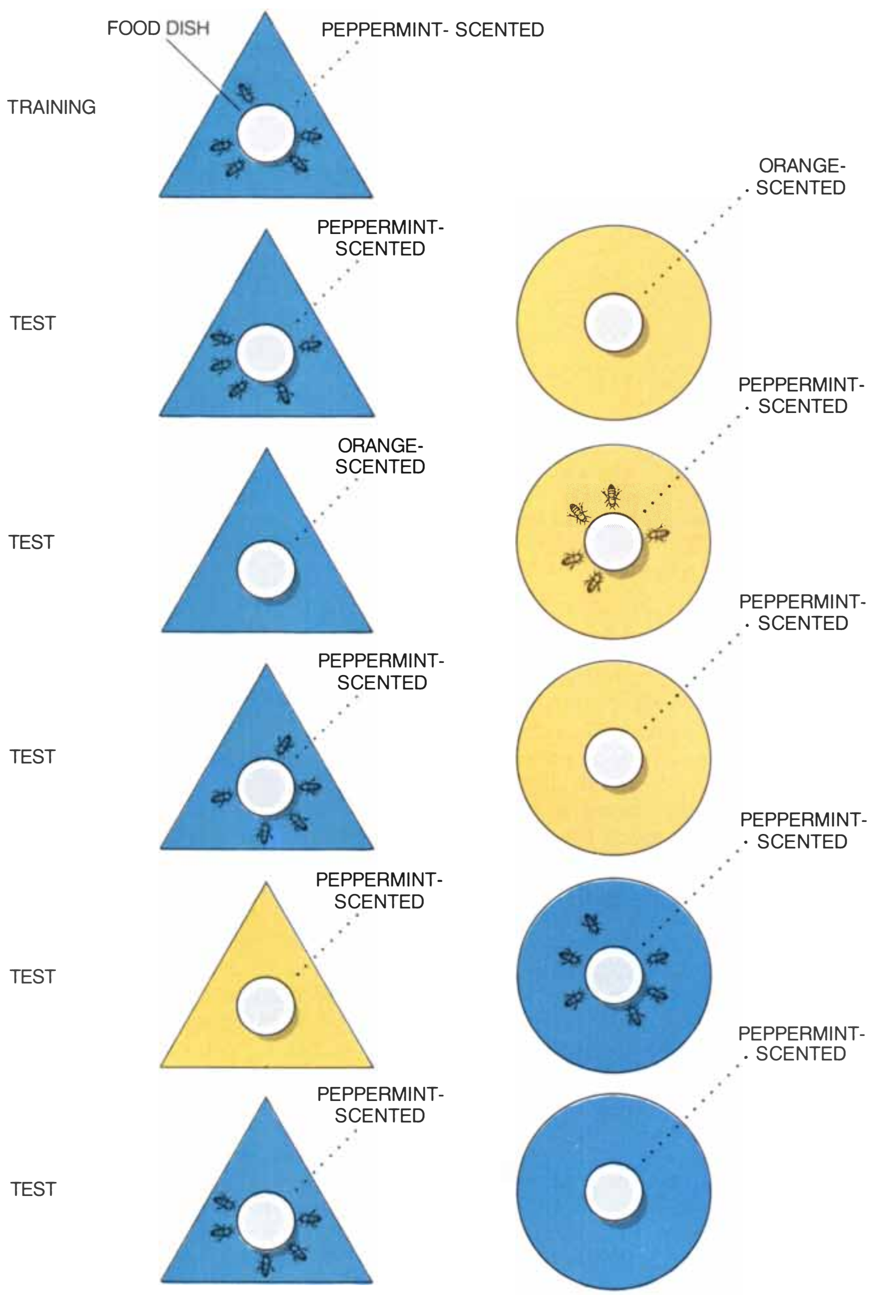
Bees learn many things about flowers, but there are some cues that cannot be stored as part of flower memory even though bees can learn them in other behavioral contexts. For example, honey bees are famous for their exquisite sensitivity to polarized light (by which they navigate), but they cannot learn the polarization patterns of flowers. They are also adept at learning which way a hive faces (to the point where rotating the hive by 90 degrees leaves most foragers unable to find the entrance until other bees provide strong chemical cues), but they will not learn in what direction a free-standing flower faces.

Organization of Bee Knowledge

The cues bees do remember about flowers, such as odor, color and pattern, are not remembered with equal weight. For example, if a bee that has been trained to feed at a peppermint-scented blue triangular target is presented with a choice between an orange-scented blue triangular target and a peppermint-scented yellow circular target, it will inevitably choose the peppermint-scented target even though that target has neither the color nor the shape the bee has been trained on. It is only when two targets have the same odor that bees pay much attention to color or shape; under those conditions color takes precedence over shape. This hierarchy corresponds to the relative reliability of the cues in nature. The odor of a flower is usually constant, whereas color can fade or appear different under different lighting conditions, and shape changes with damage from wind and herbivores, and even with viewing angle.

The hierarchy is an important factor in the organization of the bees' memory, but there is an even more important organizational element: the time of day at which each flower provides nectar. Bees learn the time at which food is available from each flower more slowly than they learn odor, color or shape, but once they have learned it, that knowledge serves to organize their use of the rest of their memory.

The organizational role of time was clearly shown by Franz Josef Bogdany of the University of Würzburg. For several days he trained a set of foragers to feed at two different feeders at differing times of day. From 10:00 to 11:00 A.M., for example, he fed them at a peppermint-scented blue triangular feeder; from 11:00 A.M. until noon he fed them at an orange-scented yellow circular feeder that was placed on the site the blue feeder had occupied. One day he put both feeders out at 9:00 A.M. and noticed an interesting pattern. Trained foragers began to appear at the blue feeder at about 9:45. They foraged exclusively at that feeder for about an hour. At roughly 10:45 some foragers began to shift to the yellow feeder, and by 11:15 the blue feeder—which was still full of food—was completely abandoned.



HIERARCHICAL STRUCTURE of bees' memory is revealed in a series of tests. Bees were trained to land on peppermint-scented blue triangles. Their training was confirmed by a test in which they preferred a peppermint-scented blue triangle to an orange-scented yellow circle, even when both targets bore full food dishes. Then they were offered an orange-scented blue triangle and a peppermint-scented yellow circle. They chose the latter, showing they are more likely to be guided by the memory of a scent than by the memory of a shape or color. Later tests, interspersed with controls, showed bees are more likely to rely on a remembered color than a remembered shape. The structure of bees' memory shows that instinct guides a bee's use of the knowledge it gains through learning.

Bees behave as though they have an appointment book by which they schedule their visits; no more than one entry can be made for any specific time. The resolution of the book is about 20 minutes; that is, bees cannot remember two separate appointments if they are less than 20 minutes apart. Bees have been able to remember as many appointments as experimenters have tried to teach them. The standing record, set by R. Koltermann of the University of Frankfurt, is nine appointments in eight hours.

Another experiment by Bogdany shows finer details in the structure of honey-bee memory. After days of being trained to the peppermint-scented blue triangular feeder, bees were presented with an orange-scented blue triangular feeder. The foragers learned the new odor in one visit, but they completely forgot the color and shape, even though these characteristics had not been changed. On the other hand, when bees were trained to an odorless blue triangle and then presented with a peppermint-scented blue triangle, they learned the new odor without forgetting the color and shape. Apparently the appointment book has an entry for each cue; the entries are structured in such a way that blanks can always be filled in but that if even one item is changed, the entire entry is erased.

Such results suggest that honey bees, guided to particular targets by instinctively recognized cues, memorize certain specific features of the targets and store that memory in a "prewired," hierarchical memory array. The cues that are memorized, the speed with which each cue is memorized and the way the memorized data are stored are all innate characteristics of the bee.

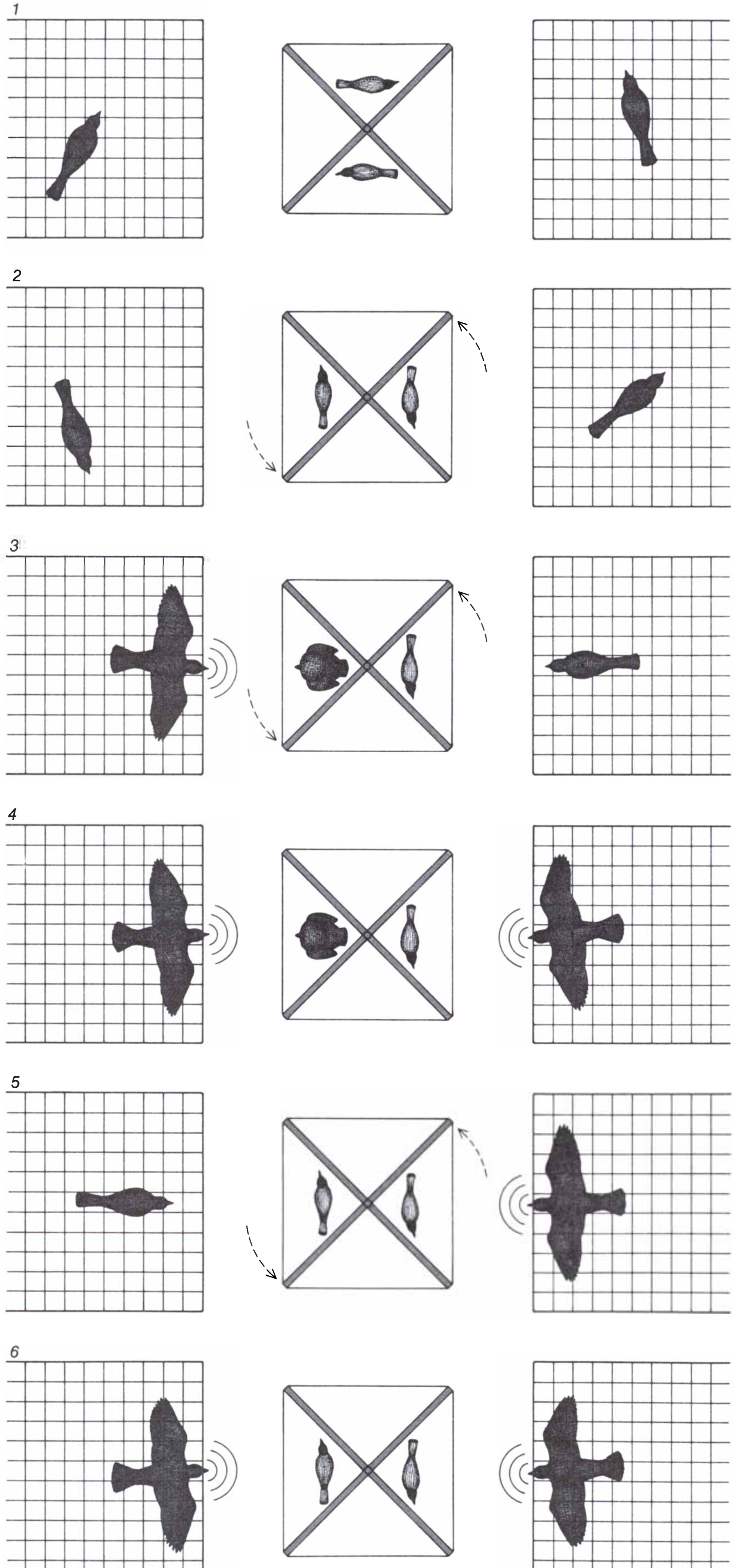
Learning about Enemies

Animals need to learn many things other than how to find food. For example, they must learn how to recognize and respond to various kinds of predators and enemies. For some animals it is enough merely to identify a very general class of predators. Flying moths and crickets automatically begin evasive maneuvers when they hear the high-pitched sounds characteristic of hunting bats. Other animals must be able to make finer distinctions among friends and potential foes. Nesting birds are a particularly apt example. They must learn to distinguish harmless birds, such as robins, from birds such as crows and jays, which hunt for eggs and nestlings. The "fill in the blanks" strategy adopted by bees as they learn about flowers is also applicable to this kind of learning.

When nesting birds detect nest predators, they attack en masse, a phenomenon commonly known as "mobbing." How do birds know whom to mob and whom to ignore? Eberhard Curio of the University of the Ruhr has shown that the process of learning which species to mob is innately guided.

In Curio's experiments groups of birds (most often European blackbirds) were kept in separate cages. Between the cages was placed a rotatable box with four compartments. At any given time the birds in one cage could see into only one compartment of the rotatable box, while the birds in the other cage saw a different compartment [*see illustration at right*]. The birds could see into each other's cages.

Curio began by rotating the central box to present a stuffed Australian honeycreeper—a harmless species—to each cage. The live birds showed no reaction. He then put a stuffed owl in one compartment and a honeycreeper in the opposite compartment. When the box was rotated so that each model was in view of one of the two sets of birds, the birds in the cage exposed to the owl began to emit the species' innate mobbing call and tried to attack the model. The other group observed the mobbing for a moment and then, responding to this powerful set of sign stimuli, began trying to attack the stuffed honeycreeper, at the same time emitting the mobbing call. On later occasions this group of birds always tried to mob honeycreepers, a species they had never seen attack a nest. Cu-



MOBBING BEHAVIOR of the European blackbird demonstrates the instinctive way the species learns to recognize predators. Between the cages in which the birds sit is a rotatable, four-chambered box (*1*). Each bird can see only one chamber of the box, but it can also see into the other bird's cage. First each bird is shown a stuffed Australian honeycreeper, a harmless species (*2*); neither bird shows any interest. Then one bird is shown a stuffed owl (a predator of small birds) and the other is shown a honeycreeper. The bird shown an owl tries to chase it away and gives the characteristic "mobbing call" (*3*). The other bird at first watches and then (*4*) joins in the mobbing behavior. It has learned to mob honeycreepers. When both birds are shown honeycreepers (*5*), it teaches the other to mob honeycreepers as well (*6*).

rio found that the baseless aversion to honeycreepers was passed on from generation to generation. The young birds learned to mob honeycreepers by watching their parents. In later experiments Curio was able to teach his birds to mob bottles of laundry detergent.

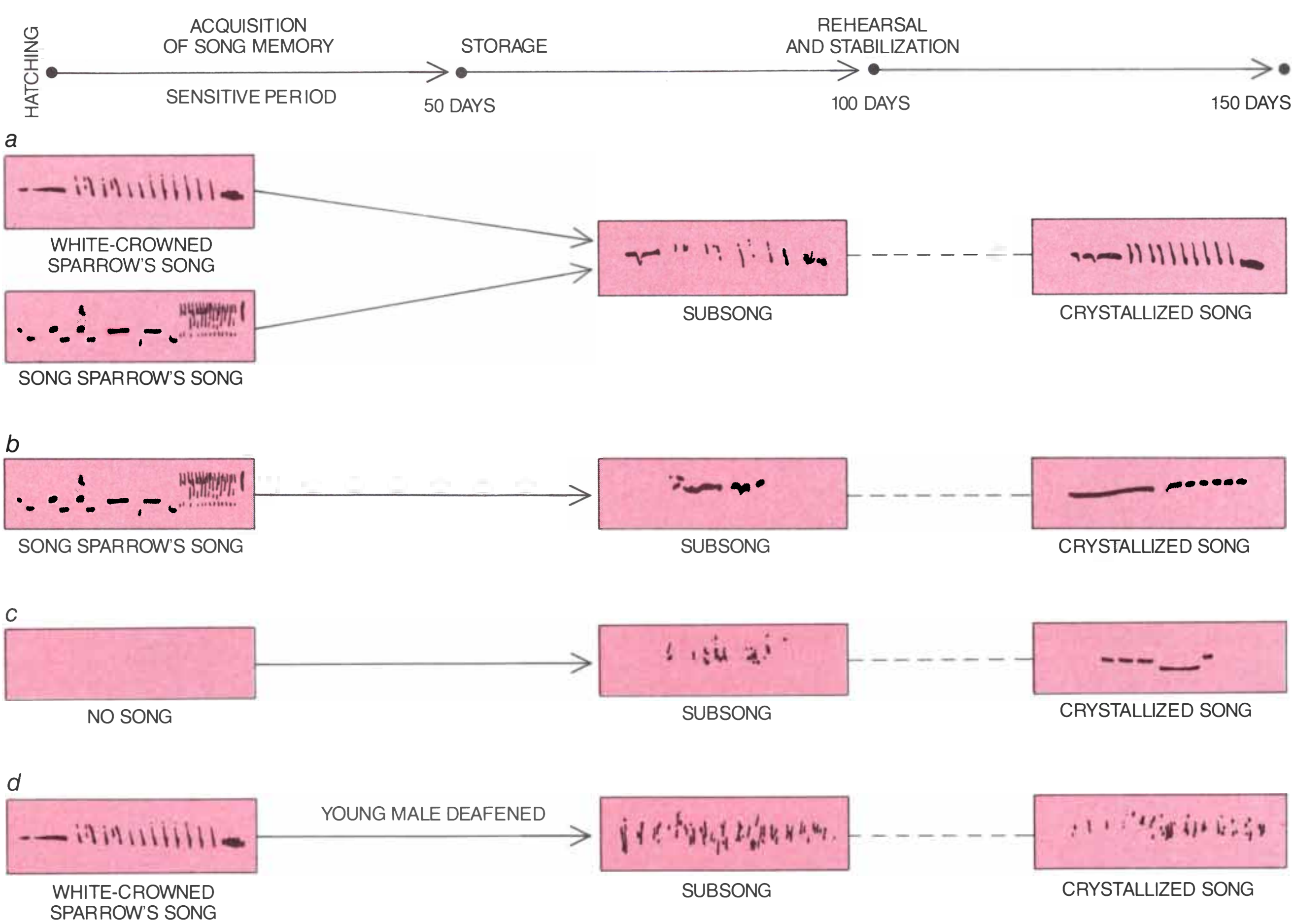
There is good reason to think variations of this strategy of learning about enemies are at work in many mammalian species as well as in birds. Perhaps the most elaborate version is found in vervet monkeys. As was shown by Robert M. Seyfarth, Dorothy L. Cheney (both now at the University of Pennsylvania) and one of us (Marler), vervets have special alarm calls for each of four kinds of predators: aerial predators such as eagles, four-legged predators such as leopards, predatory primates such as baboons, and snakes. Each alarm call elicits a different kind of response. For example, an eagle alarm sends vervets on the ground toward cover and causes those in the exposed tops of trees to drop like stones into the protective interior, whereas a snake call is ignored by vervets in trees but induces those on the ground to rear up on their hind legs and scan the ground around them.

Young vervets instinctively emit alarm calls in response to a wide but specific range of stimuli. For example, any object within certain size limits moving in free space at a certain angular velocity causes the young monkeys to give the eagle call; the call can be elicited by a stork or even a falling leaf. With time the infants learn which species cause the adults to call. Hence vervets growing up in one region might learn to give alarm calls on seeing baboons, leopards and a certain species of eagle, whereas those in another region might react to human beings, hunting dogs and a certain species of hawk. Like the bees' system of learning about flowers, this innate system is efficient for learning essential information about predictably unpredictable situations: predictable kinds of threat posed by animals whose exact species cannot be predicted.

Song Learning in Birds

Another task an animal must perform that often requires learning is recognizing others of its own species. Perhaps the richest and best understood use of learning in species recognition is the learning of songs by birds. All birds have a repertoire of perhaps one or two dozen calls that are innately produced and recognized. These calls need not be learned and can be produced even by birds hatched and reared in isolation. Several kinds of birds also have more complex vocal patterns—songs for attracting mates and defending territory—that must to some extent be learned from adults of the same species.

The white-crowned sparrow, which has been studied extensively by one of us (Marler), is a good example. Adults of this species produce a three-part or



SONG LEARNING in the white-crowned sparrow exhibits great specificity: young male birds can instinctively identify and preferentially learn the song of their own species. If a young male white-crown is played tape recordings of adult white-crown song and adult song-sparrow song (*a*), it first begins a period of experimentation, known as subsong, and then produces a crystallized song very similar to the white-crown song it has heard. If it is played only a tape recording of song-sparrow song (*b*), it will not learn the song: it still goes through subsong, but its final, crystallized song does not resemble either the song-sparrow song or the white-crown song. A bird that is played no song (*c*) also learns nothing. If the young bird hears a white-crown song but is deafened before subsong begins (*d*), it is unable to learn how to produce the song it heard; it produces an amorphous song with no melodic structure.

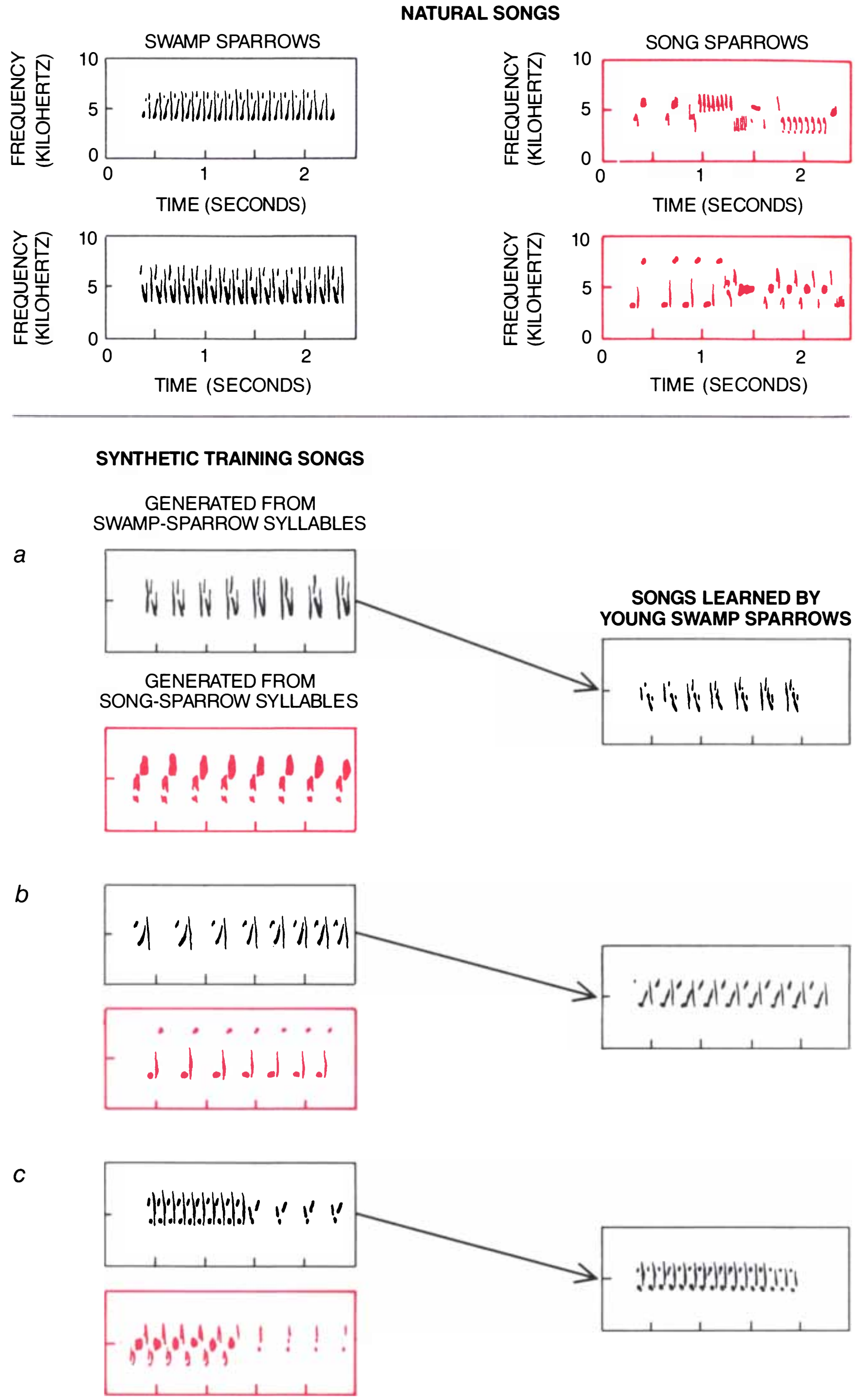
four-part courtship song rich in melodic detail. Different individuals produce recognizably different songs, but the organization of the song is common to the species. The song produced by each male white-crown is similar to (but not identical with) the songs heard near the place it was reared. (There are actually local dialects.)

Experiments in manipulating the sensory experience of young sparrows have revealed much about the organization of the process of song learning. A bird kept in auditory isolation, for example, begins to produce and experiment with song notes by the time it is about a month old. This period of experimentation, known as subsong, waxes and wanes for roughly two months. By about the bird's 100th day it "crystallizes" its song into a form that will not change significantly; the song is highly schematic, but it bears many of the basic features of normal adult white-crown song. Such experiments show that the chick is born with a basic innate song, which it learns to elaborate when it is raised in the wild.

In another experiment we play tape-recorded songs of other species to isolated young birds. Such songs have little effect on the final, crystallized song produced by the bird (although Luis Baptista of the California Academy of Sciences has shown that a live tutor can sometimes successfully indoctrinate young white-crowns). On the other hand, when we play a medley of tape recordings, one of which reproduces a real white-crown song, the young male somehow manages to pick out the white-crown song and learns to produce a tolerable imitation. If it is to produce a perfect imitation, the bird must hear the song before it is about seven weeks old. (The actual period varies with experimental conditions.) The "window" for learning (the time in which the drive to learn is high) is called the sensitive period.

Taken together, these experiments show that learning of the local song dialect in the white-crowned sparrow is innately controlled, irreversible and restricted to a sensitive period; these are exactly the characteristics of classical imprinting. Starting with some form of innate song, the young male depends on innately recognized cues to trigger the learning process and later learns to imitate the more elaborate, memorized song.

The actual process of developing the bird's own version of the local dialect seems to be done by trial-and-error learning. In the case of the white-crowned sparrow, weeks or months may pass between the end of the sensitive period (during which the bird memorizes the song) and the bird's



SYNTHETIC SONGS identified the cues young swamp sparrows rely on in identifying their own species' song. The natural swamp-sparrow song (*top left; two songs are shown because the exact song varies from bird to bird*) consists of one syllable repeated at a steady rate. The song-sparrow song (*top right*) contains several syllables, beginning with an accelerating trill. In one pair of synthetic songs (*a*) both songs had the steady rate of swamp-sparrow songs but one was constructed from swamp-sparrow syllables and the other from song-sparrow syllables. As expected, the swamp sparrows learned the song made up of swamp-sparrow syllables. In another pair (*b*) each song still consisted of a single syllable, but the syllables were repeated at an accelerating rate; in one song the syllable was from swamp-sparrow song and in the other it was from song-sparrow song. Swamp sparrows learned the swamp-sparrow syllable and restructured the tempo to match the natural swamp-sparrow song. In another pair (*c*) each song consisted of two syllables, arranged as they would be in a song-sparrow song; again one song was made up of swamp-sparrow syllables and the other of song-sparrow syllables. Young swamp sparrows learned one of the syllables in the song generated from swamp-sparrow syllables and delivered that syllable at their characteristic rate. The experiments show young swamp sparrows rely on cues within individual syllables when determining which song to learn.

first experiments with recognizable imitations, which come at the end of subsong. Masakazu Konishi, now at the California Institute of Technology, has demonstrated that when a young male white-crown that has been exposed to white-crown song is deafened before any crystallization, it never sings anything melodic—not even its innate song. Apparently a bird must experiment with its beak, syrinx and pulmonary muscles, listening to the sounds that result from various manipulations and trying to match them to its mental record. During the progression from subsong to crystallization the bird shapes its song to match the record stored in its brain, whether that record consists of only the rough, innate song or of a song memorized during the sensitive period. Konishi found that by the time the bird has crystallized its song the singing has become so routine that deafening has little or no effect.

The Cues within Adult Birdsong

How do the young songbirds pick out their own species' song from a world full of sounds? What specific cues do the sparrows rely on to determine which song to learn? One of us (Marler) and his colleagues have investigated this question in experiments with swamp sparrows and song sparrows, two species that nest within earshot of one another.

Of the two species, swamp sparrows have the simpler song. It consists of a single series of regularly repeated syllables; the kinds of syllables vary from theme to theme, bird to bird and region to region. The song sparrow's song is more complex: it consists of at least four types of syllables, often beginning with an accelerating trill. Although the innate songs of the two species reflect some of these structural differences, the syllables from which the innate songs are constructed are much simpler.

The auditory cues that guide learning might include elements of the syllables themselves, elements of the tempo and phrase structure or elements of both. As a first stage in determining which of these elements are important, we gave hand-reared sparrows of both species a chance to learn from tapes of their own species or tapes of the other species. As we had expected, the birds learned almost exclusively from tapes of their own species. The rare cross-species imitations are important, however: they show that songs of other species can physically be sung, and that the normal tendency not to learn the song of another species comes from the birds' inattentiveness to such songs rather than from inability to produce them.

To investigate further the role of different aspects of song structure in learning preferences, we put together synthetic training songs that varied in structure and played them to young birds of both species. For example, we fabricated one song from a slow, steady repetition of song-sparrow syllables and another from a slow, steady repetition of swamp-sparrow syllables. The tempo of these songs was like that of swamp-sparrow song. Swamp sparrows readily learned the steady song with swamp-sparrow syllables but not the other song, and song sparrows learned only the song containing song-sparrow syllables. These results indicated that cues lie within syllables.

By itself that experiment did not indicate whether or not tempo and phrase structure might be important as well. To investigate this issue we synthesized a variety of other songs. The new songs were created from syllables of both species, but the syllables were organized in a variety of patterns. Some of the patterns were like those of the swamp sparrow and others were like those of the song sparrow.

One of the songs, for example, was made up of swamp-sparrow syllables but had the accelerating rate of delivery characteristic of song sparrows. Would young swamp sparrows reject the song because it had the wrong tempo? Or would they accept it because it had the correct sign stimulus for learning (the right syllables) and so sing an abnormally paced song? In fact, the young swamp sparrows did neither. They learned the syllables of the song, but in their own singing they actually changed the tempo, so that they delivered the learned syllables at the constant rate typical of their own species.

In another variation we fabricated songs even more similar in structure to song-sparrow songs. Each song had two segments; each segment consisted of a different type of syllable, and in one segment the syllables were delivered at an accelerating rate. When the songs were made up of different kinds of swamp-sparrow syllables, young swamp sparrows learned to sing a steady repetition of one of the two syllable types, regardless of the temporal pattern in which the syllables had been presented. Swamp sparrows thus seem to focus entirely on the syllabic structure when searching for cues, paying scant attention to the organization of the song as a whole.

Song sparrows are different. They are readier to accept the alien syllables of the swamp sparrow if the syllables are presented in song models that have complex phrase structures (although they reject swamp-sparrow syllables when they are delivered at a steady tempo). The two attributes—syllable type and syntactic structure—apparently have additive effects.

These experiments show that although the song sparrow and the swamp sparrow are closely related, the innate mechanisms that control learning in the two species are different. No doubt the white-crowned sparrow is different in its own way.

Speech Learning in Humans

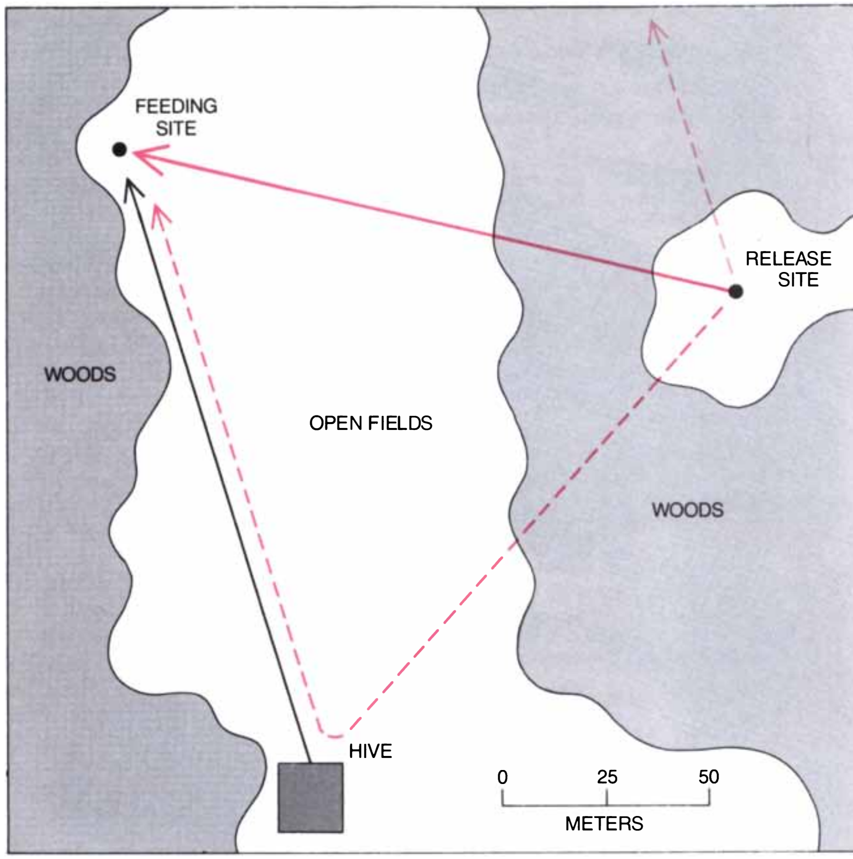
The learning of songs in birds has a number of parallels with the learning of speech in human beings. In swamp sparrows, song learning involves the innate recognition of certain elements in the species-specific syllables. There is now abundant evidence that human infants innately recognize most or all of the more than two dozen consonant sounds characteristic of human speech, including consonants not present in the language they normally hear [see "The Perception of Speech in Early Infancy," by Peter D. Eimas; SCIENTIFIC AMERICAN, January, 1985]. The innate ability to identify sign stimuli present in consonants confers several advantages: it allows the infant to ignore a world full of irrelevant auditory stimuli in order to focus on speech sounds, it starts the child on the right track in learning to decode the many layers of meaning buried in the immensely complex and variable sounds of speech, and it provides an internal standard for the child to use in judging and shaping speech sounds.

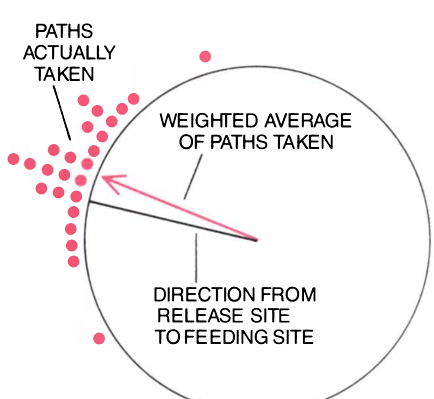
Another aspect of human speech learning parallels the subsong phase, the period during which birds of species that learn to sing begin to experiment with sound production. The subsong phase begins right on schedule even if the bird has been deafened (although such birds learn nothing from their vocal experimentation). Human infants also have a phase of babbling, in which they develop, through trial-and-error learning, the ability to produce the set of consonants found in their own language. As with birds, babbling begins and ends on schedule even in deaf children.

Birds have an innate, templatelike structure that specifies the rules for producing syllables in song. On a vastly different scale, there is good reason to believe that the rhythms in which words and sentences are assembled in speech and the set of rules known as grammar (in particular the division of words into such categories as nouns, verbs, adjectives and adverbs) are at some deep level also innate. This idea, argued most persuasively by Noam Chomsky of the Massachusetts Institute of Technology, helps to explain why the learning of speech proceeds so easily compared with the learning of such inherently simpler tasks as addition and subtraction.

Sophisticated Learning

Although much of animal learning (and probably more of human learning than is yet suspected) is innately guided by learning programs, much human behavior clearly cannot be explained so simply. For example, imagining a solution before exploring it physically is a behavior outside the two traditional forms of learning originally studied by behaviorists. This kind of cognitive learning, called cognitive trial and error, comes much closer than programmed learning to





ABILITY OF BEES to construct mental maps, rather than simply remembering routes as strings of landmarks, was demonstrated in experiments in which bees were trained to feed at a specific site (*top*). After the training period a number of bees were caught as they left the hive to go to the feeding site. They were transported, in the dark, to another site (*right*) and released. If the bees had not noted their new surroundings, they might have flown off at the compass bearing that would normally have taken them from the hive to the feeding site (*light colored broken arrow*). If bees remember routes simply by remembering strings of landmarks, they might have followed landmarks back to the hive and flown from there to the feeding site (*dark colored broken arrow*). Instead the bees were able to use their knowledge of the area as a whole to devise a new route from the release site directly to the feeding site (*solid colored arrow*). The bees' ability to form maps indicates that even creatures with limited mental faculties can be instinctively equipped to manipulate and relate separate bits of learned information.

one's intuitive sense of what intelligence is. It requires the ability to recall and combine separate bits of learned information and from such mental recombination to formulate new behavioral solutions.

The first evidence that animals have such an ability came in 1948 in a series of experiments by Edward C. Tolman of the University of California at Berkeley. In one experiment Tolman allowed rats to explore a maze that had two alternative goals, a white box and a black box, both containing food; rats learned the routes to both boxes, and chose them with equal frequency. Later Tolman took the rats to another room, where a black box and a white box were put side by side, and he gave them a shock when they entered the black box. When the rats were released in the maze the next day, they entered only the white box. Tolman concluded that they had combined information from the two entirely different experiments, generalizing about black boxes and remembering that one route led to a black box. Tolman also found that rats have an ability to form mental maps of familiar areas and from the maps to plan novel routes. Tolman's finding was subsequently confirmed and explored by David S. Olton of Johns Hopkins University [see "Spatial Memory," by David S. Olton; SCIENTIFIC AMERICAN, June, 1977].

The ability to form maps is by no means limited to rats and human beings. Emil W. Menzel of the State University of New York at Stony Brook has found the same ability in captive chimpanzees; John R. Krebs of the University of Oxford and Sara J. Shettleworth of the University of Toronto have shown that seed-caching birds can form cognitive maps registering the locations of hundreds of hidden seeds [see "Memory in Food-hoarding Birds," by Sara J. Shettleworth; SCIENTIFIC AMERICAN, March, 1983].

Mental Maps and Categories

In an effort to determine how common this sophisticated cognitive ability is, one of us (Gould) investigated whether honey bees also have mental maps. When bees travel a familiar route, they rely on prominent landmarks. The usual explanation of how bees use landmarks is that they remember the series of landmarks encountered en route to each site and can refer to the landmarks only in the same way as Hansel and Gretel would have referred to their trail of breadcrumbs. In that case bees would have no idea how one set of landmarks leading to one site is related spatially to the set that leads to any other site.

We tested whether bees do navigate this way or whether they in fact put landmarks in the context of a mental map of their home area. We trained certain bees to feed in one area and then, on subsequent days, captured them as they flew from the hive to that area and carried them (in the dark) to another location.

We thought that when the bees were released in the new area, they might adopt any of several courses of action. They could be disoriented and fly at random. Alternatively, they could fail to understand that they had been displaced; they would then fly off at the same compass bearing they would normally follow to get to the food site from the hive. If bees can navigate only by specific strings of landmarks, they might recognize the landmarks around the new location as part of a route different from the one they were on when they were captured; they would then follow that route back to the hive and from there fly to the food site. Finally, if bees do have cognitive maps, they should be able to determine where they were in relation to the food site and to select the appropriate bearing to reach the food, even though they had never flown from the hive to the food site before by such a roundabout route.

We found that bees overwhelmingly followed the last of these alternatives: when the area to which they were displaced was within their home area (the four or so square kilometers immediately surrounding the hive), they flew directly to the food site. It seems, then, that for bees cognitive map making is an innate part of route learning.

Another sophisticated ability involved in the process of learning is the formation of abstract concepts and categories. Is this ability found in animals? One suggestive hint comes from the work of Richard J. Herrnstein of Harvard University. He showed thousands of slides to laboratory-reared pigeons and rewarded them when they pecked at slides in which some specific kind of object, say a tree, was in the picture; these birds, of course, had never seen a real tree. The birds learned the task remarkably quickly, which suggests that they had a strong innate disposition to form generalized conceptual categories. When they were later tested with slides showing new species of trees, the birds reliably picked out the slides with trees, including some slides the experimenters had at first thought were treeless. The birds' occasional errors were also revealing: they sometimes identified telephone poles and television antennas as trees.

Students of human language acquisition have long known that children automatically form conceptual categories for the new words they learn. Chairs, tables and lamps are organized into a "furniture" category and the category of "chairs" is subdivided into subordinate categories such as rocking chairs and armchairs. Such categorization is essential to the rapid acquisition of words, and word storage in the brain is probably organized as a categorized filing system. The effects of small strokes, which can kill small regions of the brain, seem to reflect such a system: their victims sometimes lose an entire category of words, for instance the names of flowers.

It seems reasonable to propose that the drive to categorize is innate in at least some species. Perhaps it is the ability to make and manipulate categories that underlies the ability of animals to perform cognitive trial and error: to evaluate alternatives and formulate simple plans.

A New Synthesis

The emerging picture of learning in animals represents a fundamental shift from the early days of behaviorism, when animals were supposed to be limited to learning by classical conditioning and operant conditioning and were expected to be able to learn any association or behavior by those processes. It is now understood that much learning, even though it is based on conditioning, is specialized for the learning of tasks the animal is likely to encounter. The animal is innately equipped to recognize when it should learn, what cues it should attend to, how to store the new information and how to refer to it in the future. Even the ability to categorize and perform cognitive trial and error, a process that may be available to the higher invertebrates, may depend on innate guidance and specialization—specialization that enables the chickadee, with its tiny brain, to remember the locations of hundreds of hidden seeds, whereas human beings begin to forget after hiding about a dozen.

This perspective allows one to see that various animals are smart in the ways natural selection has favored and stupid where their life-style does not require a customized learning program. The human species is similarly smart in its own adaptive ways and almost embarrassingly stupid in others. The idea that human learning evolved from a few processes, which are well illustrated in other animals, to fit species-specific human needs helps to bring a new unity to the study of animal behavior and a new promise for understanding human origins.

Power from the Sea

Electricity can be generated by exploiting the temperature difference between warm surface water and cold bottom water. The technology will be useful when oil supplies become scarce

by Terry R. Penney and Desikan Bharathan

The ocean is the world's largest collector of solar energy. On an average day the 60 million square kilometers of tropical seas absorb an amount of solar radiation equivalent in heat content to about 250 billion barrels of oil. Even with the current glut of oil, the goal of converting a fraction of that energy into electricity continues to be pursued in several countries. The technology is called OTEC, an acronym for ocean thermal-energy conversion. If a multinational array of OTEC installations converted less than a tenth of 1 percent of the energy stored as heat in tropical surface waters, they would generate at least 14 million megawatts, more than 20 times the current generating capacity of the U.S. The same technology could also provide fresh water, refrigeration, air conditioning and a medium for culturing fish.

The source of thermal energy in an OTEC system is simply warm seawater. How can such a source generate electricity? There are two ways. In one of them the warm water evaporates a working fluid that has a low boiling point. In the other the seawater itself boils in a vacuum chamber; by lowering the pressure the chamber reduces the boiling point of the water. In each case the resulting vapor drives a turbine to generate electricity. Cold water drawn up from depths of some 600 to 1,000 meters (depending on the site) then condenses the working fluid to complete the cycle. As long as a temperature difference of about 20 degrees Celsius exists between the warm upper layer and the deep water, useful amounts of net power can in principle be generated. The plant can be on land, just offshore or on a ship that moves from place to place. The electricity can be delivered to a utility grid or can be employed locally to manufacture such products as methanol, hydrogen, refined metals and ammonia.

Two types of OTEC system are within reach of efficient power generation and ultimate commercialization. One of them (the one with the low-boiling-point fluid) operates as a closed cycle, the other (the low-pressure system) as an open cycle. In a closed-cycle operation the working fluid is enclosed in the system and recycles continuously, like the working fluid in a refrigerator. In an open-cycle system the working fluid is a constantly changing supply of seawater.

The closed cycle's requirement for a working fluid with a low boiling point calls for a fluid such as ammonia or Freon. The fluid is pumped through a heat exchanger (the evaporator), where it is vaporized by the warm seawater pumped in through a warm-water pipe. The vapor drives a turbine attached to an electric generator.

Vapor discharged from the turbine at a low pressure passes through a second heat exchanger (the condenser) cooled by water pumped up from the depths through a cold-water pipe. Pumps send the condensed working fluid back to the evaporator to begin the cycle again.

In a typical open-cycle system the working fluid is warm seawater. It boils violently in the vacuum chamber, producing low-density steam. One can visualize the process as an extension of the phenomenon that causes water to boil at lower temperatures with increasing altitude. The vacuum chamber enables an open-cycle system to operate at pressures equivalent to those between 27 and 30 kilometers above the earth. Such low pressures create unique problems, including the tendency of dissolved gases to evolve out of the seawater. Since the gases cannot be condensed, they can render the system inoperable unless they are continuously removed.

Less than .5 percent of the incoming warm water is turned into steam in such a plant. Hence large amounts of water must be pumped through the plant in order to generate enough steam to run the large, low-pressure turbine. To complete the cycle cold seawater condenses the steam. The

INTEGRATED SYSTEM based on ocean thermal-energy conversion (OTEC) is depicted. An electric power plant has three large pipes running out to the sea: a warm-

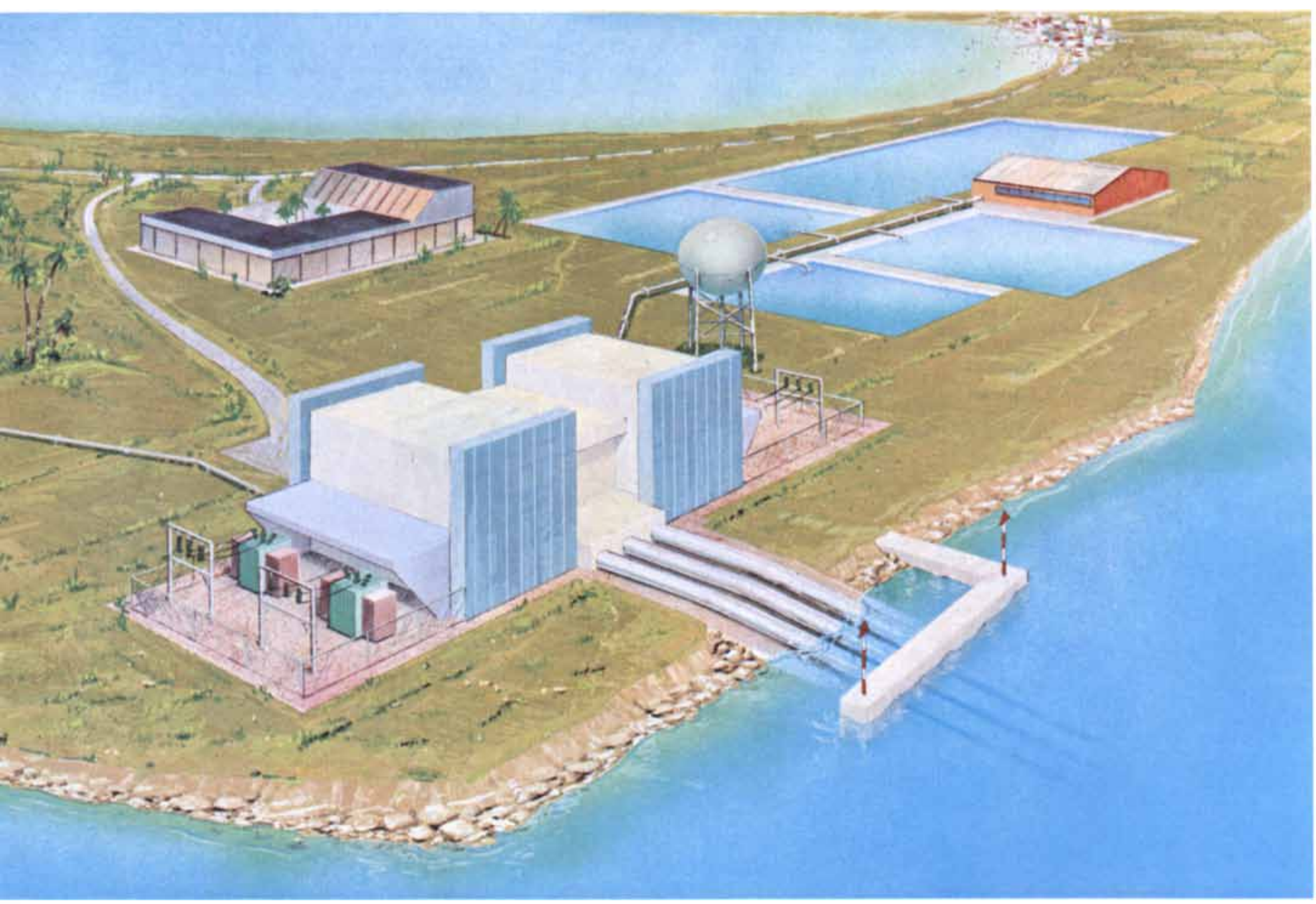
condensation can be achieved directly by mixing the cold seawater with the steam or indirectly in a surface condenser. In such a condenser the vapor and the coolant (cold seawater) are physically separated by walls; the heat must be transferred through the walls, which therefore are best made of metal because of its superior ability to transfer heat. With this system the condensed steam is free of seawater impurities, so that the operation has the added benefit of producing desalinated water.

The idea of tapping thermal energy from the ocean was put forward more than a century ago by Jacques Arsène d'Arsonval, a French engineer. He envisioned a closed-cycle system but never tested it. In 1926 his friend and former student Georges Claude—already well known for inventing the neon sign—became obsessed with what was to be a lifelong goal: making OTEC a reality. He designed an open-cycle system and tested it at Matanzas Bay in northern Cuba in 1930. The system generated 22 kilowatts of power but consumed more than that in operating. If Claude had pumped water through his cold-water pipe (1.6 meters in diameter and two kilometers long) at a higher velocity, had used a larger turbine and had exploited the greater temperature difference (24 degrees C.) available at Santiago in southern Cuba, he probably could have generated as much as two megawatts of net power. As it was, the experiment did demonstrate that cold water could be brought to the surface from a depth of more than 700 meters with low frictional losses in the pipe.

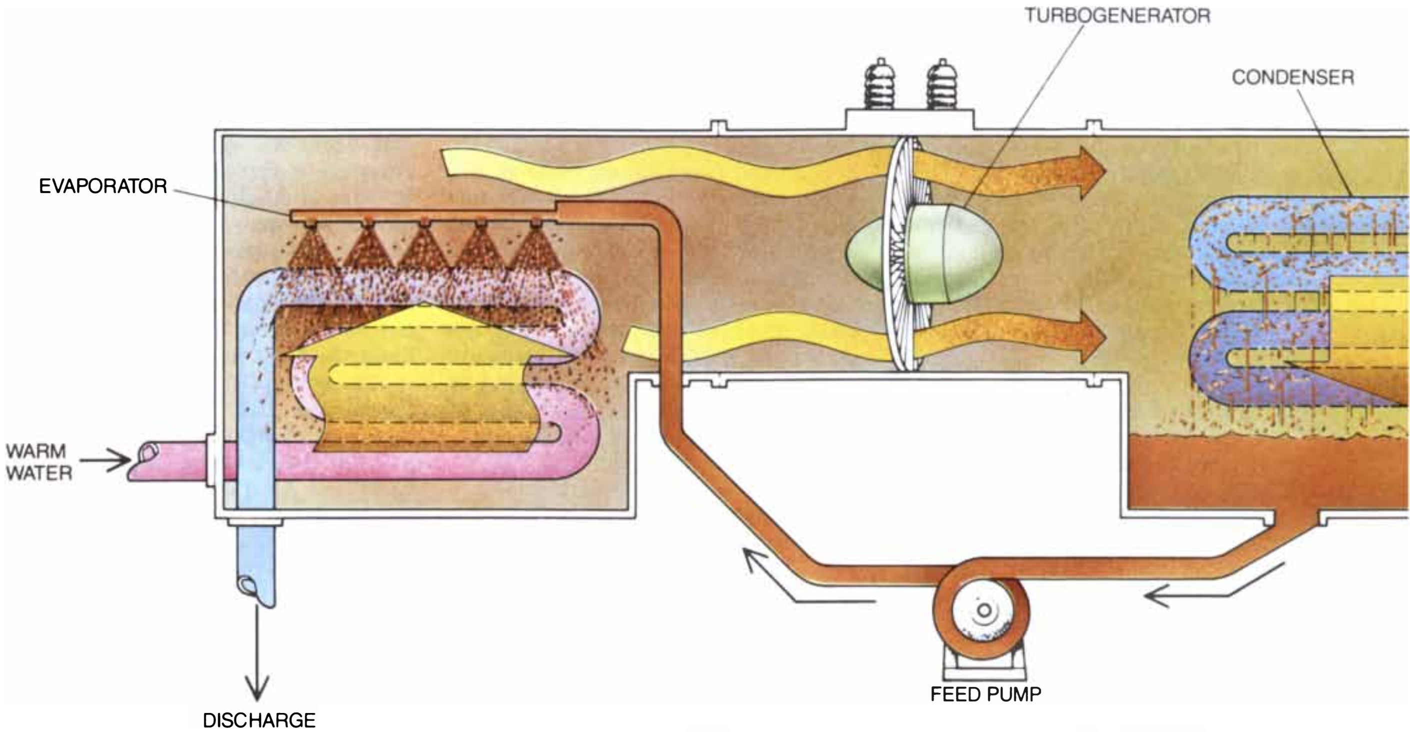
Claude's next major effort was a floating open-cycle plant installed on a cargo vessel moored off the coast of Brazil. The experiment failed because waves destroyed the cold-water pipe as it was being deployed. Claude, who had invested his own money in these projects, died virtually bankrupt and never achieved his goal of generating net power with an open-cycle system.

Influenced by Claude's work, the government of France continued research on open-cycle systems for several years. In 1956 a French team designed a three-megawatt plant to be built at Abidjan on the west coast of Africa, where a temperature difference of 20 degrees C. prevails. For various reasons, including difficulty in deploying the cold-water pipe (2.5 meters in diameter and four kilometers long), the plant was not built.

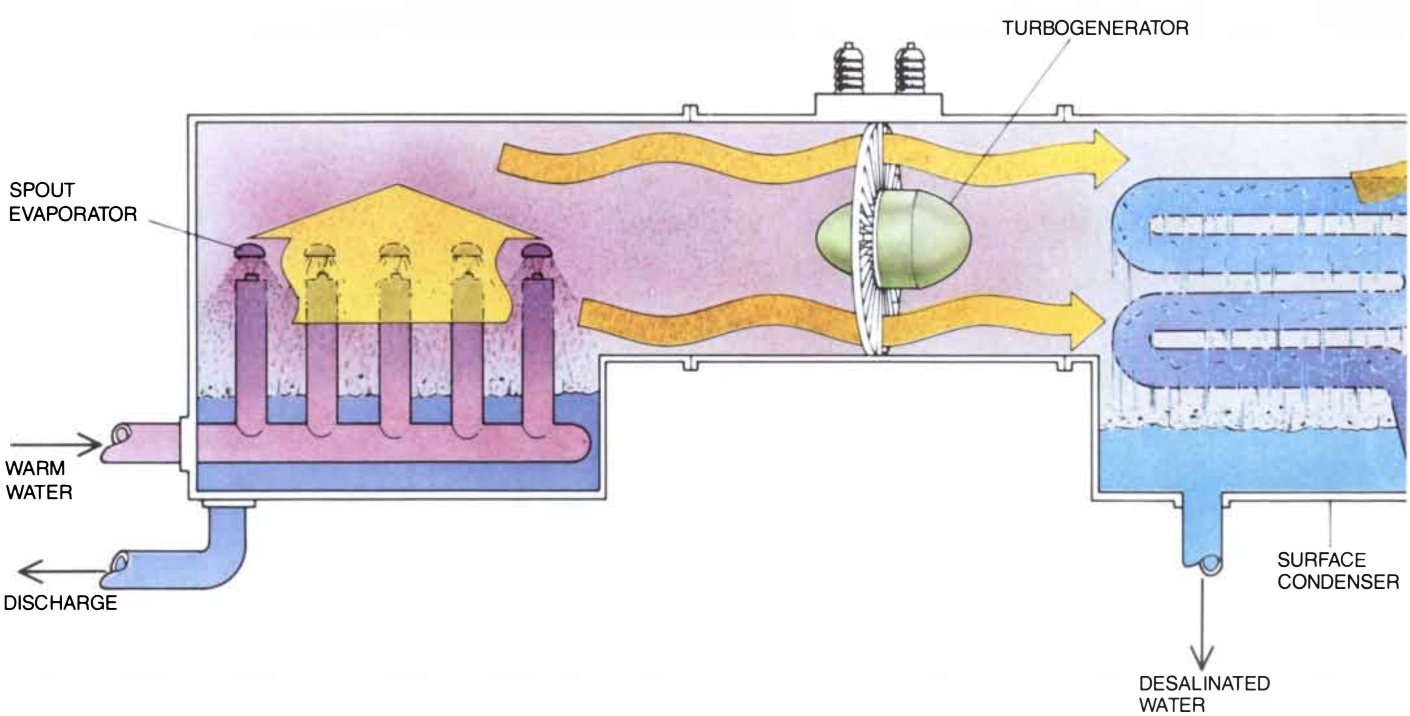
The energy crises of the 1970's forced the U.S. and several other countries to devote serious attention to OTEC. The state of Hawaii and the Lockheed Aircraft Corporation, with technical support from the Dillingham Corporation, built Mini-OTEC, the first OTEC plant to produce a net output of electric power. It was a closed-cycle system mounted on a barge moored about two kilometers off Keahole Point on the island of Hawaii. The



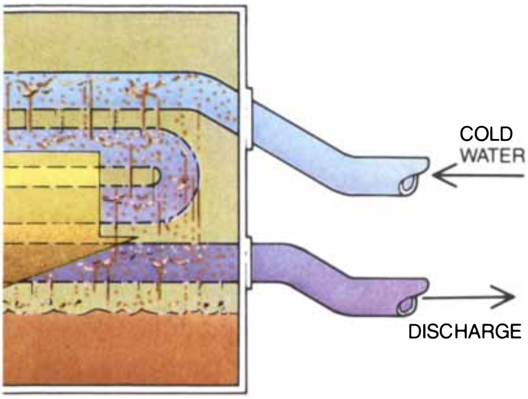
water intake that draws water from the sea surface, a pipe that draws bottom water at least 20 degrees Celsius cooler and a discharge pipe. The warm water generates steam, which drives a turbine generator; the cold water recondenses the working fluid. The spherical tank holds desalinated water produced in the plant, and the pool to the rear of the tank is for culturing fish in the nutrient-rich cold water. Cold water from the power plant also can serve the nearby buildings for both refrigeration and air conditioning.



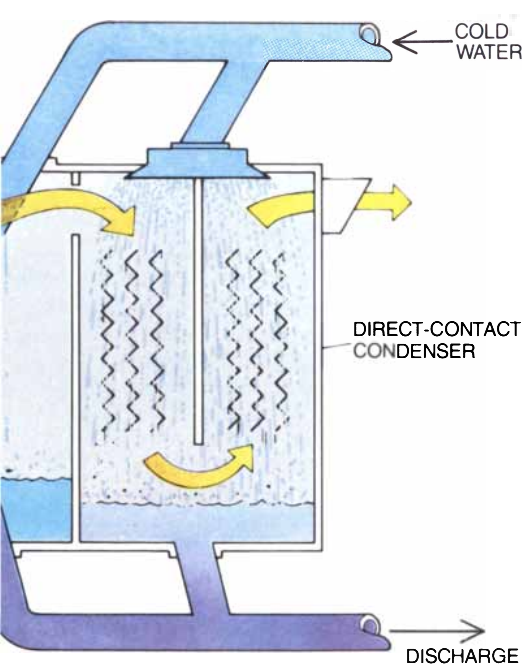
CLOSED-CYCLE OTEC SYSTEM has a working fluid (*brown*), such as ammonia or Freon, with a low boiling point. The fluid is recycled continuously. First it is pumped through an evaporator, where it is vaporized by warm seawater. The vapor drives a turbogenerator and then passes into a condenser that cools and condenses it with cold water from the deep sea. The condensed fluid is



OPEN-CYCLE OTEC SYSTEM has seawater as the working fluid. Warm seawater is evaporated to steam in a vacuum chamber, where the low pressure reduces the boiling point. The steam drives a turbogenerator and then passes through two condensers. One, a surface condenser, can produce desalinated water. The condensed water from a direct-contact condenser is discharged to sea instead



returned to the evaporator to renew the cycle. Seawater from the evaporator and the condenser is discharged into the sea.



of being returned to the evaporator. The system is often called the Claude cycle, after Georges Claude, an early developer.

plant operated for up to 10 days at a time over a period of four months, generating 50 kilowatts of gross power and 15 kilowatts of net power.

During the same years the U.S. Department of Energy supervised the construction of OTEC-1, an experimental system installed on board a converted U.S. Navy tanker. The system was designed to test closed-cycle heat exchangers of commercial scale as well as a bundle of three cold-water pipes each 1.2 meters in diameter and the necessary pumps. The operation produced significant results, demonstrating the validity of the heat-exchanger design and proving that an OTEC plant can operate by "grazing," or moving at low speed through tropical waters.

Somewhat later the Tokyo Electric Power Company and the Toshiba Corporation (with 50 percent financial support from the Japanese Ministry of International Trade and Industry) built a closed-cycle plant in the island republic of Nauru in the Pacific. With Freon as the working fluid, the plant ran intermittently from October, 1981, to September, 1982. It generated 100 kilowatts of power (35 kilowatts net). All these pilot plants were designed to test OTEC systems and were not expected to achieve the ratio of net to gross power that would be typical of commercial OTEC plants.

The Nauru operation was the last field test of an OTEC system, but work on OTEC components is continuing. In the U.S. the research on closed-cycle systems has focused on improving heat exchangers, which are expected to account for at least 20 percent of the cost of a closed-cycle plant. The work includes efforts to reduce corrosion of the apparatus by seawater and fouling by marine organisms.

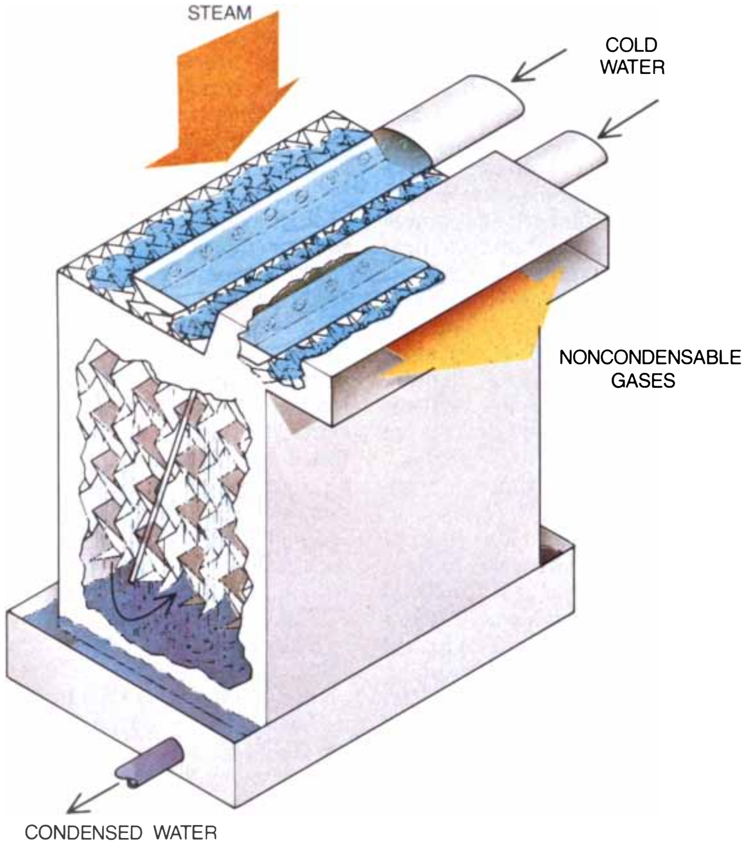
A conventional heat exchanger for a closed cycle has what is usually called a shell-and-tube configuration. Seawater flows through the tubes and the working fluid evaporates or condenses around them, within an outer shell. In the interest of increased efficiency a more advanced plate-fin design embodies an array of parallel plates. The plates are arranged so that one of them carries seawater, the plate next to it carries working fluid and so on throughout the apparatus. Fins between the plates contribute to the transfer of heat.

Titanium was originally chosen as the material for closed-cycle heat exchangers because of its resistance to corrosion. It is an expensive option for a closed-cycle OTEC plant, however, because of the volume of material required. Therefore the Argonne National Laboratory recently adopted modified brazed aluminum heat exchangers, which are common in refrigerators. Tests suggest they will last for 30 years or more in the corrosive marine environment if they have a thin protective coating. Aluminum heat exchangers should cost about a third as much as titanium ones.

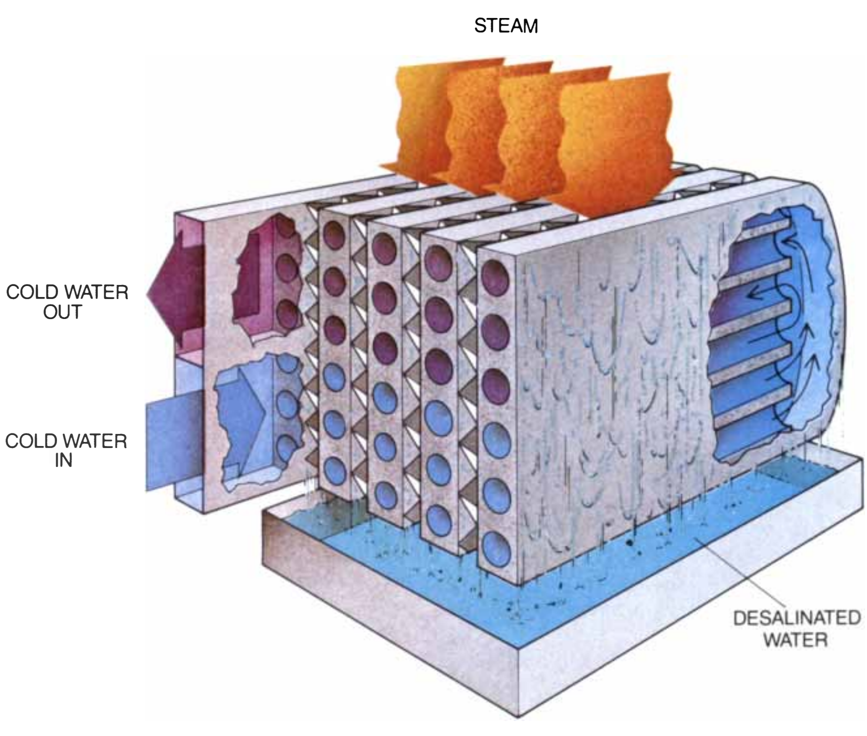
The Argonne workers have discovered that fouling will not be a problem for the parts of a plant exposed only to cold seawater, an environment in which chemical and biological reactions proceed slowly. As for warm seawater, laboratory experiments established that fouling can be controlled by intermittent chlorination totaling one hour per day. The chlorine levels are well below the current standards specified by the U.S. Environmental Protection Agency.

These results have yet to be applied in a closed-cycle pilot plant. Plans are under way for several such plants, including an American plant in Hawaii, a French one on Tahiti, a Dutch one on Bali and a floating British plant. The problem in each case has been obtaining the financing. U.S. estimates are that a 50-megawatt plant would cost from $200 to $550 million, depending on the site and the components. These estimates translate to a cost of from $4,000 to $11,000 per kilowatt of installed capacity and delivered costs of from five to 14 cents per kilowatt-hour. (A generating station fueled by oil at $20 per barrel delivers electricity at 5.6 cents per kilowatt-hour.) The capital cost of an OTEC plant is significantly more than a conventional steam plant, and in the present economic climate the money is not forthcoming.

Some Federal support is also going into research on open-cycle systems, particularly the low-pressure vacuum-chamber technique now often known as the Claude cycle. Although open-cycle systems have not been developed as fully as closed-cycle schemes, they appear to offer at least four advantages. First, with seawater as the working fluid they eliminate the possibility of contaminating marine environments with such toxic fluids as ammonia and Freon. Second, an open-cycle system has direct-contact heat exchangers that are cheaper and potentially more effective than the ones required for a closed cycle. Hence open-cycle plants may be more efficient at converting ocean heat into electricity, and they should also be less expensive to build. Third, direct-contact heat exchangers could be made of



SURFACE CONDENSER in an OTEC system provides a means of producing desalinated water in the Claude cycle. The condenser consists of a series of plates in a plate-and-fin configuration. Cold water flows through one plate, steam through the adjacent plate. Water condensing on the surface of the plates is free of seawater impurities and is fresh.



DIRECT-CONTACT CONDENSER puts steam and cold water in direct contact on a dense packing material; the material gives rise to an extensive surface area of liquid.

plastic rather than metal; they would be less susceptible to corrosion and might be less susceptible to fouling in warm seawater. Finally, a Claude-cycle system with a surface condenser can produce desalinated water as a by-product.

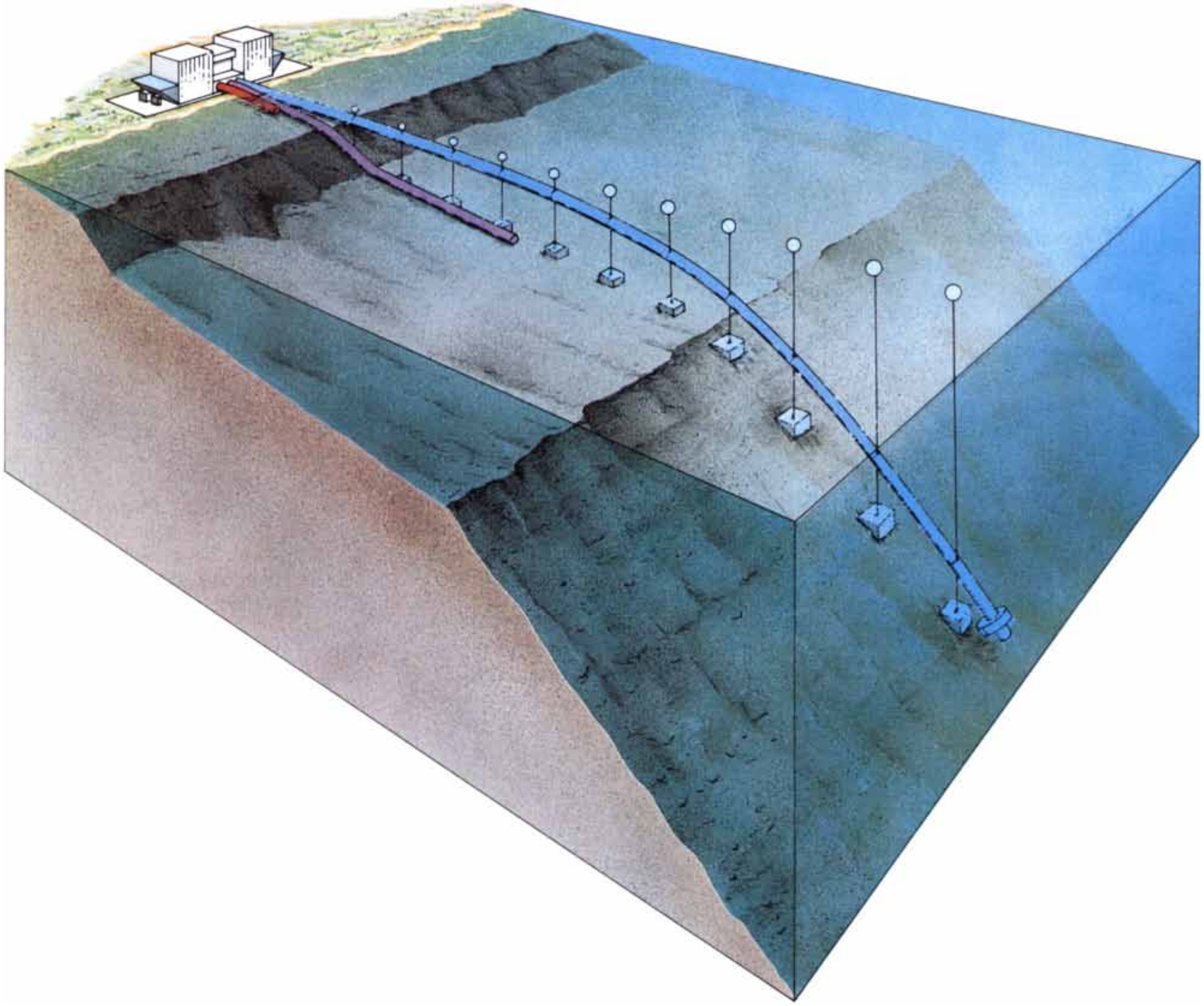
At the same time open-cycle systems present some distinctive technological problems. One is that the turbines must be much larger than the ones in a closed-cycle system because of the low density of the steam. Large turbines working in low-density steam require further research and development. Another problem is that an open-cycle system must provide smooth flow paths for the low-density steam, and so the vacuum chamber has to be quite large. A potential third problem is that the degassed seawater discharged from an open-cycle plant could change the chemistry of the nearby water and adversely affect the organisms there.

For the past five years a number of workers have sought to develop or improve such crucial Claude-cycle components as evaporators, condensers and turbines. The results of this work include a compact and simple vertical-spout evaporator and a direct-contact condenser developed at our institution, the Solar Energy Research Institute in Golden, Colo. Both devices have the virtue of operating with very low loss of pressure.

Because only .5 percent of the warm seawater entering the plant is converted into steam, the plant needs a large flow of warm water (from two to four cubic meters per second, equaling from 31,700 to 63,400 gallons per minute). Pumping that much water means pressure losses must be held to a minimum in the warm-water loop in order to ensure that the net output of power is at a maximum.

To help avoid pressure losses the vertical-spout evaporator has relatively simple intake and exit arrangements that make it easy to separate steam from spent seawater. The evaporator increases the heat transferred in a given volume of water by an order of magnitude over commercially available models.

Another feature of this open-cycle device affects the amount of surface area it provides for evaporation. Flash evaporation (characteristic of a low-pressure open-cycle operation) takes place only at the surface of the seawater working fluid. Hence the evaporator should break up the water in such a way as to provide a large and continually renewed surface area. The new evaporator effectively breaks up the incoming seawater with its array



PIPE ARRANGEMENT for an OTEC plant is indicated. The short pipe is the warm-water intake. The longer pipe accommodates discharges. The longest pipe takes in cold water; it goes through rings on cables anchored to the bottom and supported by buoys.

of vertical spouts. The resulting droplets not only collide with one another, keeping the water well mixed, but also break apart when they hit separation screens positioned below the spouts.

The work on condensers has focused on direct-contact heat exchangers, which can be more effective than surface condensers and yield higher outputs of power. Direct-contact condensation is much like flash evaporation. It works best when the cold seawater is distributed quite uniformly through the condenser, with as much liquid surface area as possible available to the spent steam.

One type of direct-contact condenser consists of two open-ended cylinders filled with a packing material of a kind typically used in cooling towers. This material distributes the water evenly throughout the condenser and creates the liquid surface area needed for condensation. About 80 percent of the steam is condensed as it travels in the same direction as the cold seawater through the first stage of the condenser. The remaining steam, along with any noncondensable gases present, is then routed into the bottom of the second cylinder and flows through the second stage in a direction opposite to that of the seawater. An exhaust system at the top of the cylinder sweeps out the noncondensable gases along with any uncondensed steam.

A drawback of direct-contact condensation is that it does not produce desalinated water. To obtain desalinated water the plant must keep the condensed, desalinated steam separate from the cold seawater. The process makes it necessary to have a fairly large surface condenser.

A two-megawatt (net) open-cycle plant with a surface condenser can produce some 4,320 cubic meters of water per day (792 gallons per minute or 1.1 million gallons per day). Unless the requirement for fresh water exceeds the capacity of the plant, such a plant will employ both surface and direct-contact condensers. A second direct-contact condenser will be employed after the desalination stage to concentrate noncondensable gases and reduce the size of the vacuum exhaust system, thereby increasing the ratio of net to gross power.

The turbine is perhaps the single most important component of a Claude-cycle OTEC system. It is also the least tested component. The requirement for large turbines may be met in small plants by employing the largest steam turbines in use today, which are 4.5 meters in diameter and serve in the secondary, low-pressure stages of some conventional generating systems. Since the number of turbines is not likely to increase plant costs by more than 10 percent, the output of power can be increased to a limited extent by running several turbines in parallel. (The French design for a 20-megawatt plant on Tahiti envisions four such units.)

The alternative, which is to build a single turbine of larger diameter, would require significant advances in turbine technology. Workers at the Westinghouse Electric Corporation have concluded that a plant producing 100 megawatts of net power would need a turbine 43.6 meters in diameter. The stress on the long blades of such a turbine would probably rule out the metal-alloy blades that now serve in low-pressure turbines. Fiber-reinforced plastics may provide the solution; they are strong, light in weight and easy to mold into the complex shapes required. Workers at Westinghouse, Advanced Ratio Design, Inc., and the University of Delaware have designed composite blades that are long and slender like the blades of a helicopter and yet are highly twisted and cambered like the blades of conventional low-pressure steam turbines.

Federally sponsored research indicates that a 10-megawatt open-cycle plant would cost about $7,200 per kilowatt of net power. At that price it would be competitive in Pacific island markets only if the price of oil became much higher than it is now. In the Caribbean and other water-deficient areas, however, revenue from selling desalinated water could make an open-cycle OTEC system competitive at much lower oil prices. Two recent economic and technical analyses of potential sites by French and American groups indicate that small Claude-cycle plants, producing from five to 15 megawatts net, will be feasible for certain island communities in the fairly near future.

The unique feature of any OTEC system is the cold-water pipe, which unlocks the door to a vast and unexploited ocean resource. The cold water not only helps to generate electricity but also contains the nutrients needed to sustain large quantities of the plankton and algae that support animal life in marine ecosystems. Natural upwellings of this deep ocean water, which occur over only .1 percent of the ocean, produce 44 percent of the world's fish. Cold water pumped up by OTEC-generated electricity could support large-scale marine farming operations. The water has many secondary uses, including refrigeration and air conditioning.

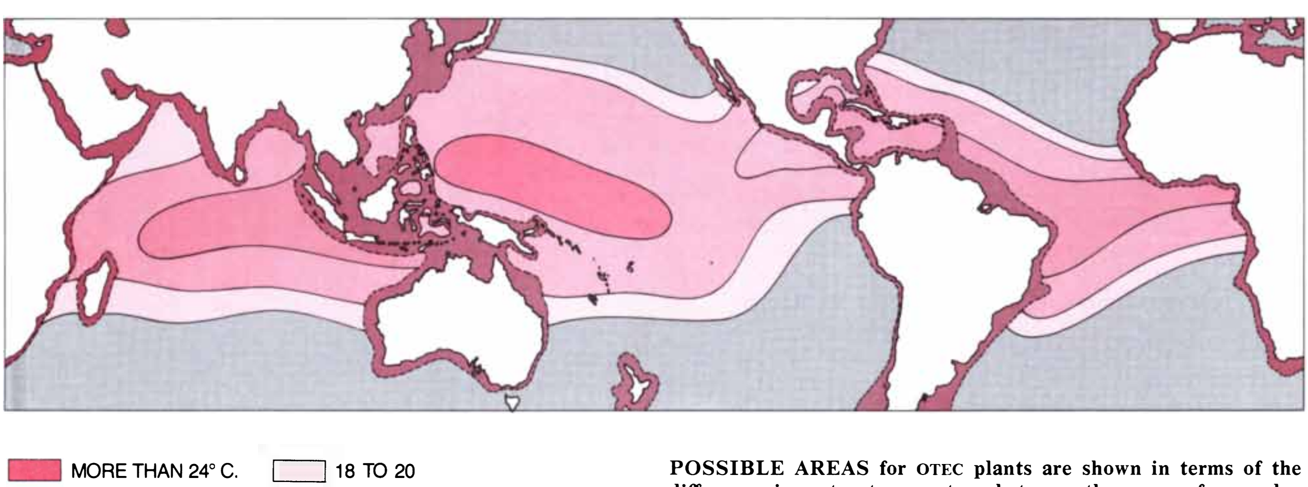
This great potential remains unrealized because of the difficulty of building pipes more than one meter in diameter and deploying them 1,000 meters below the surface of the sea. For nearly 40 years after Claude's pioneering experiment all further efforts to build and operate deep-water pipes failed. Then Mini-OTEC and OTEC-1 successfully deployed pipes made out of flexible polyethylene. These pipes were shorter, however, than those required for a shore-based facility.

In 1981 Hawaii deployed a small but long polyethylene pipe (.3 meter in diameter and 1.5 kilometers long) at the National Energy Laboratory of Hawaii. Today that pipe provides the world's only manmade, continuous supply of cold water from the ocean deep. Hence the laboratory is the primary research and development center for this resource, conducting experiments on both open- and closed-cycle OTEC systems. To supply more water for these experiments and for nearby commercial ventures in aquaculture the state and Federal governments and private industry are working with the laboratory to install a new polyethylene pipe. It will be about as long as the existing pipe but one meter in diameter. A number of other pipe materials, including concrete, steel, fiberglass-reinforced plastic, elastomers and composites, are under study.

One way to install a cold-water pipe of large diameter is to lay it on the sea bed. The procedure is tricky and difficult. Another way is to suspend a buoyant pipe above the sea floor by any of several methods. The French plan for the Tahiti plant is to install a series of anchored cables that will hang from buoys on the surface to the sea floor; the pipe will run through guides mounted on the cables. The pipe will thus be able to span discontinuities in the offshore shelf and rock outcroppings that are typical of volcanic islands.

The Tahiti plant is designed to put its cold-water resource to multiple applications: aquaculture, refrigeration and desalination. In view of the difficulties the French government faces in financing the plant, it is possible that the installation will not even generate net power. Water and aquaculture appear to be more promising economically.

The people involved in the Hawaiian OTEC activity believe commercial aquaculture will provide the economic incentive for private industry to develop an OTEC power plant in the near future. By fostering the development of cold-water pipes and other aspects of ocean-thermal systems the work could push OTEC technology to maturity by the time petroleum becomes costly or scarce.



POSSIBLE AREAS for OTEC plants are shown in terms of the difference in water temperature between the sea surface and a depth of 1,000 meters. In open water the plant would be on a ship. In other places a facility could be built on land or just offshore.

We can blow out forty candles, but only you can make our wish come true.

This year the National Multiple Sclerosis Society celebrates its fortieth birthday. Forty years of serving people with MS, with programs that help them get up, get around, and get on with their lives. And forty years of searching for a cure for this crippling disease of the central nervous system.

Progress? Yes, we're making progress. But not enough to stop and celebrate now.

Optimism? We've got plenty. As long as you're willing to support our continuing efforts with your contributions. Because, like it or not, money is the key to continuing the search for answers that wishing, alone, can't bring.

Yes, Multiple Sclerosis is incurable today. But it's your support that makes us incurably optimistic about finding a cure tomorrow.

THE NATIONAL MS SOCIETY AT 40. INCURABLY OPTIMISTIC.

205 East 42nd St., NYC, NY 10017

Fractal Growth

Growth in nature can produce the sprawling, tenuous patterns called fractals. One kind of fractal growth explains such diverse phenomena as how some solids crystallize and how air bubbles move in fluids

by Leonard M. Sander

Investigators who study bulk matter face a problem of bewildering complexity. Each macroscopic bit of the world contains gigantic numbers of atoms and molecules that are often arranged in complicated, disorderly patterns. In the case of a perfect crystal or a smoothly flowing liquid, the pattern is uniform on large scales. Even though a great deal can be understood about such systems, the vast majority of complicated natural phenomena—such as turbulent flows of fluid or air, the accretion of metal particles in an electrolytic bath and the formation of mountain ranges—have virtually defied understanding.

In the past 10 years both scientists and mathematicians have made much progress toward gaining such comprehension. Central to many of the new insights is the revolutionary concept of a fractal, a term coined by Benoit B. Mandelbrot of IBM's Thomas J. Watson Research Center in Yorktown Heights, N.Y. A fractal is an object with a sprawling, tenuous pattern. As the pattern is magnified it reveals repetitive levels of detail, so that similar structure exists on all scales. A fractal might, for example, look the same whether viewed on the scale of a meter, a millimeter or a micrometer (a millionth of a meter). Mandelbrot pointed out that many disorderly objects in nature have this property.

An increasing amount of evidence suggests that nature's love of fractal shapes is a deep one. Fractals known as percolation clusters have been identified with the pattern of a fluid flowing through a solid matrix, such as water seeping through the soil or coffee through ground coffee beans. Soot, colloids and some polymers seem to be fractals. Fractals also occur in the motion of air bubbles in oil, the growth of some crystals and the behavior of electrical discharges resembling lightning bolts. The random patterns of clouds and coastlines are almost certainly fractals as well.

As empirical evidence for the existence of fractals in nature mounts, investigators have begun to explore the question of how fractals form. In 1981 Thomas A. Witten III of the Exxon Research and Engineering Company and I proposed a mechanism for fractal growth that we call diffusion-limited aggregation. According to our model, a particular kind of fractal can result from a disorderly and irreversible growth process. The theory is attractive for two reasons. First, it is conceptually simple and easily modeled on a computer. More important, it seems to explain how a variety of real fractals form.

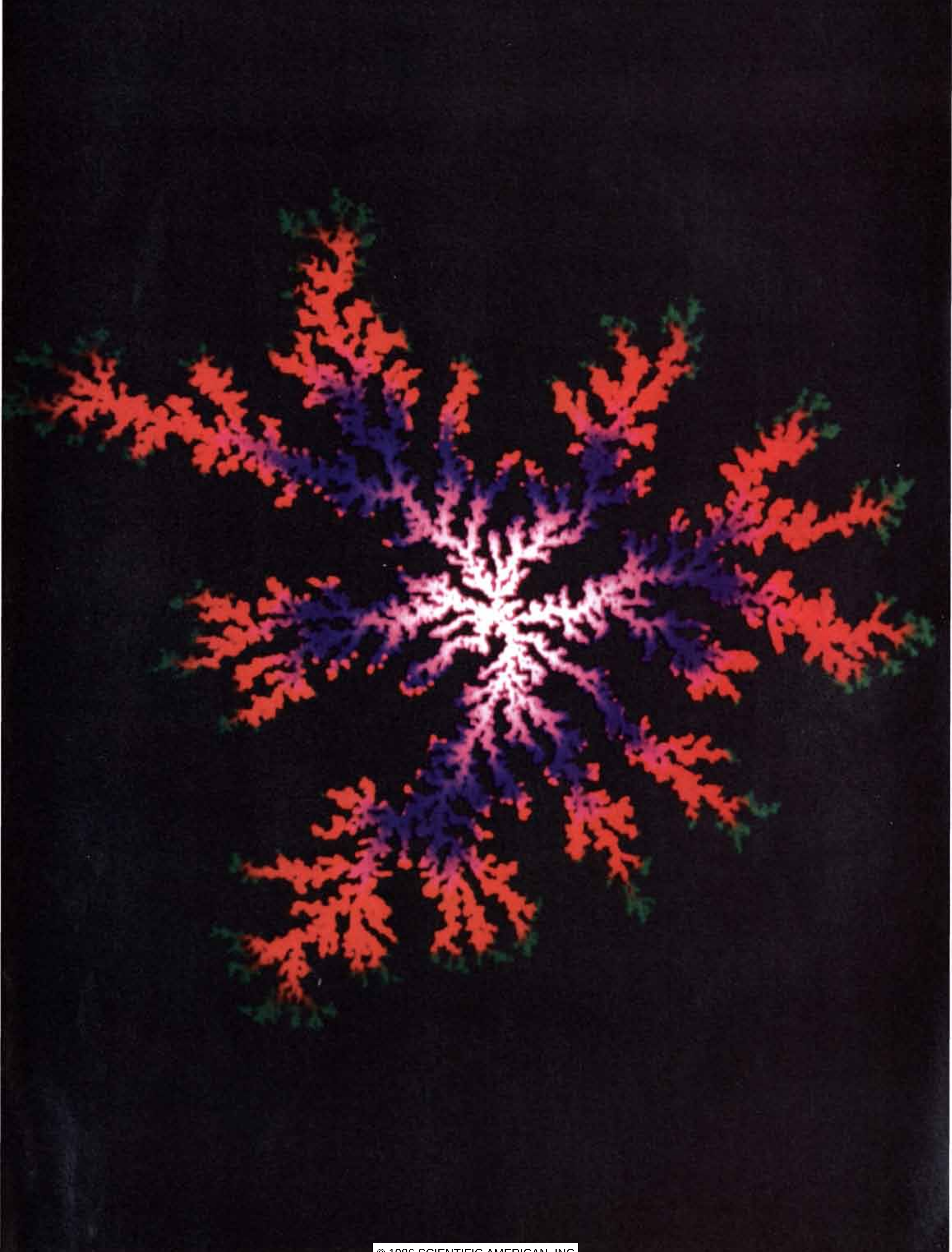
What are the properties of a fractal? Abstract discussions of objects now called fractals were given long before Mandelbrot by other mathematicians, who viewed them as "monsters" of only academic interest. Yet a typical fractal reminds people more of a snowflake than a monster. The reason is found in the repetition of a pattern. Each unit of the fractal shown at the bottom of the illustration on page 96, for instance, consists of five identical subunits. Five of the large units can then be put together to make a still larger unit, and so on. Every generation contains holes scaled to the size of that particular generation. The pattern is also scale-invariant: at each stage any part of the pattern that has one-third the diameter of the whole looks exactly like the whole. Scale-invariance is a "symmetry" of fractals. Just as round objects are symmetric under rotations, fractals are symmetric under dilations, or changes of scale.

It is useful to have a quantitative measure of how a fractal scales. This is given by a number called the fractal dimension. A fractal dimension, as opposed to an ordinary dimension, is expressed not as a whole number but as a fraction. The fractal under discussion, for example, has a dimension of 1.46: it is intermediate between a one-dimensional straight line and a two-dimensional plane. The more nearly a flat fractal object fills a plane, the closer its dimension approaches 2. The pattern shown in the computer image on the opposite page, which was formed by the diffusion-limited-aggregation process, has a measured fractal dimension of 1.71. The pattern is scale-invariant in a statistical sense. One property of that fractal—and in fact of all fractals—is that its density decreases as its size increases.

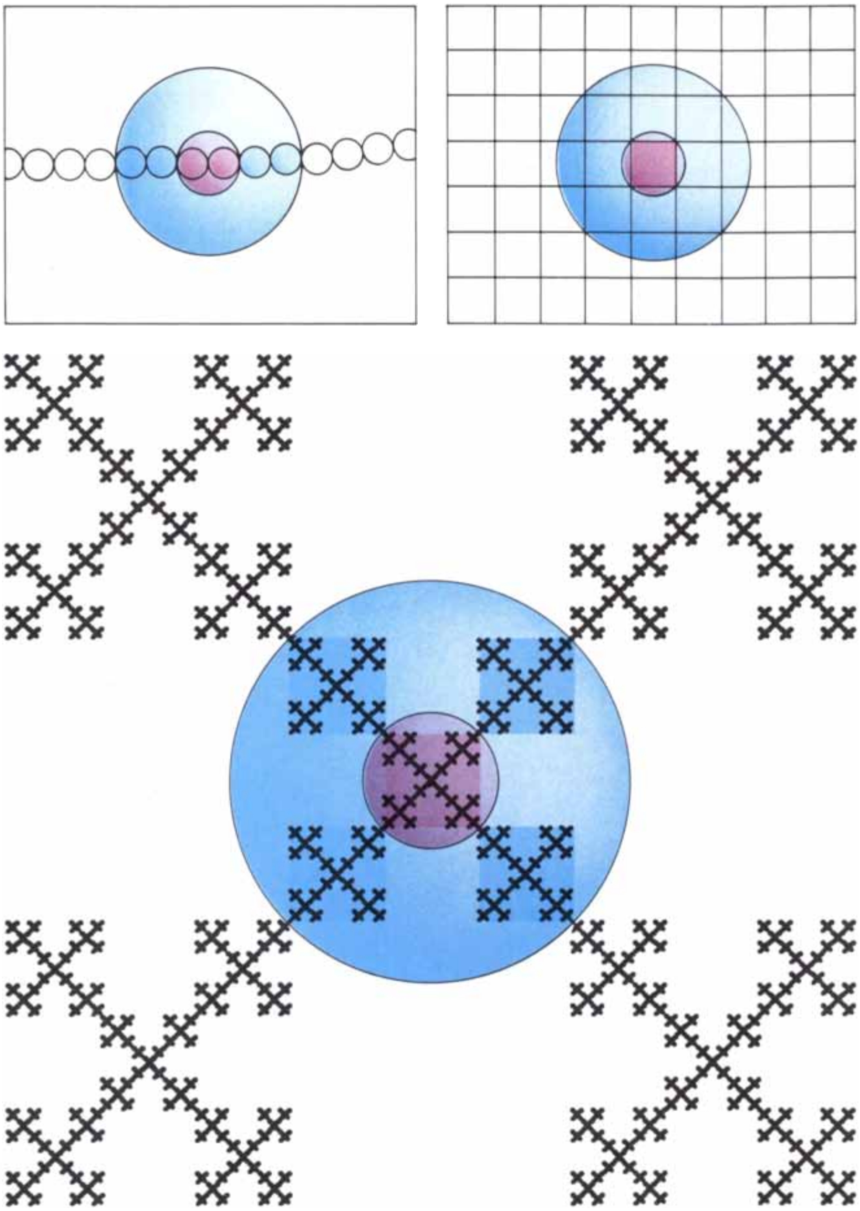
The fractal dimension of a physical object is a "universal" property: it is independent of many details of how the object is formed. The fractal dimension and other universal properties are related to behavior at large scales, where the particular details average out. As a consequence, a simple model that neglects most of the complexity of a real system should nonetheless describe the scaling properties of the system correctly.

The importance of the diffusion-limited-aggregation model is that it

FRACTAL pattern was produced by computer simulation of a process called diffusion-limited aggregation. Some 50,000 "particles" were released one at a time from an area outside the illustration and allowed to wander toward the origin. As they stuck to one another a cluster was formed, and then it grew. The color coding indicates when the particles arrived: white represents the early-arriving particles and green the late-arriving ones. The image was made by Paul Meakin of E. I. du Pont de Nemours & Company, Inc.



shows a relation between fractals and growth. There are many ways that objects in nature can grow. A perfect crystal, for instance, grows near equilibrium: it "tries" many configurations until it finds the state with the most stable structure. When a molecule is added to the growing crystal, it must in general search over many possible sites before it finally sticks in a favorable place. An equilibrium crystal forms slowly and is subject to constant rearrangement. Most real processes of growth, however, do not have the luxury of time. All biological life, for example, is out of equilibrium. The fractals I want to discuss grow far from equilibrium. (Some fractals grow near equilibrium, but that is beyond the scope of this article.)



FRACTAL DIMENSION differs from an ordinary dimension in that it is not expressed as a whole number but as a fraction. To determine the precise fractal dimension of an object one counts the average number N of fundamental units of repetition found within a sphere of a certain radius r centered somewhere on the object. According to Euclidean geometry, the number of fundamental units then equals a constant C multiplied by the radius raised to the value of the dimension D ($N = C \cdot r^D$). In the case of a line the dimension is, of course, 1: tripling the radius of the sphere triples the number of units inside (*top left*). For ordinary (nonfractal) bulk matter in two dimensions, tripling the radius of the sphere multiplies the number of units by 9 (*top right*). For a fractal of dimension 1.46 (*bottom*), in contrast, tripling the radius multiplies the number of units by 5. In other words, the number of units grows faster than it does in the case of a line but not as fast as in the case of ordinary bulk matter. In this sense the object is intermediate between a line and a plane. The fractal shown here was invented by Tamás Vicsek of the Research Institute for Technical Physics of the Hungarian Academy of Sciences in Budapest.

Imagine growing a cluster by adding one particle at a time so that as each particle comes in contact with the growing object, it sticks and never tries another site, that is, it simply stays put. Such a process is called aggregation. It is an extreme example of a nonequilibrium process, because no rearrangement takes place at all. Now suppose the particles diffuse to the cluster by means of a random walk: a sequence of steps whose magnitude and direction are determined by chance. (In a one-dimensional version of a random walk, for instance, a person flips a coin and takes one step forward if the result is heads and one step backward if the result is tails.) The aggregation of particles by means of random walks is what Witten and I called diffusion-limited aggregation.

Small clusters are quite readily "grown" on a personal computer. A cluster is started by placing a particle at an origin. Then another particle is released some distance away and allowed to take one random step after another until it approaches within a particle diameter of the first. When the second particle sticks, another is released at a random location far from the aggregate, and so on. The simulations suggest that diffusion-limited-aggregation clusters are fractals.

Although diffusion-limited aggregation is simple to describe and to simulate, the process is still not well understood at a deeper level. Why, for example, does diffusion-limited aggregation give rise to fractals rather than to, say, amorphous blobs with no symmetry at all? Why are loops rarely formed? How does the fractal dimension depend on the dimension of space? Answers to these questions remain at large, and they pose a remarkable problem to the theoretical physicist because none of the ordinary mathematical tools seems to work when applied to them.

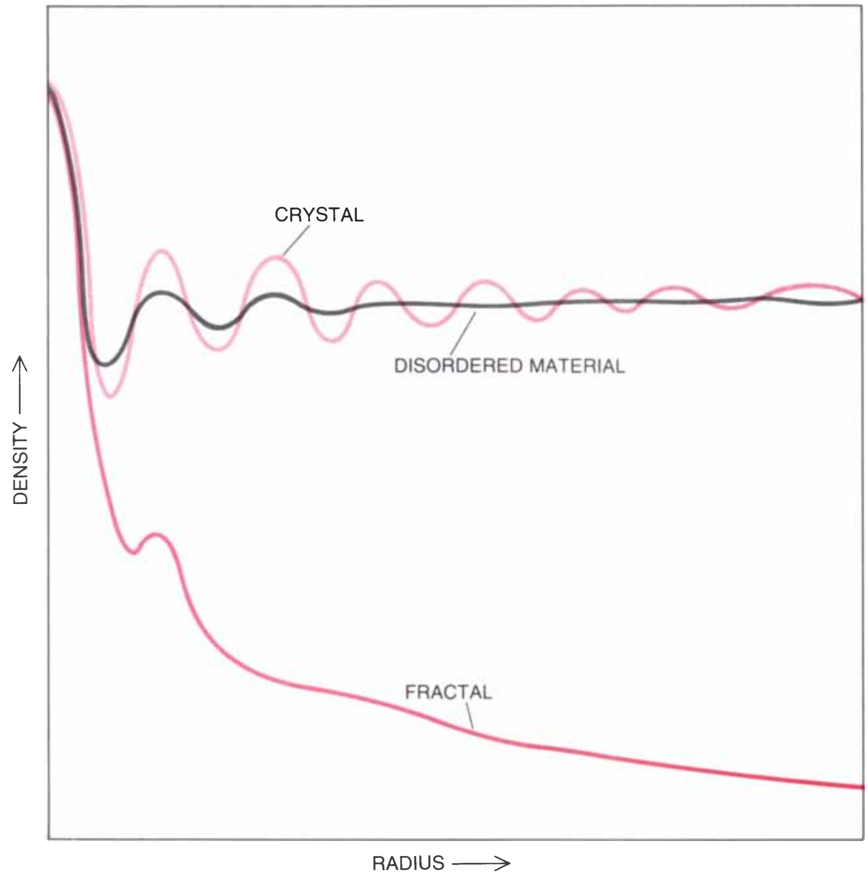
It is possible, however, to understand in a qualitative way some important features of the process. Imagine beginning with a smooth cluster onto which diffusing particles aggregate. When the cluster is small, several incoming particles may stick—purely by chance—on one part of the surface. In other words, tiny bumps and holes will form on the surface owing to "noise," or random statistics, in the behavior of the particles.

Once the surface has bumps and holes, the bumps will grow much faster than the holes. The reason is that a random-walking particle, moving tortuously from the exterior, will probably stick at or near the peak of a bump; it will almost certainly stick on the side before it could fall into a hole. Since the particle sticks near the peak, the bump grows even steeper; the filling of a hole becomes even more unlikely. As a result a cluster that is initially slightly distorted will become even more distorted, an effect called a growth instability. Eventually the growth and splitting of the protruding ends presumably give rise to a fractal. Even though the details of the process remain unknown, it is the interplay of noise and growth that is the source of the complexity and richness of diffusion-limited-aggregation clusters.

In the past five years diffusion-limited aggregation has been the subject of extensive inquiry. A great deal of excitement springs from the fact that the model seems to describe the real world: actual particles often do wander to sites, where they stick. Robert M. Brady and Robin C. Ball of the University of Cambridge pointed out in 1984 that diffusion-limited aggregation is, for instance, a reasonable idealization of what happens when a metal is deposited from an electrolytic solution of diffusing ions. Although the details of what happens when an ion sticks to the deposit are surely quite different from what happens in a computer simulation, they do not seem to affect the overall resulting shape. Nor do the details affect the resulting fractal dimension.

A deposit of zinc metal produced in an electrolytic cell [*top left in illustration on page 100*], for example, bears a striking resemblance to the computer-generated fractal pattern on page 95. The zinc deposit has a measured dimension of 1.7, which, taking experimental error into account, agrees with the measured dimension of the computer-generated fractal (1.71). The agreement is a remarkable instance of universality and scale-invariance: the computer simulation used about 50,000 points, whereas the number of zinc atoms in the deposit is enormously large, almost a billion billion.

In fact, varying the rules for the computer simulation reveals several types of universality. Suppose, for instance, a particle sometimes bounces off the aggregate instead of sticking. That rule, which is a simple representation of one of the many complications that might develop in a real physical situation, turns out to lead to a thickening of the branches of the aggregate, but it does not change the fractal dimension.



DENSITY OF A FRACTAL decreases with increasing size, as shown here. The densities of both ordered crystals and amorphous blobs, in contrast, approach fixed constants.

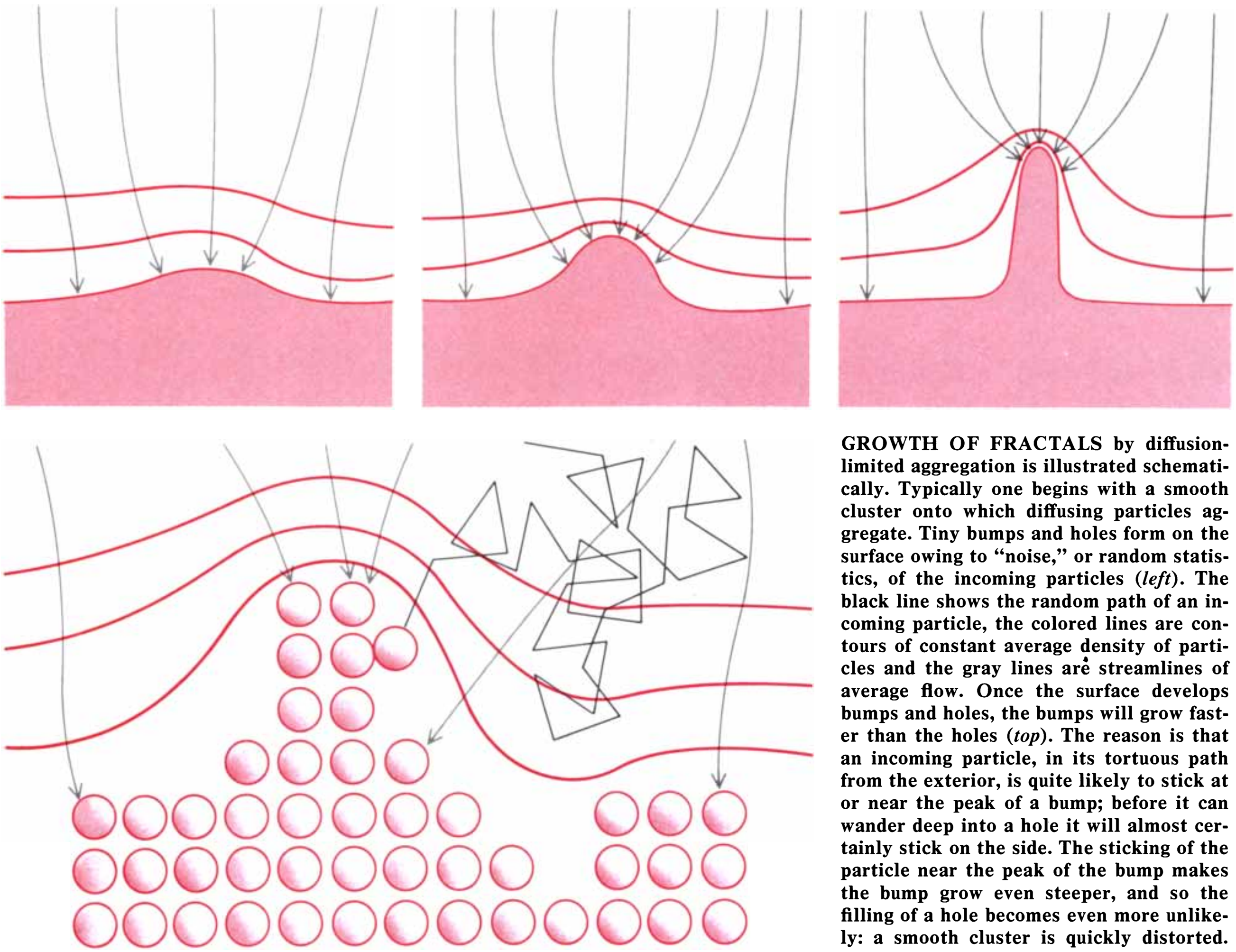
It is certainly easy to believe the deposition of a metal onto an electrode is described by diffusion-limited aggregation. Remarkably, however, the model also appears to account for a wide range of phenomena. One phenomenon involves an experimental apparatus called a Hele-Shaw cell, adapted from the work of a 19th-century British naval engineer, Henry S. Hele-Shaw. The cell consists of a viscous fluid, such as glycerine, confined between two parallel plates. When a less viscous fluid, such as air, is injected into the middle, the glycerine is displaced. An air bubble forms, sprouting a number of "fingers," or protruding ends [*top right in illustration on page 100*]. The phenomenon is called, appropriately enough, viscous fingering.

Viscous fingering is of practical interest because it also occurs when water is injected into the middle of an oil field during enhanced recovery operations. The efficiency of the recovery is greatly decreased by the fingering pattern. Unless special techniques are employed, one can displace only a small amount of the viscous oil toward the wells at the edge of the field.

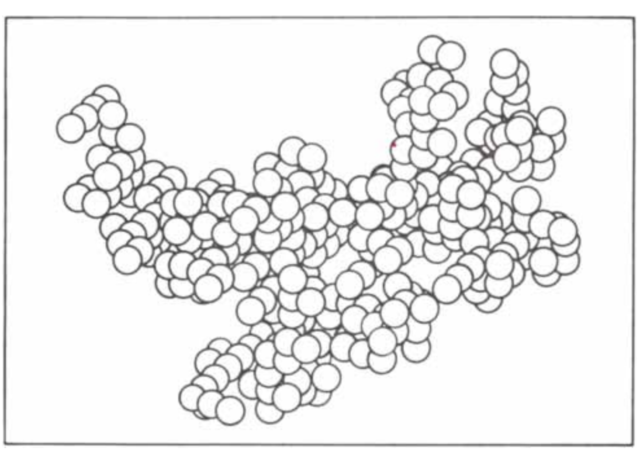
The patterns that result from viscous fingering show a strong similarity to computer-generated images of diffusion-limited-aggregation clusters. Why is this so? The answer was provided recently by Lincoln Paterson of the Commonwealth Scientific and Industrial Research Organization in Australia. Paterson points out that the processes in both diffusion-limited aggregation and viscous fingering are the same in principle. In the former, growth is due to the net inward flow of random-walking particles. The flow arises from the fact that particles are more likely to wander in from densely populated regions outside the cluster than from sparsely populated regions. Specifically, the flow is proportional to the rate of change in the populations outside the aggregate.

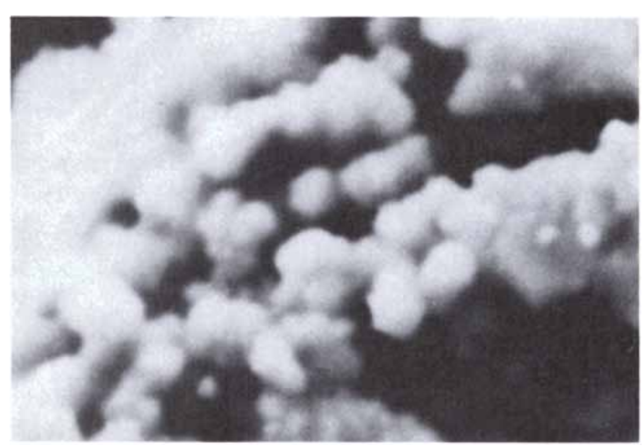
In the case of viscous fingering, the fluid pressure in the glycerine is analogous to the populations of particles. The pressure is highest at the boundary between the air bubble and the glycerine. The pressure is relieved by the flow of glycerine away from the bubble. The rate of flow is proportional to the rate of change in the pressure at points outside the bubble. The fingers grow because fluid flows away from them most easily. Since the boundary moves when the fluid flows away, the tips grow longer. As a result there is a growth instability similar to that of diffusion-limited aggregation.

Yet another system has been analyzed in much the same way. When a voltage is applied to an electrode that touches a piece of photographic emulsion or a fine powder spread on the surface of an insulator, a jagged, stringy electrical-discharge pattern resembling lightning results [*bottom left in illustration on page 100*]. The pattern is called a Lichtenberg figure, after the 18th-century German physicist Georg Christoph Lichtenberg. In 1984 a



GROWTH OF FRACTALS by diffusion-limited aggregation is illustrated schematically. Typically one begins with a smooth cluster onto which diffusing particles aggregate. Tiny bumps and holes form on the surface owing to "noise," or random statistics, of the incoming particles (*left*). The black line shows the random path of an incoming particle, the colored lines are contours of constant average density of particles and the gray lines are streamlines of average flow. Once the surface develops bumps and holes, the bumps will grow faster than the holes (*top*). The reason is that an incoming particle, in its tortuous path from the exterior, is quite likely to stick at or near the peak of a bump; before it can wander deep into a hole it will almost certainly stick on the side. The sticking of the particle near the peak of the bump makes the bump grow even steeper, and so the filling of a hole becomes even more unlikely: a smooth cluster is quickly distorted.



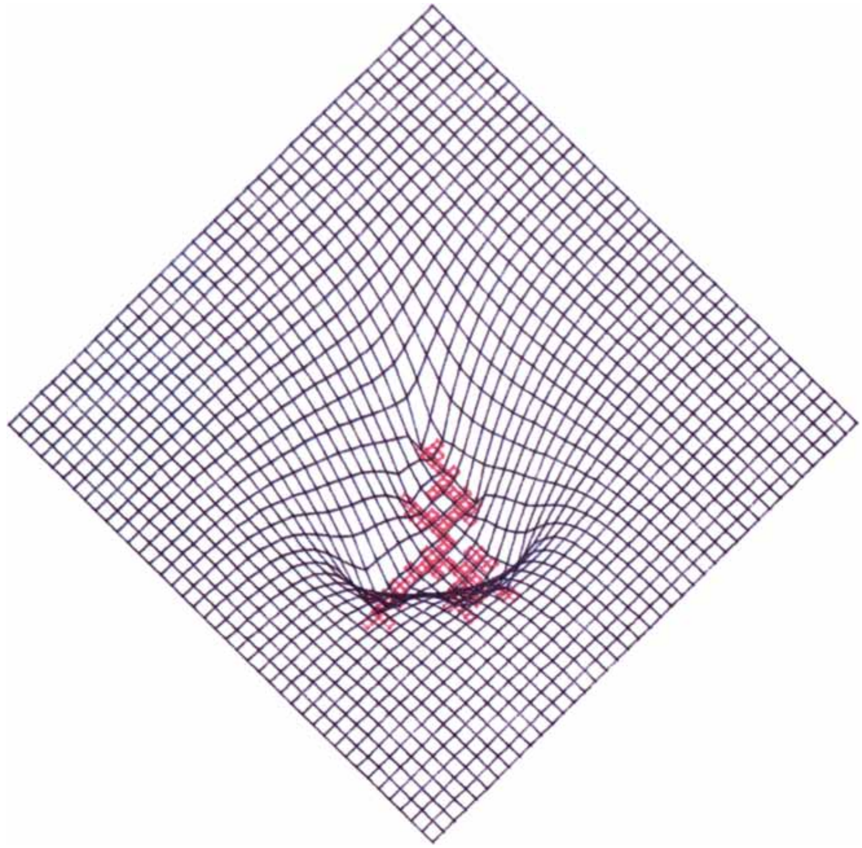


DIFFUSION-LIMITED AGGREGATION, simulated in a three-dimensional space by computer, gives rise to a fractal with a dimension of 2.4 (*left*). The pattern shows a close similarity to that of a copper cluster (*right*), which has the same dimension. The computer image was generated by Roy Richter of the General Motors Research Laboratories, and the copper cluster, which was deposited from a solution of copper sulfate, was prepared by Nancy Hecker and David G. Grier of the University of Michigan.

group of investigators at Brown, Boveri & Company, Limited, in Switzerland reported that diffusion-limited aggregation seems to underlie the growth of Lichtenberg figures.

Suppose the applied voltage is initially high enough near the electrode to damage a small region of the emulsion and create a channel of electrical conduction. Outside the channel there is an electric field; the strength of the field is given by the rate of change in the voltage at various points in the material. The Brown, Boveri group made the reasonable assumption that the growth of the channel will most likely occur where the electric field is strongest, which turns out to be at the sharp tips of the discharge. As a result the tips grow longer and proliferate: fractal growth takes place.

The common framework that unites the deposition of a metal on an electrode, the fingering of a viscous fluid and the formation of a Lichtenberg figure is most easily expressed in the abstract language of partial differential equations. It is not difficult, however, to gain some appreciation of the underlying effect by making a simple analogy to a rubber sheet, taut along all four edges and pressed down in the middle by a growing fractal.



STRETCHED RUBBER SHEET, taut along all four edges and pressed down in the middle by a growing fractal, serves as a simple model of diffusion-limited aggregation. The fractal grows fastest at the places where the rubber sheet descends most steeply: the tips of the fractal. At the next stage still sharper tips will press on the sheet, and so on.

Functions such as the probability of where a particle will randomly walk, the pressure in a Hele-Shaw cell and the voltage near a breakdown channel are all examples of harmonic functions, which are solutions to a set of partial differential equations. A harmonic function has a net zero curvature: if it curves up in one direction, it curves down in the perpendicular direction (somewhat like a horse saddle). The rubber sheet also has zero curvature. Thus the height of the sheet can be thought of as a mapping of probability, pressure or voltage, and its slope at the edge of the fractal is the growth rate. The steepest slopes occur near the sharp tips, which grow preferentially. At the next stage still sharper tips press on the sheet, and so on.

It is tempting to speculate how far such an analysis can be pushed. For example, the random branching patterns of blood vessels, air passages to the lungs and coral reefs are certainly reminiscent of fractal patterns formed during diffusion-limited aggregation. Although some investigators have attempted to model the growth of these processes, to the best of my knowledge none of the approaches has explicitly invoked fractal geometry. Whether it will prove fruitful to analyze some kinds of biological growth in similar terms remains to be seen.

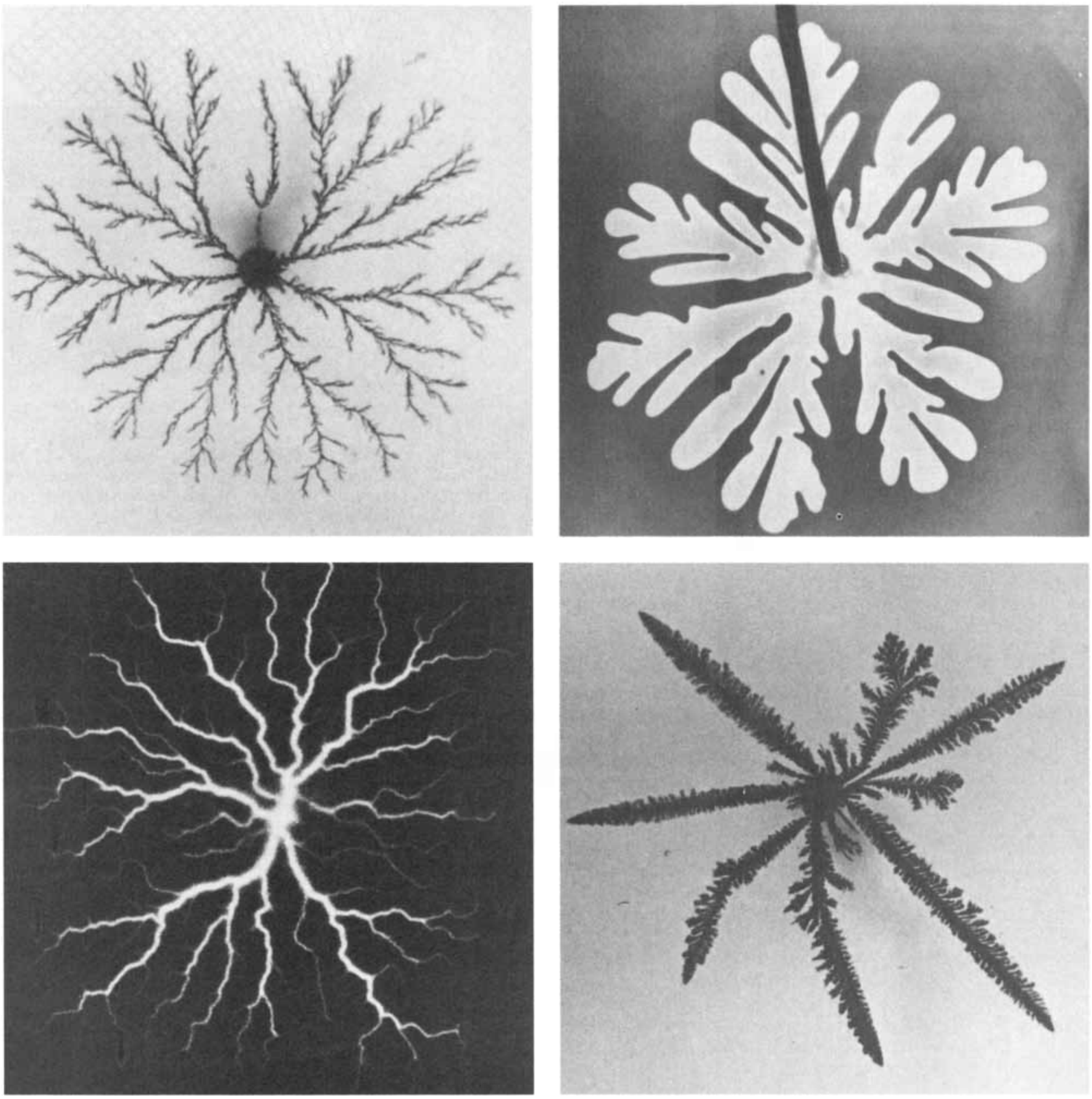
The model of diffusion-limited aggregation has also been used to describe other physical systems, such as surface crystallization of amorphous films. In addition a generalization of the model, called cluster-cluster aggregation, has been shown to describe colloids and aerosols such as soot. In this process, which was proposed by Paul Meakin of E. I. du Pont de Nemours & Company, Inc., and Max Kolb, Rémi Jullien and Robert Botet of the University of Paris at Orsay, many clusters can form and can themselves move and combine. In short, aggregation models have been extremely useful in describing physical systems.

At the same time I must also point out that fractals certainly do not account for all sprawling patterns found in nature. Snowflakes, for instance, are probably not fractals. Snowflakes are complex, to be sure, but they possess much more evident symmetry than diffusion-limited-aggregation clusters, and as such they are part of a family of crystals called dendrites. The beautiful macroscopic structure of a snowflake reflects the underlying microscopic anisotropy of the hexagonal lattices in which its atoms are arranged. One might then wonder why zinc, which also has a hexagonal form, deposits itself as a fractal in an electrolytic cell [*top left in illustration on next page*]. The answer is that the growth, albeit out of equilibrium, is so slow that the splitting of the tips wipes out the lattice anisotropy. Interestingly enough, when the growth rate is increased by raising the voltage of the cell, the anisotropy makes itself felt and a dendritic, or snowflakelike, pattern results [*bottom right in illustration on next page*]. The transition from fractal patterns to dendritic patterns is being explored by several groups of investigators.

So far I have concentrated on a particular growth process, diffusion-limited aggregation, that produces fractals. Can the knowledge be put to good use? In particular, do the scaling properties I have discussed lead to a useful understanding of physical characteristics of clusters other than their geometry?

Early signs indicate that they do. Over the past several years Raoul Kopelman and his colleagues at the University of Michigan have, for example, explored various chemical reactions that take place on percolation clusters (which are fractals formed in equilibrium). The investigators have shown that if the reaction is confined to the cluster, it acts oddly. Unlike a typical reaction, which occurs at a constant rate, the reaction rate on the percolation cluster appears to depend on time. The basic reason is that chemical species wandering on a fractal do not diffuse nearly as efficiently as they do in free space. They are hampered in finding one another because they are trapped in a structure that has many dead ends.

The reaction rate associated with a fractal depends both on the fractal dimension and on the way the reacting species move on the cluster. The combination of the two factors gives rise to another parameter called the spectral dimension. The quantity was introduced by Shlomo Alexander of the Hebrew University of Jerusalem and Raymond L. Orbach of the University of California at Los Angeles to describe diffusion and dynamics on a fractal. Although there is as yet no experimental evidence about the spectral dimension on nonequilibrium fractals, there is every reason to expect that it exists. In the future the very existence of the fractal geometry is likely to open still more frontiers of physics.



FRACTALS IN NATURE appear to grow by means of diffusion-limited aggregation. Shown here are a zinc deposit formed in an electrolytic cell (*top left*), a "viscous fingering" pattern of an air bubble in glycerine (*top right*) and an electrical-discharge pattern called a Lichtenberg figure (*bottom left*). The thick line that ends at the center of the bubble is an air tube. The zinc cluster at the bottom right shows what happens when the voltage in the electrolytic cell is increased: the growth pattern shifts from a fractal pattern to a dendritic, or snowflakelike, pattern. The zinc deposits were produced by Grier and the viscous-fingering pattern was produced by Eshel Ben-Jacob of the University of Michigan. The Lichtenberg figure is from L. Niemeyer and H. J. Wiesmann of Brown, Boveri & Company, Limited, in Switzerland, and Luciano Pietronero of the State University at Groningen.

LIPOSOMES, which are fluid-filled lipid spheres, form spontaneously when appropriate concentrations of certain fatty molecules are mixed with water or with an aqueous solution. In order to prepare this electron micrograph, Pieter R. Cullis and Michael Hope of the University of British Columbia froze and fractured a batch of vesicles whose average diameter was .00015 millimeter.

Liposomes

These lipid spheres may offer a novel way to deliver medicines to diseased tissues. Instead of being diluted in the blood, drugs put in the vesicles would reach target sites in concentrated doses

by Marc J. Ostro

Any prescribed drug dose is the result of compromise. On the one hand, all drugs are potentially poisonous, suggesting that the least possible amount should be administered. On the other hand, drugs become diluted in the blood and large amounts are degraded, taken up by healthy tissues or excreted without ever reaching the site of disease. Such wastage increases the need for high doses. Physicians balance these opposing pressures by prescribing doses they think will be high enough to control the patient's problem but low enough to avoid causing unacceptable damage to healthy tissues.

To reduce the risk and inefficiency associated with such guesswork, many laboratories are now developing drug-delivery systems that alter the pathways by which drugs travel through the body. The goal is to deliver the needed dose of medicine to diseased tissues but to bypass healthy ones, thereby improving the drug's ratio of effectiveness to toxicity. One highly promising approach to the goal is the loading of medication into liposomes, which are microscopic sacs made of the very phospholipids that constitute cell membranes.

Liposomes can be filled with a variety of medications and, because of their similarity to cell membranes, are not toxic. They also protect their loads from being diluted or degraded in the blood. As a result, when the liposomes reach diseased tissues, they deliver concentrated doses of medication. Although investigators, including my colleagues and me at the Liposome Company, Inc., have yet to perfect the ability to target liposomes to diseased tissues at will, liposomes containing a variety of drugs have been shown in many animal studies, and in some clinical tests, to be more effective and less toxic than free drugs. Indeed, no fewer than 15 new therapies based on liposome technology may be ready to enter the market within the next decade.

Liposomes have reached the clinic only recently, but they are not a new invention. Alec D. Bangham of the Agricultural Research Council's Institute of Animal Physiology in Cambridge, England, inadvertently produced the first liposomes in 1961 while evaluating the effect of phospholipids on blood clotting. When Bangham put water in a flask containing a phospholipid film, the water forced the molecules to arrange themselves into what he later discovered were microscopic closed vesicles composed of a bilayered (two-molecule-thick) phospholipid membrane surrounding water entrapped from the environment.

Phospholipids form closed, fluid-filled spheres when they are mixed with water in part because the molecules are amphipathic: they have a hydrophobic (water-insoluble) tail and a hydrophilic (water-soluble), or "polar," head. Two fatty acid chains, each containing from 10 to 24 carbon atoms, make up the hydrophobic tail of most naturally occurring phospholipid molecules. Phosphoric acid bound to any of several water-soluble molecules composes the hydrophilic head. When a high enough concentration of phospholipids is mixed with water, the hydrophobic tails spontaneously herd together to exclude water, whereas the hydrophilic heads bind to water.

The result is a bilayer in which the fatty acid tails point into the membrane's interior and the polar head groups point outward. The polar groups at one surface of the membrane point toward the liposome's interior and those at the other surface point toward the external environment. It is this remarkable reactivity of phospholipids to water that enables workers to load medications into liposomes. As a liposome forms, any water-soluble molecules that have been added to the water are incorporated into the aqueous spaces in the interior of the spheres, whereas any lipid-soluble molecules added to the solvent during vesicle formation are incorporated into the lipid bilayer.

Liposomes employed for drug delivery typically range in diameter from 250 angstrom units to several micrometers (the diameter of a red blood cell is roughly 10 micrometers) and are usually suspended in a solution. They have two standard forms: "onion-skinned" multilamellar vesicles (MLV's), made up of several lipid bilayers separated by fluid, and unilamellar vesicles, consisting of a single bilayer surrounding an entirely fluid core. The unilamellar vesicles are typically characterized as being small (SUV's) or large (LUV's).

From the standpoint of the 1980's, drug transport seems an obvious application for Bangham's lipid bubbles, but in the mid-1960's their first allure was as a research tool. Physiologists, for instance, promptly enlisted liposomes as simplified cells for the study of ion transport across cell membranes, which are themselves lipid bilayers. Like membranes of cells, membranes of liposomes offer little or no barrier to water but are relatively impermeable to most solutes.

It did not take investigators long to realize that liposomes interact with cells in ways that could prove useful for drug delivery if the liposomes were able to reach diseased tissues. Under appropriate circumstances liposomes can adsorb to almost any cell type. Once they have adsorbed the spheres may be endocytosed, or swallowed up, by some cells. Adsorbed liposomes can also exchange lipids with cell membranes and may at times be able to fuse with cells. When fusion takes place, the liposomal membrane is integrated into the cell membrane and the aqueous contents of the liposome merge with the fluid in the cell [*see illustration on page 106*]. Most medical applications now under study exploit adsorption and endocytosis.

The ability of liposomes to adsorb to virtually any type of cell and to then release their contents slowly makes them excellent candidates for time-release drug-delivery systems. Multilamellar liposomes are particularly effective in this regard because the fluid in each successive layer is released only after the lipid membrane around it is degraded by the body or otherwise punctured. How quickly a drug is released from an adsorbed liposome depends on many factors, including the composition of the liposome, the type of drug encapsulated and the nature of the cell. Once it is released, a drug that normally crosses the membrane of a cell will enter the cell; other drugs will not enter.

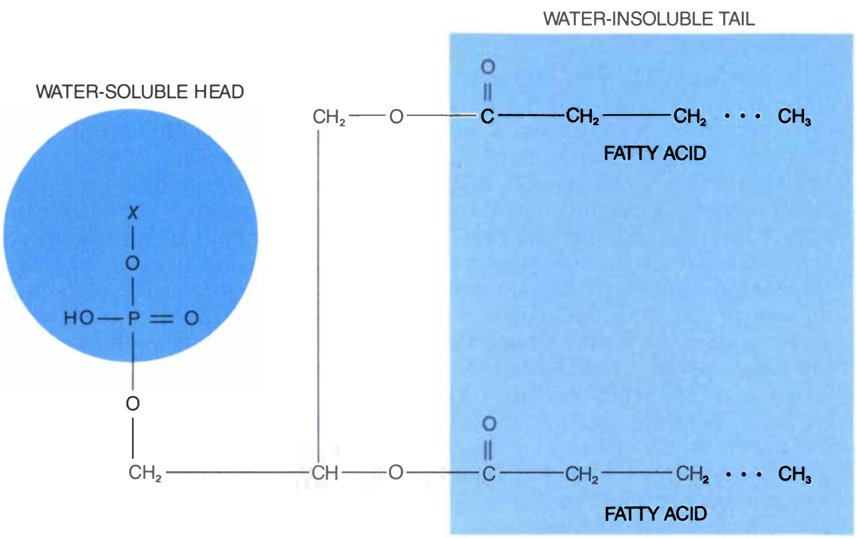
Endocytosis of liposomes occurs in a limited class of cells: those that are phagocytic, or able to ingest foreign particles. When phagocytic cells take up liposomes, the cells move the spheres into subcellular organelles known as lysosomes, where the liposomal membranes are thought to be degraded. From the lysosome the liposomal lipid components probably migrate outward to become part of the cell's membranes and other liposomal components that resist lysosomal degradation (such as certain medications) may enter the cytoplasm.

Lipid exchange involves the transfer of individual lipid molecules from the liposome into the plasma membrane (and vice versa); the aqueous contents of the liposome do not enter the cell. For lipid exchange to take place the liposomal lipid must have a particular chemistry in relation to the target cell. Once a liposomal lipid joins the cell membrane it can either remain in the membrane for a long time or be redistributed to a variety of intracellular membranes. If a drug was somehow bound to such an exchangeable lipid, it could potentially enter the cell during lipid exchange.

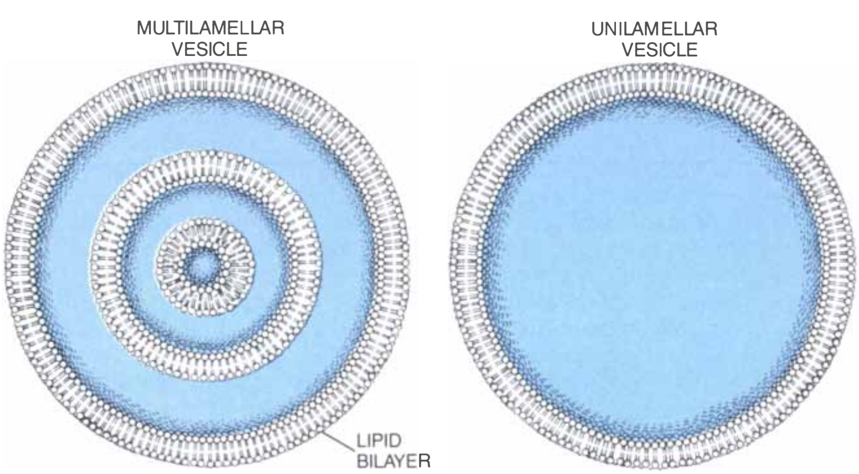
Fusion may or may not occur; the evidence is controversial. Some workers suspect that a fraction of the liposomes adsorbed to cells do fuse with those cells. Several years ago, when my colleagues and I at the University of Illinois Medical Center put liposomes containing RNA coding for rabbit globin in a cell culture, the liposomes seemed to insert the RNA into the cells. Moreover, the cells proceeded to synthesize rabbit globin, which indicated that the RNA had remained active. If the liposomes had entered the cells by endocytosis, lysosomes would have digested them and the RNA would have been destroyed.

In the late 1960's many investigators began to think liposomes might well prove to be efficient, targeted drug-delivery systems. Because liposomes could be made of the same phospholipids present in cell membranes, it seemed reasonable to assume that the spheres would be nontoxic and would escape recognition and removal by the body's immune system; the vesicles might therefore have the opportunity to interact with cells in diseased tissue in ways that would cause the vesicles to release their drug cargoes. If these assumptions were correct and if liposomes could, as expected, be readily loaded with drugs, all that would be needed to complete an ideal drug-delivery system would be the coupling of tissue-specific molecules to the outside of the drug-laden vesicles. Such molecules would then tow the liposomes to the target tissues.

By the mid-1970's liposomes made of lecithin (the phospholipid that has choline in its head group) had been shown to have no overt toxicity. Gerald Weissmann of the New York University School of Medicine and Deme-



PHOSPHOLIPID MOLECULE, shown in a generalized form, is the typical constituent of liposomes. Phospholipids are amphipathic: they have a hydrophilic (water-soluble) head and a hydrophobic (water-insoluble) tail. The head can be capped by any of several hydrophilic molecules (*X*), such as choline or serine. The tail consists of two fatty acid chains, each of which has from 12 to 24 carbons. When phospholipids encounter water, the hydrophobic tails associate to exclude the water and the hydrophilic heads face it. The result is water-filled vesicles formed by a two-molecule-deep membrane of lipid (a bilayer). The membrane of a living cell is also a bilayer and is rich in phospholipids.



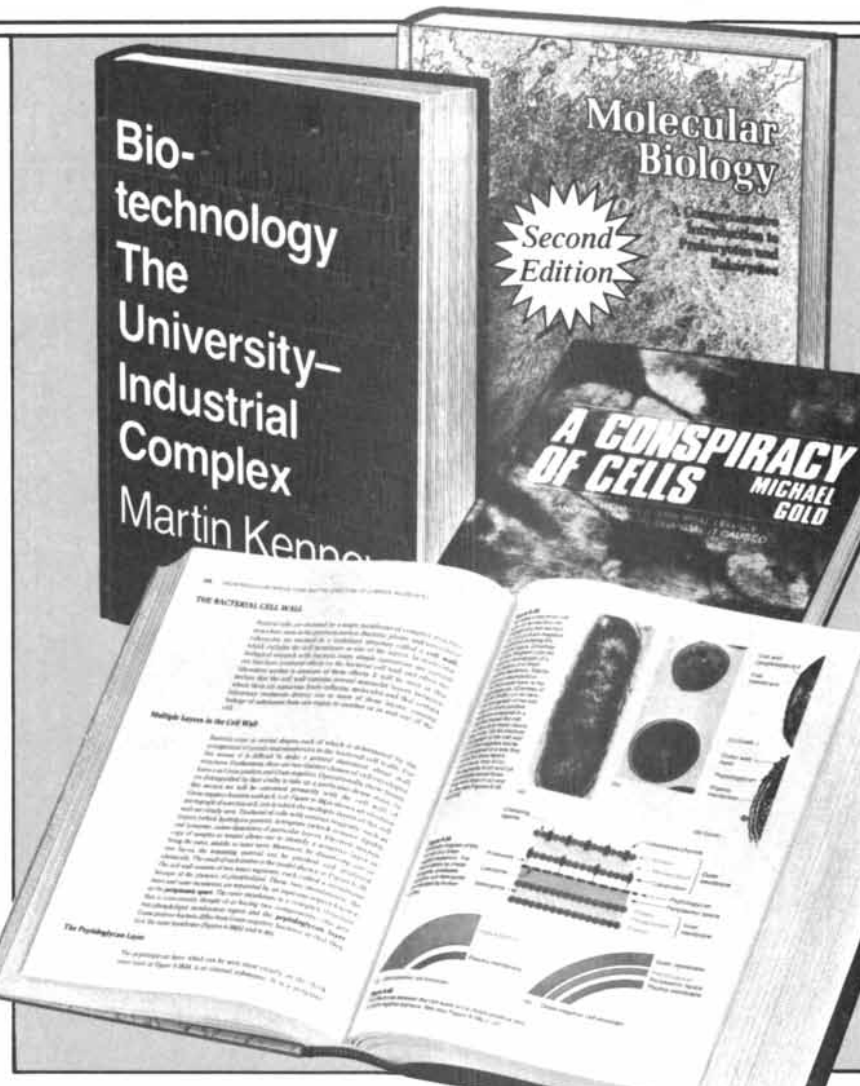
LIPOSOMES formed by mixing amphipathic lipids and an aqueous solution can be multilamellar (*left*) or unilamellar (*right*). Multilamellar vesicles have an "onion skin" structure in which concentric lipid bilayers are separated by aqueous layers. Unilamellar vesicles consist of a single lipid bilayer surrounding an aqueous interior. As liposomes form, water-soluble substances in the solution enter the aqueous spaces and lipid-soluble substances merge into the lipid bilayers, allowing one to load varied drugs into liposomes.

A $93.40 value

TAKE THE BIOTECHNOLOGY LIBRARY

for only $2.95 when you join THE LIBRARY OF SCIENCE.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.



MOLECULAR BIOLOGY—A Comprehensive Introduction to Prokaryotes and Eukaryotes, Second Edition *By David Freifelder*

"Remarkably successful...the coming years will see a long line of new editions."—*Nature*

This volume has undergone a complete revision. It's larger, more comprehensive and up-to-the-minute. The text, diagrams, charts and photographs have been carefully selected and prepared to give the best possible background in molecular biology. Rigorous, yet accessible, it covers basic molecular processes; the synthesis of DNA, RNA, and protein; and genetic phenomena in prokaryotic and eukaryotic cells, plus biotechnical techniques.

Retaining the winning qualities of the popular first edition, this second edition continues to offer both professional biotechnicians and biology students a crucial understanding of the chemical and physical components of macromolecular structure and synthesis...*without advanced training in chemistry or physics.*

Publisher's Price: $39.50

BIOTECHNOLOGY: THE UNIVERSITY-INDUSTRIAL COMPLEX *By Martin Kenney*

"Penetrating...an important contribution to the study of social and economic institutions."—*Sheldon Krimsky, Tufts University*

Who should control the new technology? Should safeguards and ethical standards be imposed? What does this industry's awesome potential pose for the future?

To answer these increasingly critical questions, Martin Kenney provides a timely, provocative overview of the entire biotechnology industry.

In the wake of continued controversy and confusion, Kenney reveals how the burgeoning biotech industry, born in the laboratories of American universities, has increasingly come under the powerful domination of big business interests.

Publisher's Price: $23.95

A CONSPIRACY OF CELLS—One Woman's Immortal Legacy and the Medical Scandal It Caused *By Michael Gold*

"A dazzling, readable and chilling thriller...the horror story of how one dead woman's terribly virulent cancer cells have adversely affected the 'war on cancer.'"—*Science Books and Films*

On October 4, 1951 Henrietta Lacks died of cancer. That is, most of Henrietta died. In a laboratory dish, a few cells taken from her cervical tumor continued to live—thrive, in fact.

Now, *A Conspiracy of Cells* presents the first full account of how her cells spread, until they infiltrated and destroyed a sensitive East-West scientific effort.

Publisher's Price: $29.95

The Library of Science is the nation's oldest, most respected science book club.

Dedicated to serving both professionals and serious amateurs, it surveys the entire field of published sources on scientific subjects to make the very finest works available to club members at substantial savings.

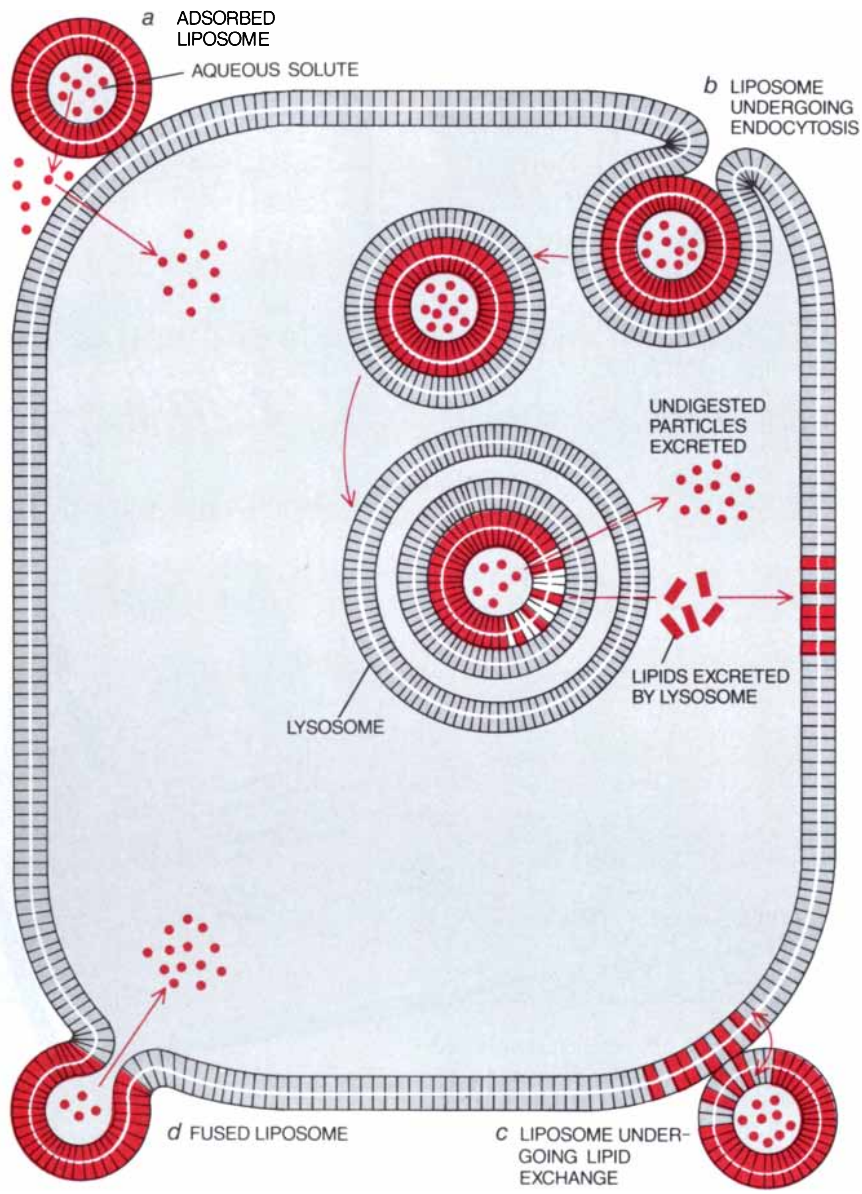
MEMBERSHIP BENEFITS • In addition to getting the Biotechnology Library for only $2.95 when you join, you keep saving substantially on the books you buy. • Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings of up to 65% off the publishers' prices. • At 3–4 week intervals (16 times per year), you will receive the Library of Science News, describing the coming 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 Science, Dept. 2-BS8-00754, Riverside, N.J. 08075, to obtain membership information and an application.

Scientific American 1/87

trios P. Papahadjopoulos, then of the Roswell Park Memorial Institute, had developed efficient ways to load all kinds of molecules, including entire enzymes, into liposomes. Individual liposomes were shown to be capable of carrying tens of thousands of drug molecules, and the goal of coupling cell-specific molecules to the outside of the drug-laden vesicles seemed to be close at hand: antibodies had been bound to the surface of liposomes and were shown also to bind tightly to target cells in culture.

Unfortunately two major problems were not solved, and they persist today. First, liposomes are for the most part unable to leave the general circulation and hence are unlikely to reach most cell types. Unencapsulated, or free, drugs typically diffuse through capillary walls into tissues, but liposomes are too large to pass through the walls of capillaries in most organs. Second, the body's immune system turned out somehow to recognize the liposomes and remove them from circulation. When liposomes enter the blood, circulating proteins (opsonins) apparently identify them as foreign and mark them for clearance by other parts of the immune system.



INTERACTION of a liposome and a cell can take several forms. A liposome can adsorb to almost any type of cell. An adsorbed liposome may be endocytosed, or internalized, by certain kinds of cells, may exchange lipid with the cell membrane or may fuse with it. After adsorbing to a cell (*a*) a liposome is likely to slowly release its contents, some of which may enter the cell, depending on the nature of the contents and the type of cell involved. An endocytosed liposome (*b*) is processed by a lysosome, an intracellular digestive organelle, after which the lipid components of the liposome are thought to be incorporated into the cell's membranes, whereas the aqueous solutes that escape lysosomal degradation may be incorporated into the cytoplasm. A liposome that undergoes lipid exchange (*c*) takes up lipid from the membrane of the cell and in return gives up some lipid to the cell. When a liposome fuses with a cell (*d*), the liposomal membrane merges with the cell membrane and the liposome's cargo becomes part of the cell's cytoplasm.

Liposomes are then removed from the circulation by cells that make up the reticuloendothelial system (RES). This system consists of macrophages and other highly phagocytic cells that are sprinkled throughout the lymph nodes, liver, spleen, bone marrow and lungs and that circulate in the blood and the lymph. Such cells take up liposomes from the circulation regardless of the vesicles' composition and size; moreover, liposomes somehow enter, and accumulate in, the extracellular spaces of reticuloendothelial organs, where the vesicles may persist for days or weeks before phagocytic cells endocytose and degrade them. Uptake by reticuloendothelial cells and organs probably prevents liposomes from circulating long enough to reach many targeted cells and tissues efficiently.

Although these obstacles may limit the ability of liposomal drugs to treat certain disorders, the lipid vesicles nonetheless show promise for treating many diseases and for boosting the effectiveness of certain experimental vaccines. For instance, when reticuloendothelial cells themselves become infected by bacteria or parasites, liposomes can be efficient, targeted drug-delivery systems. Few free drugs are able to enter such cells, making infections there difficult or impossible to treat by conventional means.

To explore the potential of liposomal treatment for infections of reticuloendothelial cells, Carl R. Alving of the Walter Reed Army Institute of Research and Chris D. Black of the Zoological Society of London independently tested the vesicles against leishmaniasis, a parasitic disease infecting an estimated 100 million individuals throughout the world. The disease, a major public-health problem, can be lethal if the parasites invade cells of the liver and the spleen and the infection goes untreated. Even with therapy many people die because the commonly prescribed drugs, known as antimonials, are related to arsenic: at high concentrations they damage the heart, liver and kidney. As with many other treatments for parasites, physicians often find themselves in the tragic position of hoping the drug will eliminate the infection before killing the patient.

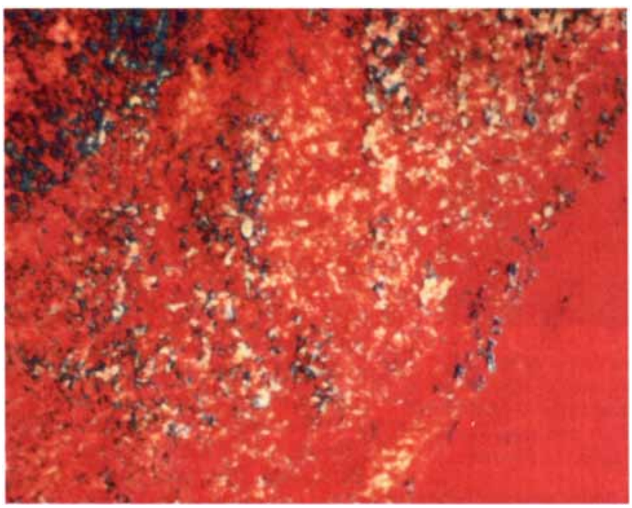
Alving and Black found that encapsulating antimonial drugs in liposomes dramatically reduced the dosage needed to treat leishmaniasis; one liposomal antimonial drug administered to leishmaniasis-infected hamsters was about 700 times as effective as the free drug. (In all studies described in this article the drugs are injected into a vein or into the peritoneal cavity.)

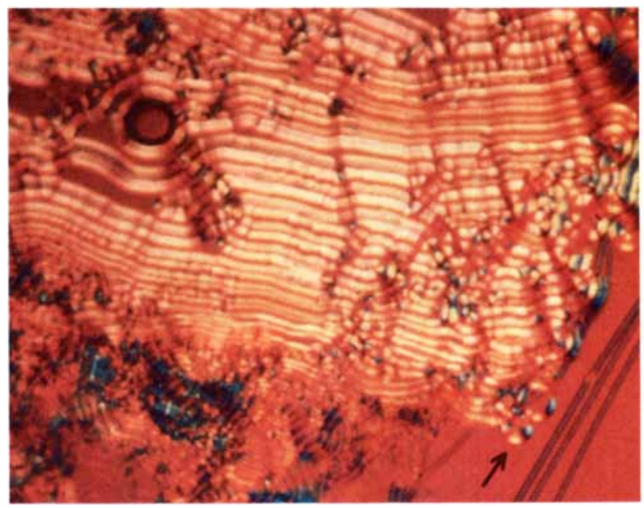
To Alving's surprise, electron micrographs of the liposomes in Kupffer cells (reticuloendothelial cells in the liver) revealed that after being swallowed by the cells at least some of the antimonial liposomes were somehow taken up, and presumably digested, by the parasites within the cells. In other words, the liposomes went directly to their parasitic targets, where the vesicles released their potent drug, destroying the parasites.

In the past five years workers investigating liposomes have treated a wide variety of reticuloendothelial infections in laboratory animals. They find that liposomes containing commonly prescribed antibiotics are particularly effective for treating brucellosis, listeriosis and acute salmonellosis, all bacterial diseases. Whereas no amount of streptomycin (an aminoglycoside antibiotic) corrected brucellosis in guinea pigs, one milligram of a liposomal version of the drug did produce a cure. Whereas it took 48 milligrams of free penicillin to cure listeriosis in mice, just .54 milligram did the job when it was encapsulated in liposomes. In other words, packaging drugs in liposomes can sometimes change ineffective treatments into effective ones.

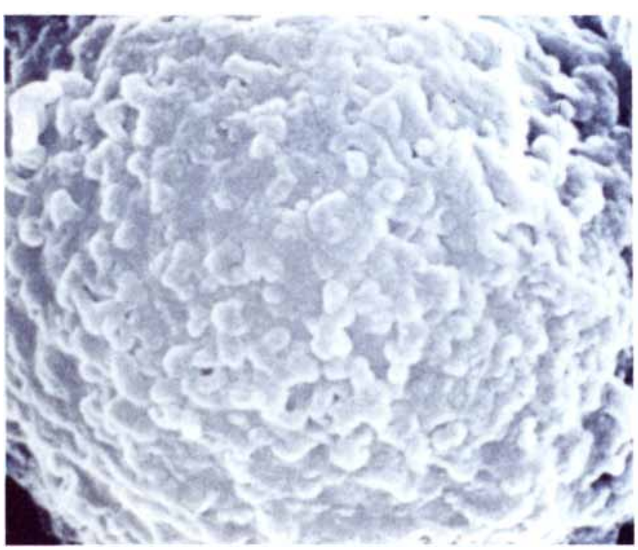
Liposomal drugs are effective against diseases of the reticuloendothelial system because its cells gobble up the liposomes. How well do the lipid spheres perform against diseases of cells and organs that are not a part of the reticuloendothelial system? For reasons that are not completely clear, liposomes can improve the effectiveness and reduce the toxicity of drugs administered for such diverse diseases as pyelonephritis (a bacterial infection of the kidney), systemic fungal infections and cancer.

Unlike brucellosis, pyelonephritis is

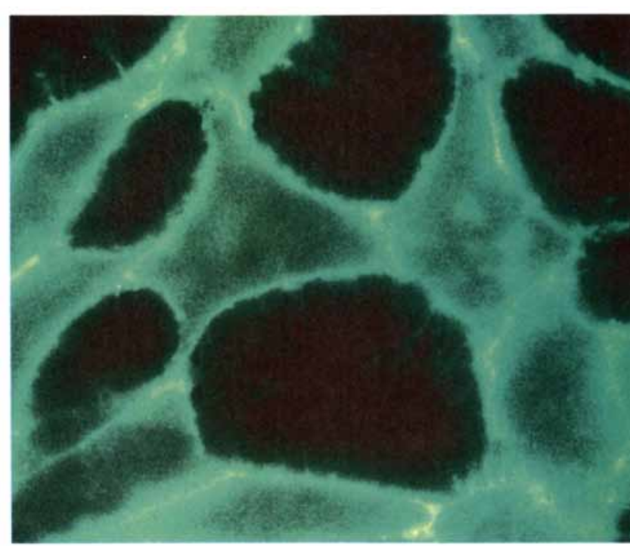




FORMATION OF LIPOSOMES begins immediately after a lipid film (*left*) is mixed with water (*right*). Two emerging liposomes can be seen in the bottom right-hand corner of the second micrograph. Eventually the entire film will form liposomes. The polarized-light micrographs were made by Leonard Estis of the Liposome Company, Inc.; the magnification is about 360 diameters.



MOUSE TUMOR CELL covered with liposomes (*white protrusions*) demonstrates the ability of the lipid vesicles to adsorb readily to cell membranes. The material is enlarged some 9,750 diameters in this scanning electron micrograph made by the author.



MEMBRANES of cells that have been incubated with liposomes are laced with green because the cells have taken up fluorescent lipids from the liposomes during lipid exchange. Richard E. Pagano of the Carnegie Institution of Washington made the image.

GAMMA-RAY SCAN, produced in the laboratory of Gabriel Lopez-Berestein of the University of Texas M. D. Anderson Hospital and Tumor Institute, reveals sites of liposomal uptake (*blue and white squares*) in a human subject 24 hours after intravenous injection. Liposomes gather in the lungs (*top left and right*), the liver (*bottom left*) and the spleen (*bottom right*). The bone marrow and lymph nodes (not shown) also take up liposomes.

treatable with free aminoglycosides, but the drugs must be taken for days or weeks. In addition physicians must control the free doses carefully because high concentrations can impair hearing or kidney function or both. (Aminoglycosides have an affinity for cells of the kidney and for the cranial nerve responsible for hearing.) In animal trials, when workers administered a single 16-milligram dose of drug per kilogram of body weight, the liposome-encapsulated medicine eradicated the infection whereas the free drug had little effect. At 32 milligrams per kilogram the free drug produced a 56 percent cure rate. Higher doses could not be administered because of drug-associated toxicity.

Systemic fungal infections are seen most often in people whose resistance is depressed by disease or by medications that suppress the immune system. Indeed, these infections can cause disability and death in victims of acquired immune deficiency syndrome (AIDS) and in cancer patients undergoing chemotherapy. The fungal infections very often defy treatment because the drugs that destroy the fungi are extremely toxic. In studies of mice, liposomal amphotericin *B,* a powerful antifungal agent, cured systemic fungal infections more effectively than the free drug did, largely because the investigators could give more of the drug to the subjects without increasing the toxicity. The free drug is highly toxic to the kidney.

Studies of human patients suffering from systemic fungal disease are even more striking. Gabriel Lopez-Berestein and his colleagues at the University of Texas M. D. Anderson Hospital and Tumor Institute at Houston found that liposomal amphotericin *B* given to 20 immunosuppressed patients for whom all standard therapies had failed effected a cure in 10 of the patients and resulted in substantial improvement in several others. Lopez-Berestein is now expanding his investigation to include a greater number of subjects.

A vast array of solid malignant tumors and leukemias are treated with a drug called doxorubicin, which attacks rapidly dividing cells. Doxorubicin's most serious side effect is progressive and irreversible damage to the heart. In addition the drug attacks hair-follicle cells, intestinal cells and cells of the immune system, leading to hair loss, nausea, vomiting and immune suppression. Laboratories in the U.S., Canada and Israel have demonstrated in rodents and dogs that liposomal doxorubicin is as effective as the free drug but is several times less toxic to the heart. More recently investigators have shown that the encapsulated drug also causes significantly fewer side effects in human patients.

No one mechanism is likely to explain all these data. A phenomenon known as RES loading, however, could well account for the effectiveness of liposomal aminoglycosides against pyelonephritis and the effectiveness of certain other liposomal drugs against many diseases caused by infectious agents. Macrophages and other reticuloendothelial cells circulate in the blood as the body's mobile defense force against invading pathogens. After a pathogen initiates an infection these cells rush to the infection site, where they internalize the pathogens and attempt to destroy them. Under normal conditions the number and strength of available reticuloendothelial cells may be inadequate to combat severe infections. It may be that if the cells endocytose liposomes containing antibiotic before heading to the infection, the cells are better able to destroy the pathogens they internalize. In essence the phagocytic cells will themselves have been effectively converted into a targeted drug-delivery system.

RES loading is a compelling theory, but it does not account for amphotericin *B*'s success as a treatment for fungal infections in immunosuppressed patients or for liposomal doxorubicin's success as a treatment for cancer. In the first instance the immunosuppressed patients lack circulating reticuloendothelial cells; in the case of cancer, macrophages are not drawn in significant numbers to cancerous cells, which often have ways of "hiding" from the body's immune system.

A feasible explanation for the success of liposomal amphotericin *B* against systemic fungal infections is that the blood vessels in the infected areas may be damaged, allowing liposomes to leak out of the circulation into the diseased tissue. Amphotericin *B,* which binds readily to the cell wall of fungi, is lipid-soluble and therefore resides in the membrane of the liposomes rather than in the aqueous spaces. Once the liposomes come in contact with fungi, the fungi probably pull the drug out of the liposomal membrane (in a process similar to lipid exchange) and are destroyed.

In the case of doxorubicin, it is conceivable that the liposomes become trapped in the intercellular spaces of reticuloendothelial organs, such as the spleen or the liver, where the lipid membrane is slowly degraded by enzymes. As the membrane degrades, doxorubicin may seep slowly into the blood, providing a small but continuous supply of drug to the tumor. There is some evidence that when free doxorubicin is infused into the body slowly, it remains effective but is less toxic than when given in intermittent high doses. Alternatively, certain tumors have been shown to be fed by weak, porous capillaries, which could allow liposomal doxorubicin to accumulate at the tumor site and slowly release free drug there.

On the whole, liposomal drugs are probably less toxic than free drugs simply because liposomes do not usually leave the general circulation except to enter reticuloendothelial cells and organs or sites where capillary networks are incomplete. If liposomes containing aminoglycosides do not reach kidney cells, and if liposomes containing doxorubicin do not reach heart cells, these cells are unlikely to be damaged by the drug within the liposomes. How then do the reticuloendothelial cells avoid being damaged by the liposomal drugs they ingest? The question has not yet been satisfactorily answered, although it appears that the cells can eventually degrade or inactivate several drugs, including doxorubicin. Indeed, such degradation may help to explain why liposomal doxorubicin, which presumably is taken up by phagocytic cells, is less immunosuppressive as a cancer treatment than the free drug.

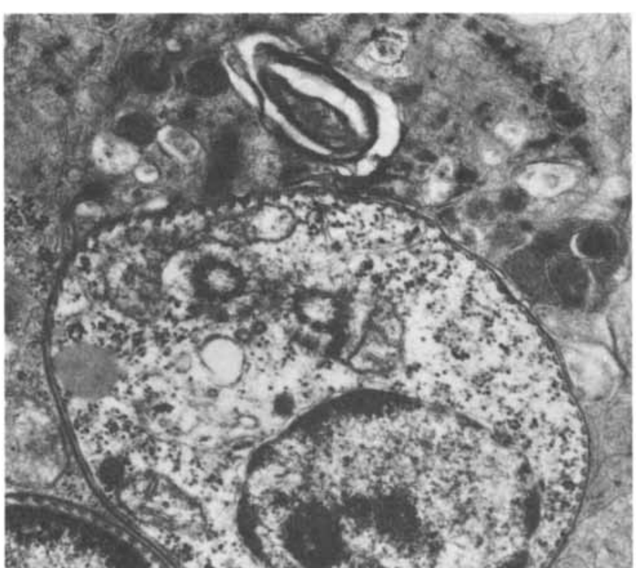
Clearly no single hypothesis will account for every success achieved by liposomal drugs. These drugs probably improve the effectiveness and reduce the toxicity of standard drugs by several mechanisms or combinations of mechanisms, some of which may not yet be known.

Disease prevention is as important as disease treatment, and liposomes may be a vehicle for producing safe and potent antiviral, antibacterial and antiparasitic vaccines. Ever since the first smallpox vaccine was produced some 200 years ago, vaccines have been made from nonvirulent or weakened strains of organisms. The body's immune system recognizes the pathogens as foreign and is thought to produce antibodies to them by at least two pathways. In one pathway, antigens on the pathogen surface presumably bind to receptor molecules on the white blood cells known as *B* cells, causing the *B* cells to become plasma cells, which proliferate and secrete antibodies specific for the pathogen. In another pathway, circulating macrophages bind to the pathogens, endocytose them and display processed antigens on their surface. *T* cells (another type of white blood cell) then bind to the expressed antigens. By way of several complex steps this binding ultimately results in further plasma-cell proliferation and increased antibody production.

Whole-pathogen vaccines apparently activate both of the described pathways of antibody production, eliciting a strong immune response, but they also pose risks. A small percentage of people injected with weakened or killed pathogens become ill with the original disease. More often, inoculated individuals suffer side effects that range from simple fever to Guillain-Barré syndrome, a relatively rare and usually temporary paralysis-inducing inflammation of the nerves. The syndrome has been associated with some influenza vaccines.

Because of the risks posed by whole-pathogen vaccines, many investigators have turned to recombinant-DNA technology and sophisticated peptide-synthesis techniques to produce small antigens found in the membranes of viruses, bacteria and parasites. After these subunits of pathogens are injected into animals, they can under some conditions stimulate the production of antibodies that later can inactivate viruses, bacteria or parasites displaying the antigens. Unfortunately the body's immune system does not respond strongly to most small proteins. In particular, macrophages do not readily ingest and process the small antigens.

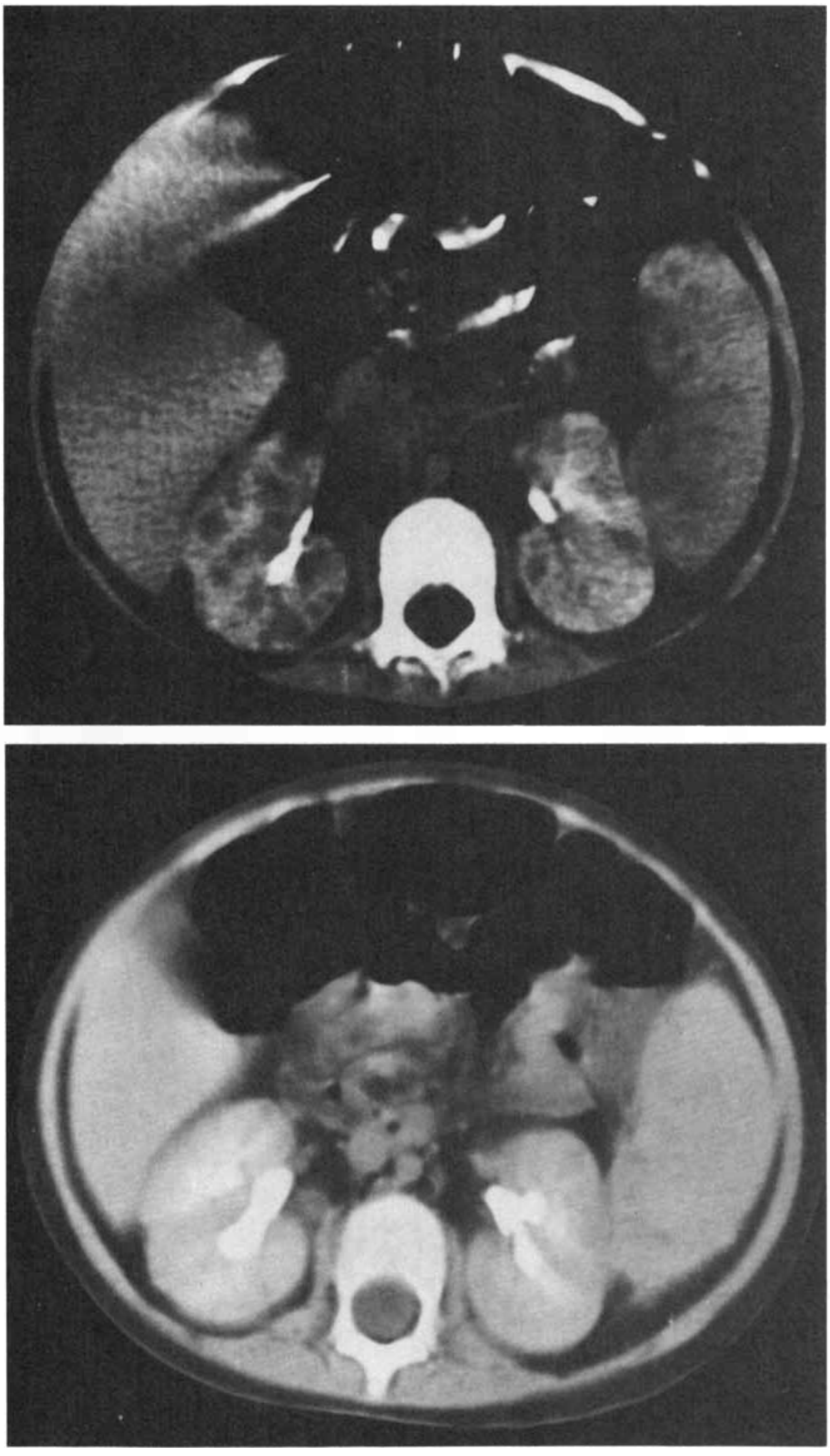
To improve the immune response of laboratory animals, investigators mix antigens with adjuvants: substances that stimulate the immune response. These adjuvants usually incorporate



PARASITE-INFECTED CELL in the liver of a hamster is seen soon after the cell endocytosed a liposome carrying an antiparasitic agent (*left*). The cell, a Kupffer cell, is part of the reticuloendothelial system, which is composed of phagocytic cells that remove large particles from the circulation. Once it was in the cell, the liposome (*small disk at top*) approached the parasite (*large sphere*) and adsorbed to it. Then the parasite somehow internalized the liposome (*right*), which in turn destroyed the organism. The thin-section electron micrographs were produced in the laboratory of Carl R. Alving of the Walter Reed Army Institute of Research.

bacterial extracts; they are acceptable for injection into test animals but cause severe local toxicity, such as inflammation, in human patients. Indeed, with the exception of one weak adjuvant, there are no effective adjuvants for human subunit vaccines.

Liposomes may solve the problem. When antigens are embedded in the surface of a liposome, they show enhanced ability to evoke a strong immune response. For example, liposomes incorporating lipid *A* (a nontoxic bacterial extract) in their membranes have been produced as adjuvants for antigens commonly found in the organisms that cause cholera, malaria, hepatitis *B* and salmonella. In every case the concentration of antibody secreted by test animals in response to the liposome-associated antigen was substantially higher than the concentration secreted when free antigen was administered. In many animals the liposomal vaccine resulted in a thousandfold increase in antibody production.

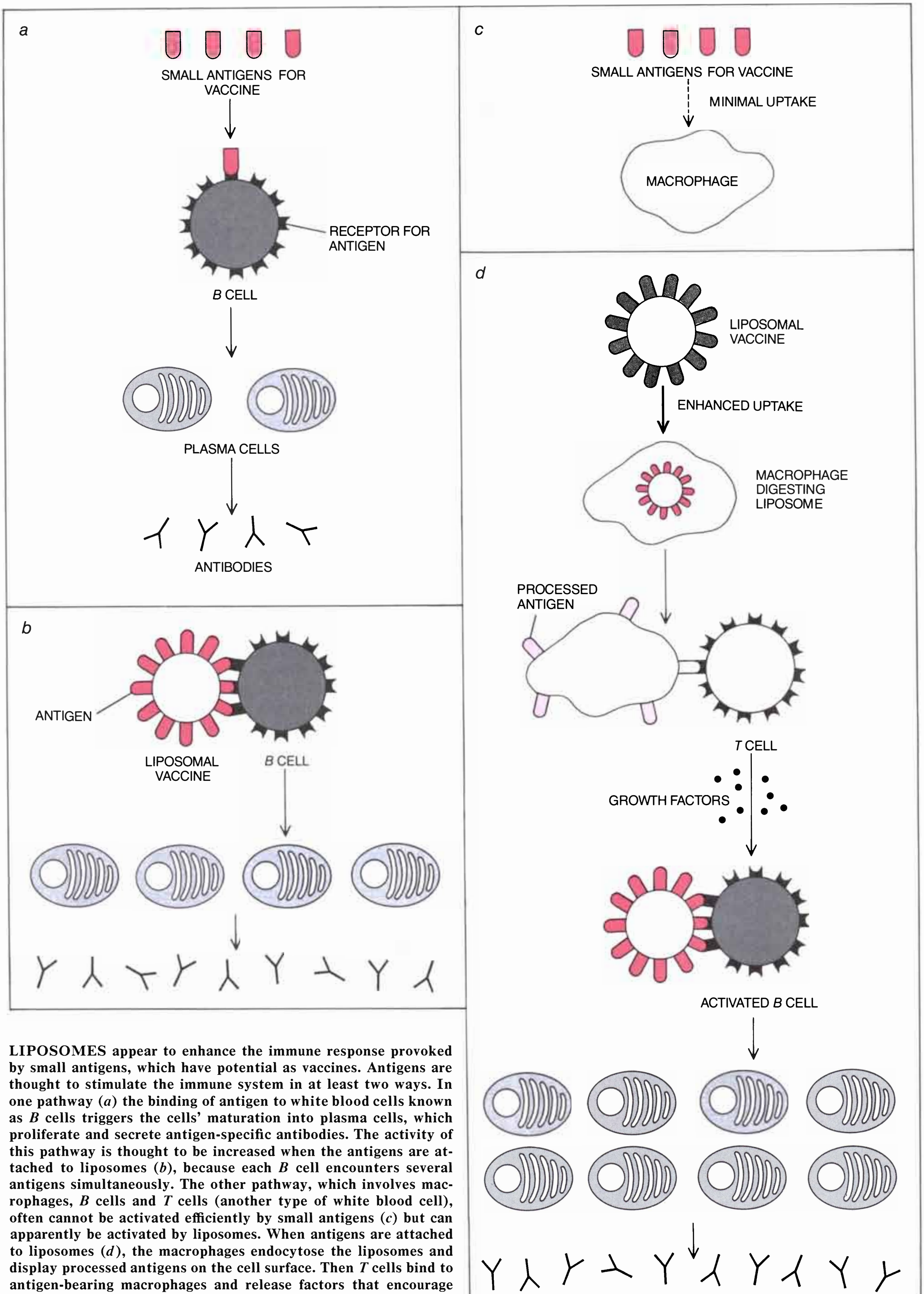


KIDNEYS (*flanking the spinal column*) of a two-year-old girl suffering from leukemia and a systemic fungal infection (*top*) seemed massively infected (*dark areas*) when examined by a CAT scanner in Lopez-Berestein's laboratory. After several courses of treatment with a liposomal antifungal agent (*bottom*) all signs of fungal infection disappeared.

Liposomal vaccines probably enhance antibody production by increasing activity along both pathways described above. When multiple antigens attached to a liposome bind to multiple receptors on a *B* cell, the resulting plasma cell probably proliferates faster than it does when it encounters a solitary antigen. Similarly, whereas a macrophage is unlikely to endocytose a small antigen efficiently, it will readily endocytose a liposome. When endocytosis takes place, antigens coupled to the liposomal surface or encapsulated within the vesicle are ingested and possibly displayed (as processed antigen) on the macrophage surface. Such presentation could result in *T*-cell activation and additional plasma-cell proliferation.

In the past two decades liposomes have progressed from being a laboratory curiosity to being a promising drug-delivery system. At one time it was thought that producing liposomes in standard sizes, stable forms and large, sterile batches might be impossible, but workers have overcome many of the obstacles to commercial application in several creative ways. For instance, liposomal preparations, which often have the consistency of milk, can now be freeze-dried to a powder and then reconstituted before being administered. "Empty" liposomes can be prepared and then loaded with drug immediately prior to use in order to avoid drug leakage. New techniques have also enabled workers to control the size of liposomes and to produce large (multiliter), sterile, reproducible batches of liposomal drugs.

In spite of these advances and extremely encouraging laboratory and clinical studies, liposomal drugs will not be widely available until investigators learn more about how the spheres are processed by the body and until drug-laden liposomes are proved to be safe and effective in extensive trials with human subjects. This will take time, but the work continues to progress. Liposomes hold great promise for vastly improving the effectiveness and safety of an array of important but potentially toxic medications.



LIPOSOMES appear to enhance the immune response provoked by small antigens, which have potential as vaccines. Antigens are thought to stimulate the immune system in at least two ways. In one pathway (*a*) the binding of antigen to white blood cells known as *B* cells triggers the cells' maturation into plasma cells, which proliferate and secrete antigen-specific antibodies. The activity of this pathway is thought to be increased when the antigens are attached to liposomes (*b*), because each *B* cell encounters several antigens simultaneously. The other pathway, which involves macrophages, *B* cells and *T* cells (another type of white blood cell), often cannot be activated efficiently by small antigens (*c*) but can apparently be activated by liposomes. When antigens are attached to liposomes (*d*), the macrophages endocytose the liposomes and display processed antigens on the cell surface. Then *T* cells bind to antigen-bearing macrophages and release factors that encourage antigen-bound *B* cells to become plasma cells, which proliferate.

The Evolution of the Fleece

Features of the sheep's coat and the rise of textile crafts interacted to motivate the thousands of years of selective breeding that led to modern fleece types

by Michael L. Ryder

The thick, woolly fleece of the domestic sheep is its distinguishing feature and the source of much of its economic importance. Yet only a moment, in evolutionary terms, has passed since the domestic sheep had a coat resembling that of many other wild animals. As recently as 8,000 years ago it was covered not in a white, continuously growing mass of wool but in a brown coat consisting of an outer array of kemps, or coarse hairs that are shed annually, and a fine woolly undercoat that also molted. Such an animal could not have supported the technology that has grown up around the domestic sheep—the shearing, dyeing, spinning and weaving of wool—any better than could a wild sheep such as the Bighorn of North America.

Much of the selective breeding that led to the fleece types known today took place in prehistory, and even the later developments went largely unchronicled. Yet other kinds of record survive, in three forms. Specimens of wool from as long ago as 1500 B.C. have been found, mostly as ancient textiles but also occasionally as sheepskin remains. Antique depictions of sheep in sculpture, relief and painting give even earlier clues to the character of ancient fleeces. The longest line of evidence takes the form of certain primitive breeds that are still tended in remote areas or that escaped from captivity long ago and now live in the wild. They retain the characteristics of ancient sheep, providing living snapshots of the process that gave rise to modern fleeces.

The evidence shows that the development of the fleece cannot be understood without reference to the rise of textile crafts, technologies that developed in step with biological changes in the coat. In some cases innovations in technology led to selective breeding for particular coat characteristics. In other cases the influence flowed the opposite way, as particular features of the sheep's coat inspired new ways to make use of it.

Because the coat of wild sheep looks no more promising as a source of textile fibers than the coat of other wild animals, sheep are not likely to have been domesticated for their wool. Like most other game, wild sheep served a range of human needs: they provided meat and fat for food, bone and horn for tools, gut for containers and skin for clothing. It is reasonable to speculate that their value as game led to an increasingly close association between Stone Age peoples and the wild flocks.

A form of game management, in which hunters followed individual flocks and culled them rather than killing wholesale, may have been the first step in this gradual and unwitting process of domestication. Eventually the herds came to be penned and protected. The process was complete at least 11,000 years ago; bones thought to be from domestic sheep are found at sites of the Neolithic period (the time of the first agricultural settlements) along the foothills of the Zagros Mountains, which straddle the border between Iran and Iraq. Along with the goat, which was domesticated at about the same time, sheep became the first livestock.

Once sheep had been domesticated their biological destiny was shaped by human needs. Those influences were slow to produce a woolly fleece. On Corsica and Sardinia there survives a breed of sheep, the Mouflon, that was once believed to have always lived in the wild but now is seen as the feral progeny of domestic sheep that accompanied the first settlers to those islands in about 6000 B.C. The coat of the Mouflon consists of fine underwool and coarse kemps, and it differs little from that of truly wild sheep such as the Bighorn. It appears that not much change in coat structure took place for 3,000 or 4,000 years after the domestication of the sheep in the Middle East. (Indeed, the so-called hair sheep of Africa and India still bear a kempy coat like that of the Mouflon, in spite of long domestication, presumably because there is no demand for wool in a tropical climate.)

During that earliest period of domestication sheep almost certainly served primarily as a source of meat. They also yielded skins for clothing, and it is conceivable that a certain amount of unintentional selective breeding for a softer coat took place even at this stage. Animals with an exceptionally fine, soft coat may have been prized and kept for longer before they were slaughtered than those with a coarse coat, enabling them to leave more offspring.

Even if it does not provide some renewable product such as wool, a domestic animal is more valuable alive than dead. Alive it confers prestige and represents economic security in the form of a movable store of food; dead it is worth no more than its flesh, bone, horn and skin. Perhaps it was in an effort to conserve their flocks that the early pastoralists began to take nourishment from them by milking or bleeding the animals rather than slaughtering them. The same impulse may have led to the first use of textiles as a substitute for skins.

Wool felt was almost certainly the earliest fabric made from animal fibers. When it first appeared in the archaeological record is a matter of dispute. Excavators at the site of Çatal Hüyük in Turkey, which dates from 6500 B.C., have reported finding samples of felt, but the cloth I have examined from that site is linen. The earliest generally accepted felt samples appear much later. Still, the craft of feltmaking is far simpler than that of spinning and weaving. Indeed, it may

MOUFLON AND MERINO stand at opposite poles of fleece development. The wild Mouflon sheep (*top*), native to Corsica and Sardinia but shown here in West Germany, is a living relic of early domestic sheep. It is thought to have accompanied the first human settlers to the islands in about 6000 B.C. and then to have escaped and survived in the wild. Its brown coat, consisting of an outer layer of coarse hair called kemps and a fine undercoat, resembles that of truly wild breeds. Like those sheep, the Mouflon molts annually. The Merino (*bottom*), a breed that emerged in Spain during the Middle Ages and now is raised in many parts of the world, represents a pinnacle of fleece specialization. Its wool is uniformly fine and white, and the breed does not shed its coat.

have had a direct natural inspiration.

Feltmaking depends on the microscopically scaly surface of wool fibers. When the fibers are rubbed together in a mass, the scales allow individual fibers to move in one direction only, and the mass becomes irreversibly tangled. Warmth, which relaxes the fibers, and wetness, which lubricates them, speed the felting process. As a result it occurs readily in the fibers shed by a molting sheep. In primitive sheep such as the Mouflon, which typifies the kind of animal whose hair was first made into felt, I have seen that the molting underwool often gets tangled with the kemps of the outer coat and mats together before falling from the animal in felted masses. A similar observation may have inspired the first feltmaking.

Once sheep began to supply fibers for textiles, the impetus for selective breeding to improve the coat grew. It became even stronger when, still in prehistory, primitive peoples learned how to spin fibers into yarn, which could then be woven into fabric. Spinning and weaving surely antedate the oldest surviving fragments of wool cloth, found in a bog in Denmark, which were made in about 1500 B.C. Although older scraps of linen cloth have been found, there is reason to think the techniques originated with wool. Obtaining fiber from flax is a lengthy process, whereas the hair shed by a molting animal can be spun directly into yarn.

The crucial discovery was that twist imparts strength to a strand of fibers—a discovery that, like the invention of feltmaking, may have been aided by the sight of a primitive sheep during the molt. As the animal rubs the shedding hairs from its coat it sometimes twists them into long bundles resembling coarse yarns. I was able to collect 20 such "yarns," some of them more than a meter long, from a single Mouflon sheep as it molted. It seems reasonable to suppose Stone Age peoples, already adept at weaving plant stems into basketry, may have tried to weave such natural yarns into crude textiles. Indeed, I was able to make a rudimentary piece of cloth from the yarns I had gathered.

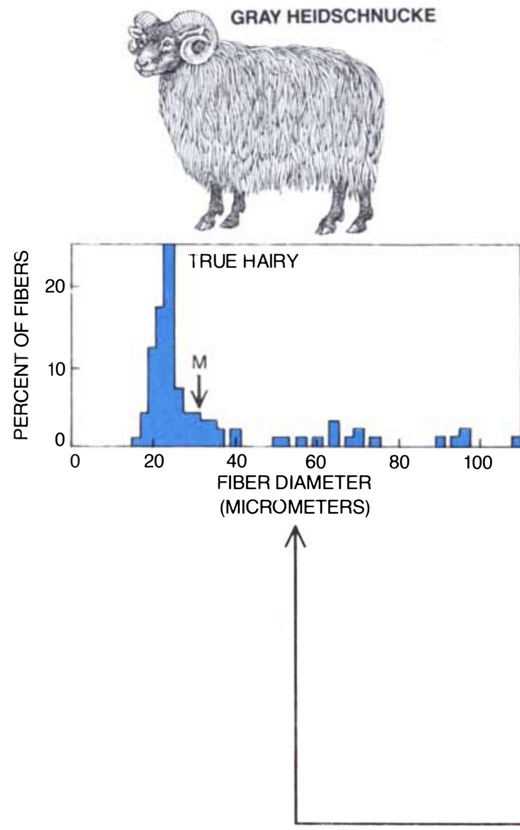
Like the origin of clothmaking, the associated biological transformation of the kempy coat of the early domestic sheep into a primitive fleece cannot be dated precisely or reconstructed in detail. The end point of the process, however, is recorded in the earliest wool textiles, in other archaeological artifacts and in a surviving breed of primitive sheep. The evidence makes it possible to reconstruct the transformation in outline.

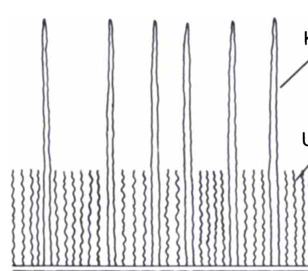
Over the several thousand years following the stage represented by the Mouflon sheep the coarse kemps of the original outer coat became progressively finer under the influence of selective breeding. At the same time, in a response that may have been biologically correlated, the fine underwool of the primitive sheep became somewhat coarser. The mean fiber diameter in the underwool of wild sheep is about 15 micrometers (15 thousandths of a millimeter); in most of the early textile remains the diameter of the fine fibers averages about 20 micrometers, which has remained the typical value for fine wool since then. The Danish specimens seem to record an intermediate stage; the kemps they contain are finer than those of the wild coat, but the average diameter of the fine underwool fibers is still about 15 micrometers.

These trends had yielded two kinds of primitive fleeces by the beginning of the Bronze Age, about 3000 B.C. in the Middle East and 1500 B.C. in Europe. The "hairy medium" fleece, in which the kemps had lost their bristly coarseness and had come to resemble hairs growing from a thicker woolly undercoat, developed first. Further reduction in the diameter of the kemps changed them into medium wool that merged with the finer undercoat, producing a "generalized medium" fleece.

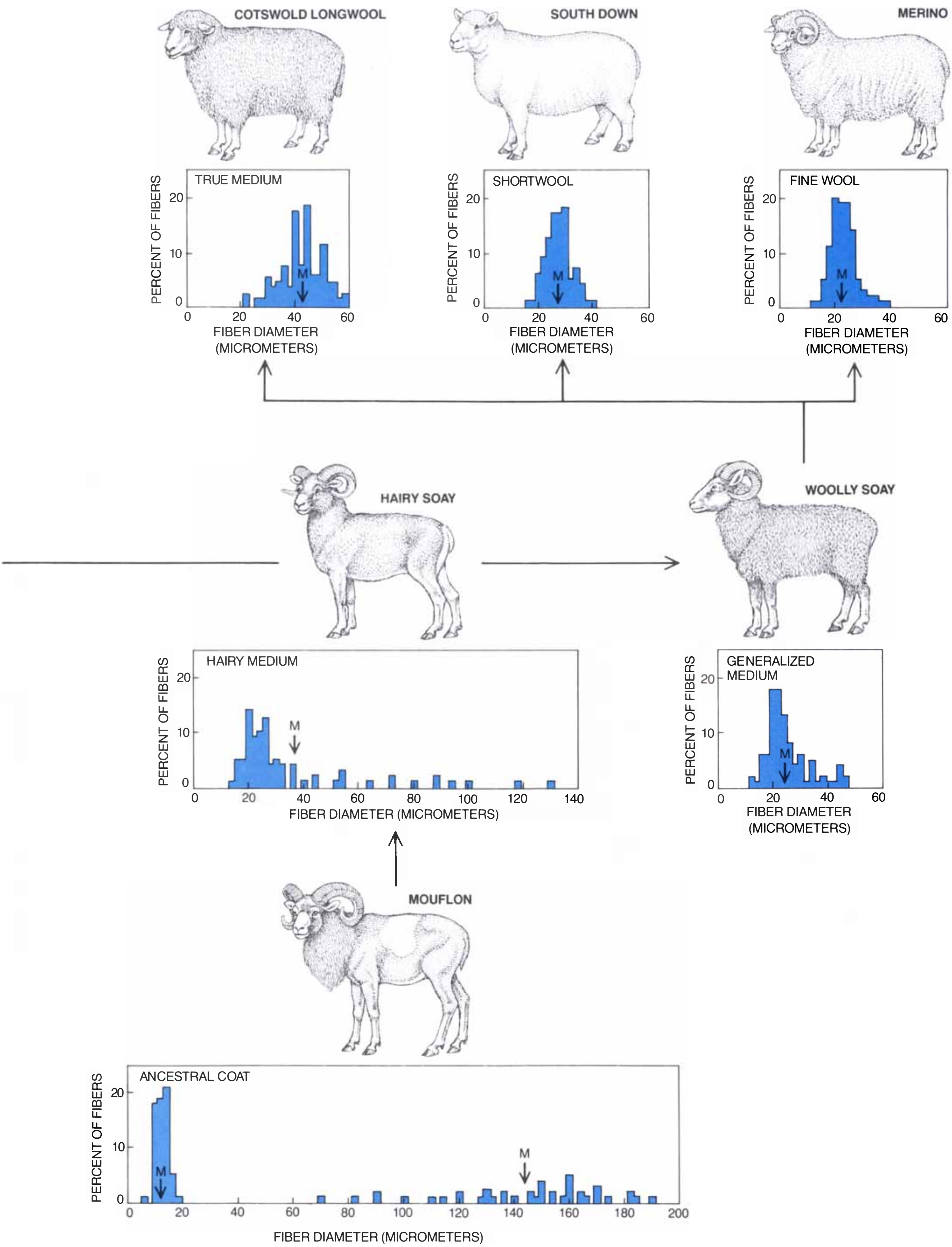
These earliest fleeces may have been spun into different kinds of yarn. Although worsted yarn, in which most of the fibers are parallel, was once thought to have been introduced in medieval England, it may well have been developed in prehistoric times along with the hairy medium fleece. Such a fleece, which contains many hairs growing in parallel, could simply have been combed to prepare it for spinning. Worsted yarn would have resulted from the highly aligned fibers. In a woolly fleece such as the generalized medium, in contrast, the fibers are tangled and difficult to comb. Instead such wool is usually carded to tease the fibers apart; spinning then yields a true woolen yarn, in which the fibers are oriented randomly.

The archaeological evidence for the emergence of these two fleece types varies from region to region. The record is longest in the Middle East, where a crude clay image of a sheep, found at Sarab in Iran and dated to about 5000 B.C., bears what may be the first hint of a fleece. The kempy coat of a wild animal is smooth in appearance, but a fleece tends to look clumpy: the wool fibers naturally twist together into distinct bundles known as staples. In sheep with a primitive, hairy fleece the staples are pyramidal





COAT STRUCTURE became transformed over perhaps 6,000 years. The graphs consist of bars indicating the percent of the coat fibers falling within two-micrometer increments of diameter in primitive and modern fleece types; a sheep from a representative existing breed surmounts each graph. The trend with respect to the original coat, diagrammed above, is one of convergence: the coarse kemps have grown finer, eventually changing to wool. At the same time the fine underwool has coarsened a little. Hence the range of fiber diameters is narrow in modern woolly fleeces (*top of opposite page*). The mean diameter (*M*) is coarsest in the true medium fleece, intermediate in the shortwool and finest in the fine wool. A fourth modern fleece, the true hairy, developed by another route. Its fiber-diameter distribution differs little from that of the hairy medium fleece, a primitive type, but the coarsest fibers have changed from kemps, which are shed annually, to hairs, which grow continuously.


COTSWOLD LONGWOOL
SOUTH DOWN
MERINO
TRUE MEDIUM
SHORTWOOL
FINE WOOL
PERCENT OF FIBERS
FIBER DIAMETER (MICROMETERS)
HAIRY SOAY
WOOLLY SOAY
HAIRY MEDIUM
GENERALIZED MEDIUM
MOUFLON
ANCESTRAL COAT
M

in shape; the point is formed by the long, coarse hairs and the base is filled out by the more numerous, shorter and finer fibers of wool. The coat of the Sarab sheep is represented by a series of V's, which could correspond to the staples of a hairy medium fleece.

Even though fleeced sheep may have appeared as long ago as 5000 B.C., relics from 2,000 years later suggest that Stone Age sheep with their kempy coat were still kept in some areas. Early figurines and reliefs from Mesopotamia depict some sheep with locks of wool that must be staples and others with what I interpret as the kempy coat, shown smooth or as a series of parallel streaks. The early dynastic period in Egypt, moreover, yields images of the Stone Age sheep only. In Egypt the first clear representations of staples—the V-shaped ones that suggest a hairy medium fleece—do not appear until much later: the first millennium B.C.

In the Middle East the remains of woven fabrics are rare until Roman times, perhaps because conditions did not favor the preservation of wool or because the early archaeologists, interested primarily in artistic treasures, paid little attention to cloth relics. In northern Europe, in contrast, evidence of Bronze Age fleeces takes the form of textile remains more often than of artistic representations. The bogs from which many remains have been excavated are acidic and low in oxygen, and both factors prolong the survival of wool. The textiles reveal not only the fiber diameters of Bronze Age wool, which are those of hairy medium and generalized medium fleeces, but also its brown color, retained from the ancestral wild sheep.

Bronze Age fleeces need not be studied through relics alone; they can be examined firsthand. On the islands of St. Kilda, which lie 40 miles west of the Hebrides, live the Soay sheep, feral descendants of sheep brought to the islands by early settlers. They display the brown upper parts and the white belly of ancestral sheep and have both hairy and woolly fleeces. Fiber diameters indicate that the Soay fleeces correspond respectively to the hairy medium and generalized medium fleeces preserved in Bronze Age textiles. Those parallels, together with skeletal resemblances between Soay sheep and ancient sheep remains, confirm that the animals represent the kind of sheep kept in Bronze Age Europe. It is significant that the relic breed survives in the extreme northwest corner of Europe, well out of range of the waves of improved sheep that spread into the continent from the Middle East.

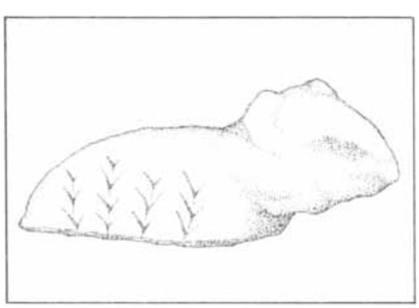
The hairy medium and generalized medium fleeces of the Bronze Age persisted into the Iron Age, which began in about 1500 B.C. in the Middle East and about 750 B.C. in northern Europe, but textile remains and surviving breeds indicate two further changes in the 1,000 years or so B.C. Sheep began showing a greater range of natural color; Iron Age textiles were woven not only of brown wool but also of black, white and particularly gray wool, from fleeces in which pigmented and white fibers are mixed. Later some breeds lost the tendency to molt and began to keep their coats from year to year. The development of new fleece colors, white in particular, and of continuous wool growth were intimately related to two technological steps: the advent of dyeing and the invention of sheep shears.

The colors of Iron Age wool can still be seen in the fleece of sheep surviving in isolated areas of Europe: the Shetland and Orkney breeds found on those two groups of islands northeast of Scotland, the Goth breed from the island of Gotland in the Baltic Sea, the native sheep of the French island of Ushant and the Drama breed of Greece. Anatomical correspondences with Iron Age sheep remains, a tendency to molt annually and the predominance of gray animals in the flocks indicate that these sheep are living relics of the Iron Age.

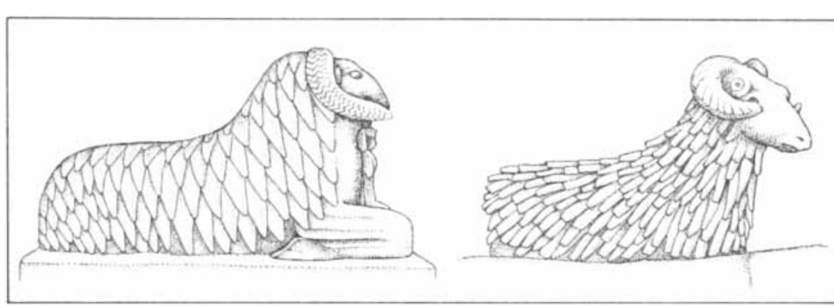
The emergence of such varicolored sheep from the uniform brown of Bronze Age animals may reflect the influence of long domestication on the genetic makeup of the species. In wild sheep variations in color may have been selected against, since a brown coat may have offered the best camouflage from predators. Under human protection, the full diversity of coat colors could show itself. The novelty of white, black or gray wool may then have led to selective breeding to maintain the variety of colors.

A new stimulus for breeding sheep with an eye to coat color arose in the Middle East by about 1000 B.C.: the development of dyeing. The origin of the craft cannot be assigned a precise date or location, but the scale of the industry into which it grew is evident. The Phoenicians, who lived in what is now Lebanon between about 1400 and 400 B.C., are even thought to have wiped out two species of murex shellfish, the source of the dye Tyrian purple, from the eastern Mediterranean; mounds of their discarded shells mark the sites of many Phoenician colonies. A dye industry implies a demand for white wool, and textiles from the first millennium B.C. in the Middle East indicate an increase in the number of white sheep. Nearly all the remains are of dyed white wool.

Even if the advent of dyeing stimulated the breeding of white sheep, the technology probably owes its exis-



FIRST GLIMPSE OF A FLEECE in the archaeological record may be a 7,000-year-old clay figurine from Sarab in Iran. The V-shaped markings on its flanks may represent staples, the locks of wool seen in a fleece but not in the coarse coat of wild animals and very primitive domestic sheep.



ANCIENT SHEEP FIGURINES show staples whose shape hints at the fleece type. An artifact from 680 B.C., found at Kawa in the Sudan, depicts pointed staples (*left*). The sheep probably had a hairy medium fleece, in which short, abundant wool fibers broaden the staple base and longer, sparser kemps extend the tip. A representation from Sumeria, dating from about 3000 B.C., seems to show the other kind of primitive fleece, the generalized medium, in which the kemps have changed into medium wool (*right*). Because the fiber length as well as thickness is more uniform in that fleece, it forms blunter staples.

tence to a biological precondition: the presence of white animals in the varicolored herds that had developed earlier. In the absence of white wool there would have been no reason to extract dyes, and as in the case of feltmaking and spinning the properties of the wool itself could have suggested the new practice. A white wool fabric accidentally stained with juice from fruits or berries could well have inspired the first experiments in dyeing.

Technology and biological change may have interacted differently in the case of the introduction of shears and the development of continuous wool growth. Before the invention of shears in the Middle East in about 1000 B.C., wool could be gathered only by plucking sheep during their molting season. Roman writers remarked on the practice among barbarian tribes, to whom the use of shears had not yet spread. On the Shetland Islands, where the local Iron Age sheep still molt in the spring, "rooing," or sheep plucking, persisted until early in this century.

In gathering wool from a primitive fleece containing coarse kemps plucking does offer an advantage over shearing: the kemps tend to be shed later than the fine wool underlying them, so that plucking can be timed to yield mostly fine wool. To get the fine fibers from a molting sheep one can also comb the fleece instead of plucking the animal by hand. The method is used in northwestern China to remove the fine underwool of cashmere goats and to harvest wool from the hairy sheep that are kept in the same region.

Combing a molting sheep is a technique that may have long antecedents. Excavations at the Iron Age village of Glastonbury in England have unearthed bone artifacts that are called weaving combs because of their resemblance to the tools with which a weaver compacts the weft. Actually their curved cross section makes that use unlikely. The combs may have served instead in the preparation of hairy medium wool for spinning into worsted yarns, but it is also conceivable that they were used to comb wool from molting sheep.

Waiting for sheep to molt is an inefficient way to harvest wool, however. Some sheep molt prematurely and lose their wool before the normal plucking time; others are not ready to be plucked until after the wool harvest. Nevertheless, selective breeding to eliminate molting is not likely to have begun until there was some other way to remove the fleece. Hence a technological step—the invention of shears—must have been the precondition for the biological alteration. Shears big enough to clip a sheep, consisting of two blades working against each other and joined by an iron bow that acts as a spring, are a common find at late Iron Age and Roman sites across Europe. Textiles and sheepskin remains from the same period record a change in the structure of the fleece.

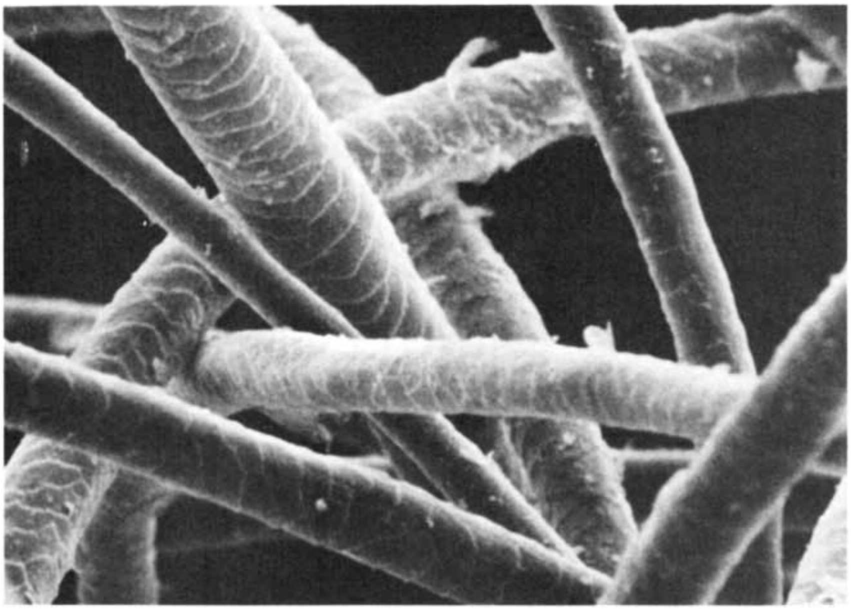
PLUCKING OF SHEEP as they molted was the only way to harvest wool until sheep shears were invented, in about 1000 B.C. The photograph, made in 1892, shows spring "rooing," or plucking, of primitive sheep kept on the Shetland Islands. Their annual molt sets that breed apart from modern breeds, whose wool grows continuously. For the sheepkeeper molting inevitably entails some wool loss. Breeding for continuous growth probably began soon after the advent of shears, which made it unnecessary to wait for molting.

The "hairs" of the hairy medium fleece, one of the two primitive fleece types, are in fact kemps. By definition they are shed annually, and their short length and coarseness make them easy to recognize in fragments of ancient cloth. A scrap of sheepskin recovered from a Scythian burial mound in Siberia, dating from about 400 B.C., and textile samples from Europe show that late in the Iron Age the kemps of the hairy medium fleece were transformed into true hairs in some breeds. Whereas kemps stop growing in winter in preparation for molting in the spring, hairs merely become thinner, continue to grow and do not shed. Hence these continuously growing fibers can be distinguished from kemps by their length and fineness.

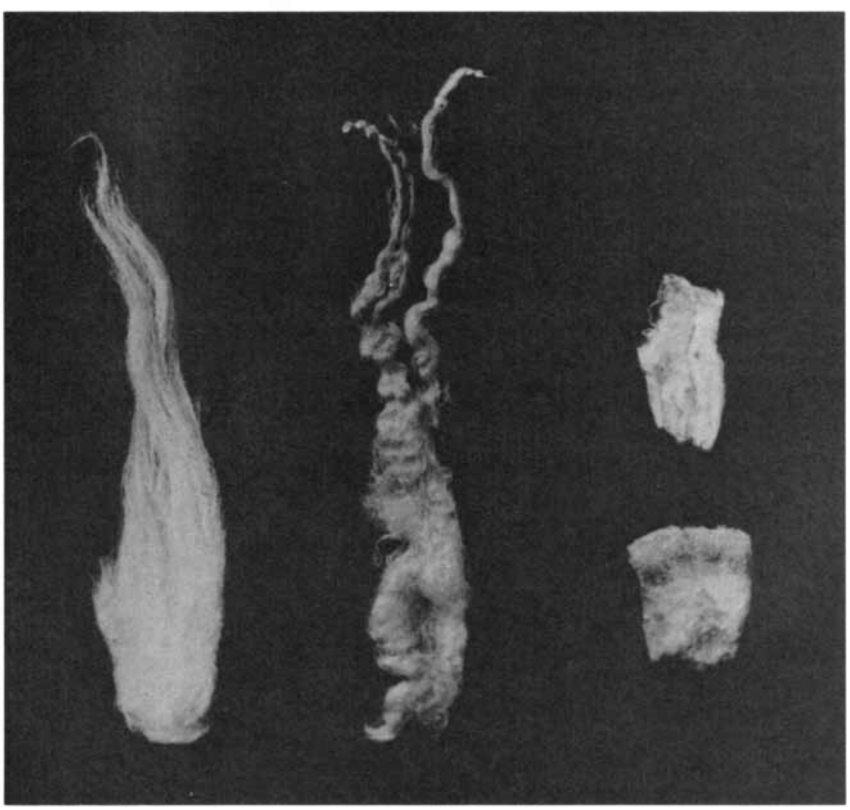
The appearance of sheep hairs marks the emergence of the kind of fleece known as the true hairy fleece, which today provides wool for carpets and tweeds. The change of the kemps of a hairy medium fleece into the hairs of a true hairy fleece is only one result of selective breeding for continuous growth. It was probably the availability of shears that also led sheepkeepers to try to eliminate molting in breeds that had the other kind of primitive fleece, the generalized medium. The kemps of the generalized medium fleece had already thinned into medium wool fibers, however, and so the change from molting to continuous growth cannot always be identified in ancient wool textiles as a change in the fiber dimensions.

Textile remains from Roman times are numerous, and they record the final transformations of the fleece. Three paths of development seem to have diverged from the ancient woolly fleece (the generalized medium), leading to the various kinds of wool produced today. Along one path the medium fibers, which had developed from the coarse kemps of the wild coat, became even finer, yielding a coat consisting only of fine fibers in a narrow range of diameters. The new fleece, known as fine wool, is today the major source of wool for clothing. Another path involved the opposite transformation: the fine fibers of the ancestral fleece became coarser, changing into medium fibers and producing a uniformly medium fleece: the true medium fleece, suited to heavy wear in blankets and rugs. Along the third path the extremes of fiber diameter in the generalized medium fleece converged, giving a semifine fleece: the modern shortwool, which is used for hosiery and some knitwear.

The fine wool seems to have been the first of these modern fleeces to emerge. It probably developed in the



MICROSCOPIC SCALES pattern wool fibers, shown here in a scanning electron micrograph of felt. The scaly surface is responsible for the process of felting: when wool fibers are rubbed together, the scales allow movement in one direction only, and eventually the fibers become locked in a dense tangle—a felt. Felting can occur naturally in the coat of a molting sheep, and the natural process could have inspired the first human attempts to make felt, probably the earliest wool fabric. The micrograph was made by Thea Gabra-Sanders of the Teaching and Research Centre at Western General Hospital in Edinburgh.



WOOL STAPLES from modern fleece types vary in the length and coarseness of their fibers. The long staple at the far left comes from a true hairy fleece, which consists of two kinds of fibers: long, straight hairs growing from a thicker, finer mass of wool. In the next staple, from longwool, a true medium fleece, the fibers are all long, coarse and woolly. The short staple at the lower right represents shortwool (a semifine fleece) and the staple at the upper right represents fine wool (the finest fleece). Shortwool and fine wool have a narrow range of fiber diameters and lengths. As a result they form squared-off staples.

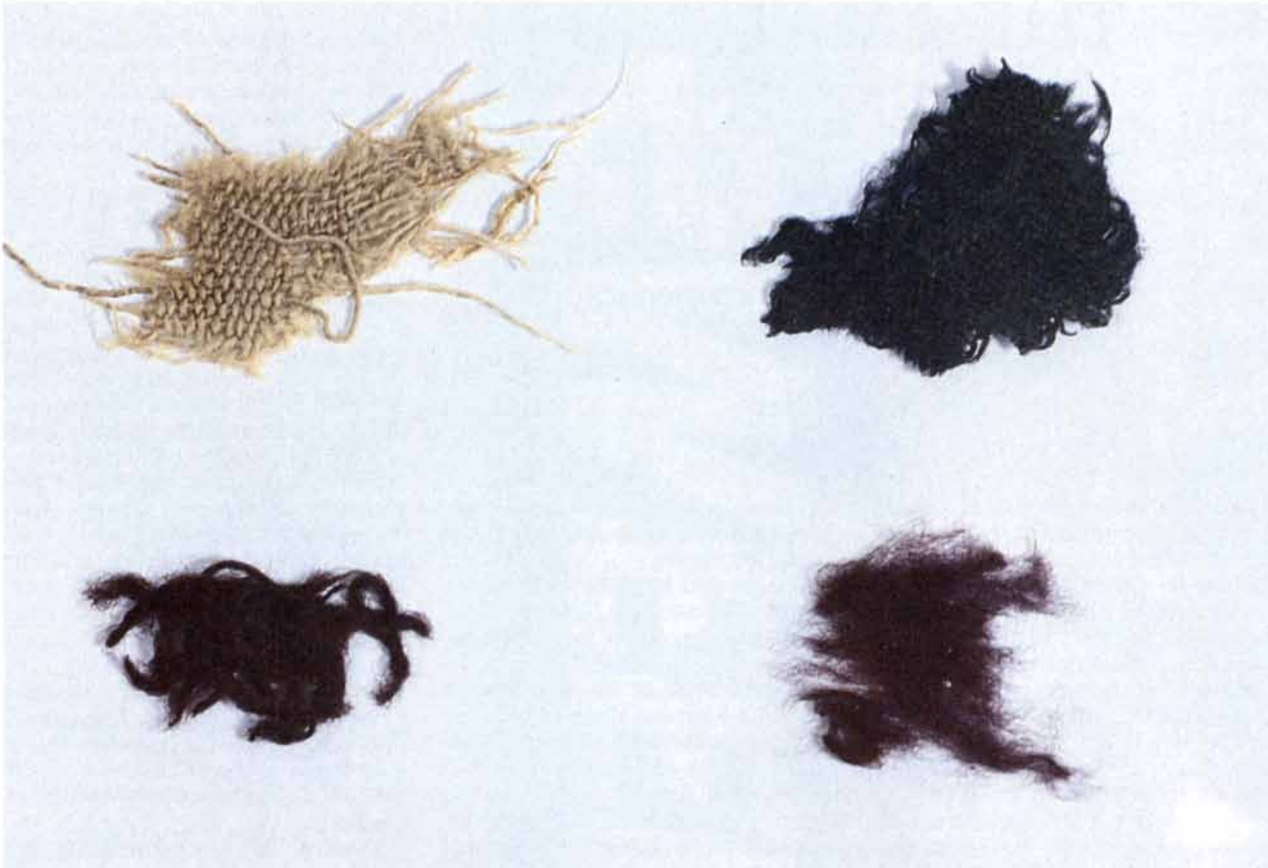
Middle East, the setting for most of the advances in wool and its associated technology, but fine-wooled sheep were soon disseminated throughout the ancient world, probably by seafaring Greeks. The earliest remains of true fine wool, dating to the fifth century B.C., come from the Greek colony of Nymphaeum in the Crimea. Images of sheep from ancient Greece and Rome also reflect the advent of relatively fine-wooled sheep: the representations show wool staples that are squared off in appearance instead of pointed like the staples in a hairy fleece. As fleeces become finer, the fiber lengths as well as their diameters become more uniform. As a result the staples that develop in a fine-wooled fleece are "blocky" rather than "tippy," to use the terms of the woolman.

The other two modern fleece types, the shortwool and the true medium, had appeared by Roman times. Vindolanda, a Roman site of the first century A.D. on Hadrian's Wall along the border between England and Scotland, has yielded evidence of the new fleeces. I analyzed fiber diameters in a collection of 57 yarns from the site. About half of the wool represented the generalized medium fleece and a third represented the older hairy medium type, but 2 percent was shortwool and 4 percent was true medium wool.

Breeds with the older fleece types predominated throughout the Roman world, however, as they did at Vindolanda. Wool remains from other Roman sites typically contain medium fibers, indicating that the generalized medium fleece remained important even as the modern fleece types were emerging. The Vindolanda specimens do contrast with other Roman specimens in their color: whereas wool from Mediterranean sites of that time is usually white, in the yarns I studied only 40 percent of the wool was white and 51 percent was gray, a proportion that is characteristic of unimproved Iron Age breeds.

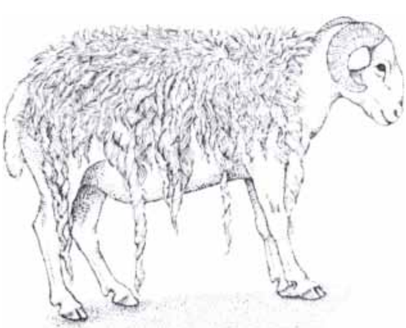
In later centuries Britain and other parts of northern Europe continued to harbor primitive breeds of sheep. Until after the Middle Ages, when shortwool and medium-wool breeds gained in importance, sheep that bore hairy medium and generalized medium fleeces remained widespread there. Around the Mediterranean, in contrast, the fine-wooled sheep became the mainstay of wool production. There the history of the fleece after classical times is dominated by the emergence of the definitive fine-wool breed, the Merino, now the world's foremost source of wool.

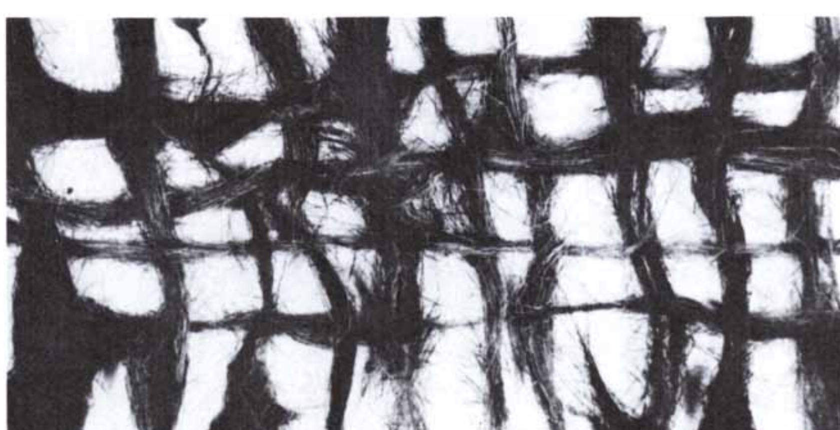
The Merino was developed in Spain



DYED WOOL in the form of unspun fibers (*lower right*), yarn (*lower left*) and woven cloth was found in the Cave of the Letters in Israel. The samples date from the second century A.D.; like most wool from the Roman period, they contain medium fibers as well as fine ones. The fleece that supplied the wool must have been the generalized medium type, ancient even then. The wool does show a feature that became common only in Roman times: it is naturally white. The advent of dyeing probably stimulated selective breeding for colorless wool. Ian R. Pitkethly of the Hill Farming Research Organisation in Edinburgh made the image.

during the Middle Ages. It became so highly valued that its export to other countries was forbidden, and the herds altered Spanish landscapes while the sheepkeepers collectively influenced Spain's political history. Spain lost its monopoly on the breed in the 18th century; in less than 100 years the Merino was introduced to other European countries and to the Southern Hemisphere, where its wool sustained the growing national economy of Australia. The story of the Merino epitomizes an aspect of the relations between sheep and human beings that a short account of fleece development necessarily neglects. Mankind transformed the sheep biologically through domestication and selective breeding, but the sheep in turn has powerfully shaped the course of human history.





FIRST ATTEMPT TO SPIN WOOL INTO YARN may have been suggested by the sight of a primitive sheep during the molt, when it sometimes rubs the shedding fibers into twisted strands (*left*). Stone Age peoples, adept at basketry, may have found such "yarns" could be woven into a crude cloth like the one made by the author using strands from a molting Mouflon sheep (*right*).

THE AMATEUR SCIENTIST

Reflections from a water surface display some curious properties

by Jearl Walker

Puddles and pools display several intriguing optical properties. Under what conditions can you see shadows on a body of water? Why are you sometimes able to see your own shadow but not that of a person standing nearby? What produces the bright streaks of light you sometimes see radiating from the shadow of your head on a pool? Why are shadows in a pool often fringed with colors or overlaid with images of the surroundings?

Water displaying moderate waves has a particularly curious property. Suppose you photograph the reflection of the mast of a sailboat. As you might expect, the mast in the photograph will look sinuous and kinked. Parts of it, however, also show up as isolated loops. What produces them? Are they always closed?

I begin with the problem of shadows. A shadow forms on a sunlit surface when an opaque object such as a book keeps light from reaching part of the surface. The rest of the surface scatters light in your direction so that you perceive it as being illuminated and the shadow as being dark. The shadow reproduces the shape of the book. The borders of the shadow are formed by the light rays skirting the edges of the book.

If the sun were a point source of light, so that its rays were precisely parallel, a shadow would be uniformly dark. If you look closely at the shadow of a book held above a sidewalk, however, you will see that the borders are slightly brighter than the interior. The region of partial illumination is the penumbra and the darker interior is the umbra. The penumbra results from the fact that the sun occupies about half a degree in the visible sky. The rays passing the edges of the book are spread through an angle of that size. Some of them illuminate the interior of the shadow along the borders, thereby forming the penumbra. Even the interior of a shadow is normally not completely dark because light from the surroundings is scattered into the shadow region.

Suppose you hold the book over a pool of still water. Do you see a shadow? If you do, where is it? The answers depend on several factors, including the depth of the water, its turbidity and what lies along the bottom. If the pool is shallow and clear and has a bottom such as concrete that does not absorb all the light or reflect it like a mirror, you can see a shadow on the bottom. Usually, however, you do not see it in its true position.

Again the cause lies with the rays that pass the edges of the book and define the borders of the shadow. When the rays travel through the air-water interface, they are refracted because the effective speed of light in water is less than it is in air. If the sun is directly overhead, the rays remain vertical; otherwise they intersect the interface at an angle to the vertical and are refracted so that they are more nearly vertical.

On reaching the bottom of the pool the rays are scattered in many directions. Some of the light moves upward and is refracted through the water-air interface. Unless the rays are vertical, the refraction increases their angle to the vertical. You extrapolate straight back along the rays and into the water without allowing for the refraction. The origins of the rays appear to be the points where the extrapolations intersect the bottom surface. Those points lie on the borders of the shadow because the intermediate region appears to be dark. Since you do not allow for refraction, the shadow you see is usually displaced from the true shadow.

The amount of displacement depends on your angle of view. If you look directly down on the true shadow, the rays scattering from its borders travel vertically through the air-water interface with no change in direction. Your extrapolation of them into the water goes to the true position of the shadow. With any other angle of view you miss the true shadow.

If the bottom completely absorbs the light, the shadow is invisible. Can you see it if the bottom reflects like an ideal mirror and the sun is the only source of light? Yes, provided you position yourself to intercept the reflected rays. From all other angles of view the illuminated regions of the bottom are as dark as the shadow.

A shadow at the bottom of a pool of water is not completely dark because the sun adds a penumbra to it and light from the surroundings is scattered out of it. In addition some direct sunlight is scattered from the bottom, reflected from the water-air interface and then scattered to you from within the shadow region. A still brighter light is scattered from the surroundings, including the sky, and then is reflected from the top of the water. Some of it reaches you along the route taken by the rays from the borders of the shadow. The composite scene is a shadow overlain by faint reflected images. The scene is puzzling because the reflected images may appear to be distant whereas the shadow seems to be on the bottom of the pool.

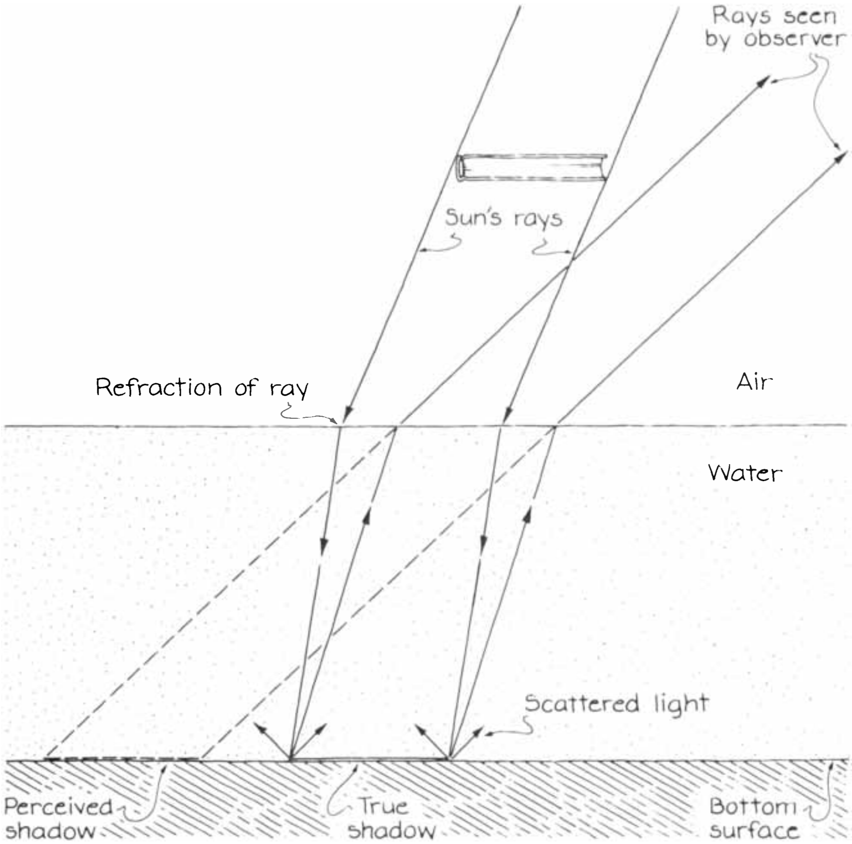
To check this observation I sat by a shallow puddle of water on a sidewalk. The sun was behind me and a building silhouetted against the sky was in front of me. Within my shadow on the puddle I saw reflections of the building and the sky. The apparent distance to the building equaled the true distance between the building and the puddle. I seemed to be looking through a hole at the bottom of the puddle. The image of the building's concrete walls was dim. The images of the windows were brighter because the windows scattered more light than the walls. The image of the sky was even brighter. I could see my reflection if I looked almost directly down on the water. Light from the surroundings was scattered from my face and then was reflected from the air-water interface. Faint colors were also visible.

Deep pools of clear water often display indistinct shadows. The penumbra is broader, decreasing the visibility of the shadow, and there is more opportunity for light to be scattered to you from within the shadow region.

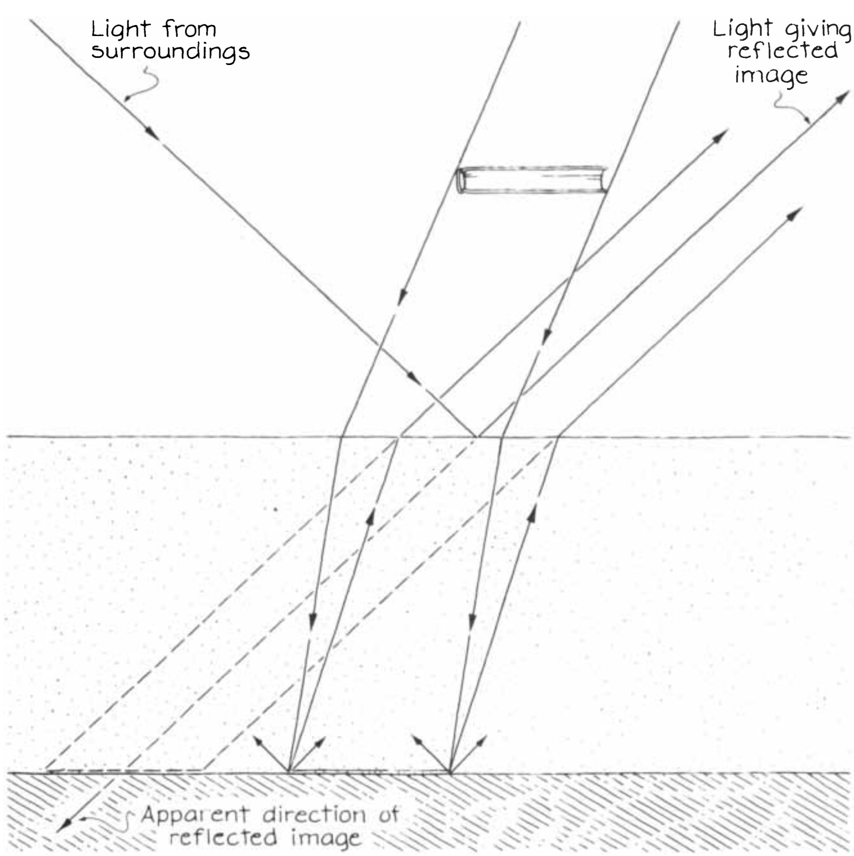
The appearance and the position of your shadow in a pool change markedly if the water is turbid. The light is scattered from the suspended particles. If the concentration of particles is moderate and the water is not too deep, light may still reach you by being scattered from the bottom. Your shadow is visible but its edges are muddled because the scattering does not take

Reflections of a sailboat's mast

Streaks of light that seem to radiate from a shadow



The origin of true and perceived shadows



How a reflected image overlies a shadow

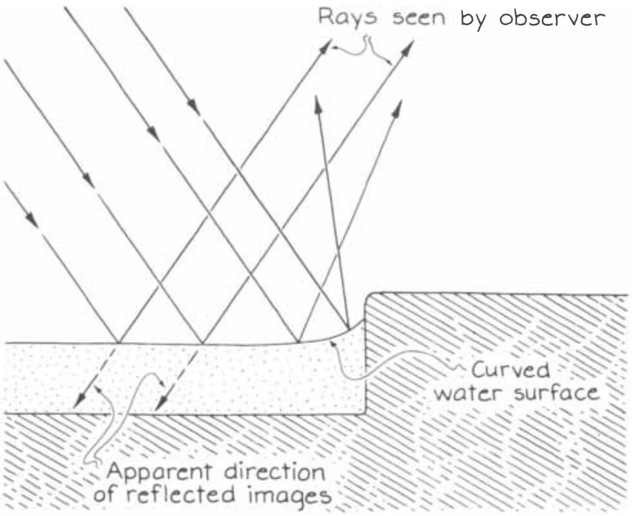
place in a single plane. The interior of the shadow is brightened by light being scattered from the shadow region. With more turbidity the amount of light reaching the bottom is insufficient for you to see a shadow there. Instead you see light that is scattered from the suspended material near the surface of the pool. Your shadow then seems to lie near the surface, probably again with muddled borders.

In *The Nature of Light & Color in the Open Air* Marcel Minnaert argues that in some conditions of turbidity you may not be able to see the shadow of someone standing farther along the edge of the pool. If the pool is so turbid that it is almost the consistency of mud, the light is scattered from the upper surface with little penetration. Your shadow and the shadows of other objects appear on the surface in their true positions. Because the scattering takes place in what is almost a single plane, the shadows have sharp borders.

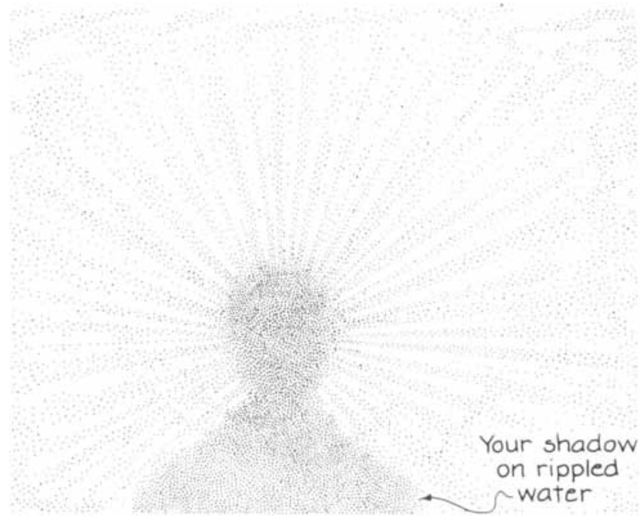
The surrounding sunlit mud is darker than dry dirt would be. Why is wet dirt darker than dry dirt? Craig F. Bohren of Pennsylvania State University recently explained the phenomenon. When the particles are surrounded by water instead of air, their scattering of light shifts forward, that is, downward into the bed. The reason is that water more closely matches the index of refraction of the particles than air does. With the shift in scattering direction the light is scattered many more times from wet particles than it is from particles surrounded by air. At each scattering point some of the light is absorbed. Hence the light that is scattered in your direction is dimmer than it would be in dry dirt.

Reflections in puddles are often distorted even when the water is still. Examine a puddle nested in a small "step" formed by two uneven concrete slabs. The flat section of the puddle reflects the surroundings as a mirror would, but the part next to the step yields distorted reflections. The surface there is curved because of the molecular attraction between the water and the concrete along the side of the step. Similar distortions are caused by the curved surfaces surrounding leaves and other objects that float in a puddle. If a pencil is inserted vertically into a puddle, the water is pulled up along its shaft. The sudden change in the reflected images creates the illusion that the pencil pulls images toward it. Actually the curved water surface reflects images to you from objects farther to one side of you than the flat water surface did.

In 1953 Stephen F. Jacobs of the University of Arizona noted a curious



An effect of a curved surface



The streak effect

pattern that can be seen when breezes blow over a pool of moderately turbid water at least a meter deep. Stand somewhat above the water and look at it near the shadow of your head. Flickering bright streaks seem to radiate from the shadow. They may create the illusion that the entire pattern rotates about your head.

The streaks are due to breeze-driven waves. In a calm, sunlight enters the unshadowed areas of water. The light is scattered from the suspended particles, and you see a uniformly bright pool. If breezes play over the pool, the variations in the shape of the surface alter the refraction of light into and out of the water. You no longer receive scattered light from the entire sunlit part of the pool. Jacobs believes the only regions that still send scattered light to you are those where the water is momentarily flat. I think some of the curved regions also contribute. In any case the water is bright in only a few places at any given instant.

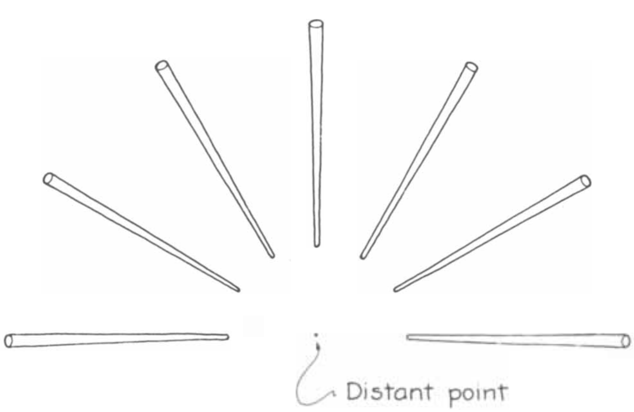
Why do the bright spots appear as radial streaks? Each streak is similar to a narrow shaft of light penetrating a dark, dusty room. What you see is light scattered from the dust particles, but what you perceive is a beam crossing the room.

The radial orientation of the streaks in water is an illusion. Suppose you were to look along the length of an array of long, parallel rods [*see illustration at left below*]. The scene offers at least two cues about depth. The outer ends of the rods seem to be near because their faces are visible. The tapering of the rods also reveals depth. Without the cues you might see the figure as a flat array of rods aligned along radial lines from the center of your perspective. You make a similar interpretation of the streaks in the water. Lacking cues about depth, you conclude that they are on the surface and that they point toward the position of the eyes in the shadow of your head.

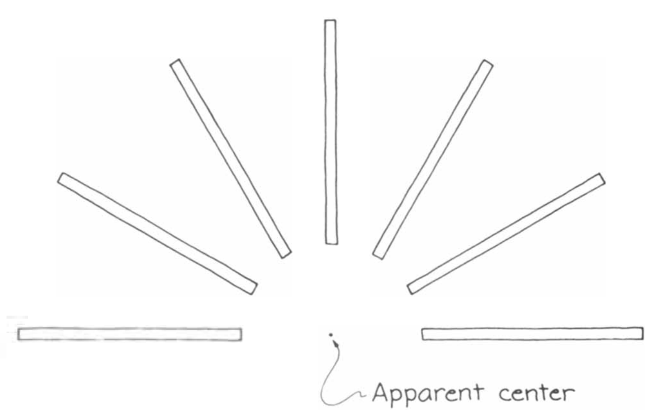
A coloring that can appear in clear water was discovered by Frank S. Crawford, Jr., of the University of California at Berkeley while he was relaxing in a hot tub. The water was dappled with sunbeams streaming through the leaves of a tree. A companion noticed that the spots of light formed by the sunbeams on the bottom of the tub were either completely white or had colored edges, depending on how they were viewed. If you sit with the sunbeams passing over a shoulder, they are white. If you sit facing the sun, the near edge of a spot is red and the far edge is blue.

The colors are due to the spread of the color components in the initial white sunlight as it is refracted into the water. The blue changes its direction of travel at the interface more than the red does; the intermediate colors green and yellow change directions by intermediate amounts. Because of the spread, the red light and the blue light are scattered from the bottom at different points.

If you sit with your back to the sun, the only rays that return to you from the bottom are those that travel almost in reverse along the initial path of the light. A ray of each color travels back to you in this way, passing through the



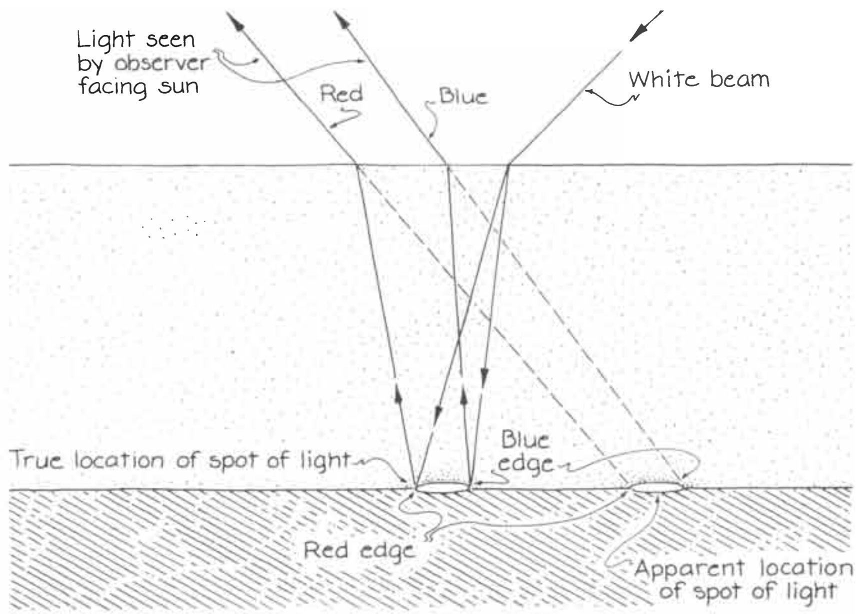
The analogy of long, parallel rods



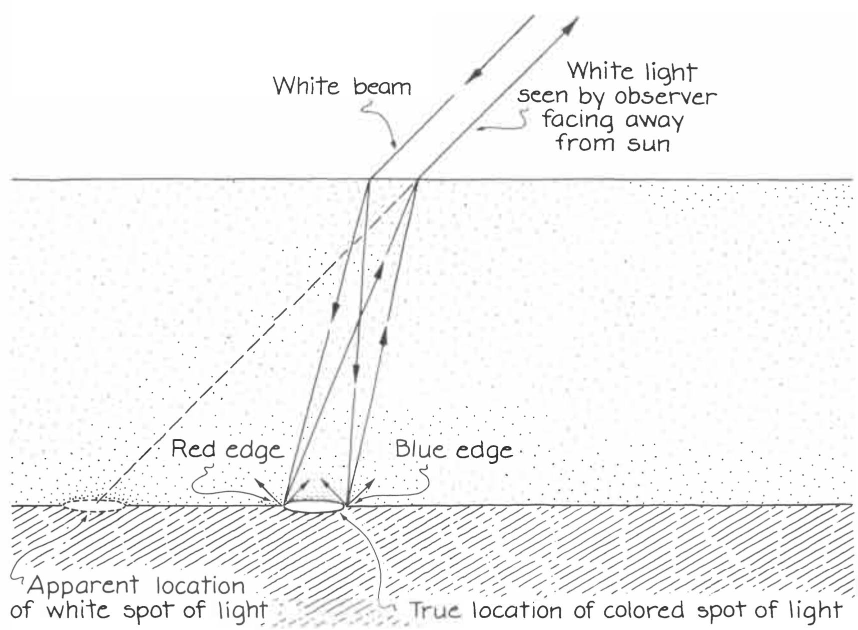
An apparently flat array of untapered rods same point on the water-air interface. You perceive the combination of colors as being white. The source of the light appears to be a single spot on the bottom. The spot is displaced from the true scattering points for the colors because you do not allow for the refraction of light at the water-air interface.

If you sit facing the sun, the rays of various colors are scattered from the bottom and pass through the interface at different points. You perceive them as originating at different but overlapping spots on the bottom. The center of the overlap is white, but the far edge is blue and the near edge is red.

Recently Thomas Gold of Cornell University rediscovered a puzzling property of a body of water covered with waves. Photographing the reflection of a ship's mast, he found the expected sinuous image. He also saw, off to one side of the image, an isolated loop that was an image of a short section of the mast. The image within the loop was the sky on the far side of the mast, and the image outside the loop was the sky on the near side of the mast. What happened was that while Gold was making the exposure a small region of the water surface tilted and curved in such a way as to reflect the



The result of facing the sun



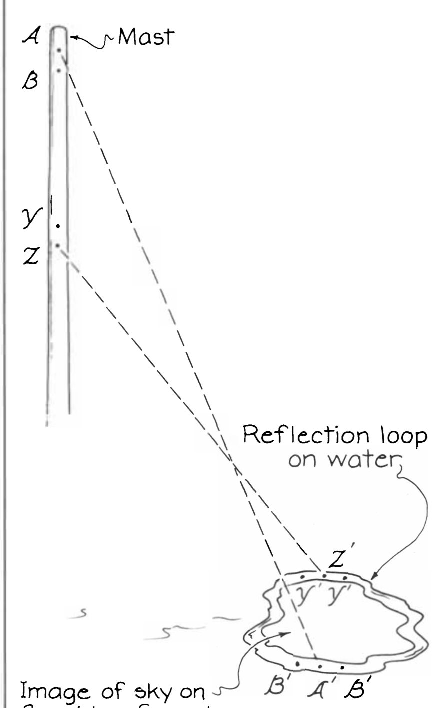
The result of facing away from the sun


BREAKTHROUGH:
QUICK COOL ALLOYS
SO THEY CAN TAKE THE HEAT.
Build a plane that can fly from the U.S. to the Orient between lunch and tea? It can be done. But to build it, new lightweight materials are needed that can survive the heat of hypersonic speeds.
Our solution: A metallurgical recipe using molten titanium alloy and a special process in which the hot liquid metal is solidified into powder by extremely rapid cooling. After the powder is processed into a solid and worked into sheets, designers have a practical material for helping get hypersonic flight off the ground.
We're creating breakthroughs that make a difference in the way things work and the way people live.
We're McDonnell Douglas.
For more information, write: Orient Express, McDonnell Douglas, Box 14526, St. Louis, MO 63178.
MCDONNELL DOUGLAS
SPACECRAFT & MISSILES
MILITARY & COMMERCIAL AIRCRAFT
HEALTHCARE SYSTEMS
INFORMATION SYSTEMS
HELICOPTERS
FINANCING
© 1987 McDonnell Douglas Corporation

section of mast. Provided the water surface has no sharp waves and no area of it is hidden from view by waves, such isolated reflections of a mast must form complete loops.

To understand the loop's formation, label the points on the section of mast and their corresponding points in the loop. Call the highest point *A* and its reflection *A'*. The next-lower point on the section, *B*, is reflected twice in the loop, once on each side of *A'*. Call the lowest point on the section *Z* and its reflection *Z'*. The next-higher point on the section, *Y*, has two reflections, one on each side of *Z'*. All the points between *A* and *Z* have two reflections in the loop. Thus the section is imaged as a closed loop.

David K. Lynch of the Aerospace Corporation Space Science Laboratory in Los Angeles elaborated on Gold's observations by noting that a pool of water covered with waves can also display "land pools" and "sky pools." These images are small, isolated patches that reflect extended sources of light such as mountains. A point source of light can also create an extended image. If it is photographed with a slow shutter speed, the image is sinuous and probably overlapping because the waves move during the long exposure. If the shutter speed is fast, the image is a short line, with ends corresponding to the opening and closing of the shutter. You might like to study further the different types of reflected images that can be photographed in these circumstances.



A loop in a reflection

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE YIELDS OF SOVIET STRATEGIC WEAPONS

SOVIET STRATEGIC FORCES: REQUIREMENTS AND RESPONSES. Robert P. Berman and John C. Baker. The Brookings Institution, 1982.

THE VERIFICATION OF A COMPREHENSIVE NUCLEAR TEST BAN. Lynn R. Sykes and Jack F. Evernden in *Scientific American*, Vol. 247, No. 4, pages 47–55; October, 1982.

NUCLEAR WEAPONS DATABOOK: U.S. NUCLEAR FORCES AND CAPABILITIES. Thomas B. Cochran, William M. Arkin and Milton M. Hoenig. Ballinger Publishing Company, 1984.

ARMS CONTROL VERIFICATION: THE TECHNOLOGIES THAT MAKE IT POSSIBLE. Edited by Kosta Tsipis, David W. Hafemeister and Penny Janeway. Pergamon-Brassey's, 1986.

WORLD ARMAMENTS AND DISARMAMENT: SIPRI YEARBOOK 1986. Stockholm International Peace Research Institute. Taylor & Francis, Ltd., 1986.

URANUS

THE NEW SOLAR SYSTEM. Edited by J. Kelly Beatty, Brian O'Leary and Andrew Chaikin. Cambridge University Press & Sky Publishing Corporation, 1982.

URANUS AND THE OUTER PLANETS: PROCEEDINGS OF THE IAU-RAS COLLOQUIUM. Edited by Garry Hunt. Cambridge University Press, 1982.

VOYAGER 2 MISSION TO URANUS. In *Science*, Vol. 233, No. 4759; July 4, 1986.

THE AIDS VIRUS

ISOLATION OF A T-LYMPHOTROPIC RETROVIRUS FROM A PATIENT AT RISK FOR ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS). F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum and L. Montagnier in *Science*, Vol. 220, No. 4599, pages 868–871; May 20, 1983.

DETECTION, ISOLATION, AND CONTINUOUS PRODUCTION OF CYTOPATHIC RETROVIRUSES (HTLV-III) FROM PATIENTS WITH AIDS AND PRE-AIDS. Mikulas Popovic, M. G. Sarngadharan, Elizabeth Read and Robert C. Gallo in *Science*, Vol. 224, No. 4648, pages 497–500; May 5, 1984.

FREQUENT DETECTION AND ISOLATION OF CYTOPATHIC RETROVIRUSES (HTLV-III) FROM PATIENTS WITH AIDS AND AT RISK FOR AIDS. Jörg Schüpbach, Mikulas Popovic, Raymond V. Gilden, Matthew A. Gonda, M. G. Sarngadharan and Robert C. Gallo in *Science*, Vol. 224, No. 4648, pages 500–505; May 5, 1984.

THE EPIDEMIOLOGY OF AIDS: CURRENT STATUS AND FUTURE PROSPECTS. James W. Curran, W. Meade Morgan, Ann M. Hardy, Harold W. Jaffe and William R. Dowdle in *Science*, Vol. 229, No. 4720, pages 1352–1357; September 27, 1985.

HUMAN T-LYMPHOTROPIC RETROVIRUSES. Flossie Wong-Staal and Robert C. Gallo in *Nature*, Vol. 317, No. 6036, pages 395–403; October 3, 1985.

NEW HUMAN T-LYMPHOTROPIC RETROVIRUS RELATED TO SIMIAN T-LYMPHOTROPIC VIRUS TYPE III ($STLV\text{-}III_{AGM}$). Phyllis J. Kanki, Francis Barin, Souleyman M'Boup, Jonathan S. Allan, Jean Loup Romet-Lemonne, Richard Marlink, Mary Frances McLane, Tun-Hou Lee, Brigitte Arabelle, François Denis and M. Essex in *Science*, Vol. 232, No. 4747, pages 238–243; April 11, 1986.

A SECOND POST-TRANSCRIPTIONAL *TRANS*-ACTIVATOR GENE REQUIRED FOR HTLV-III REPLICATION. Joseph Sodroski, Wei Chun Goh, Craig Rosen, Andrew Dayton, Ernest Terwilliger and William Haseltine in *Nature*, Vol. 321, No. 6068, pages 412–417; May 22, 1986.

LEARNING BY INSTINCT

ETHOLOGY: THE MECHANISMS AND EVOLUTION OF BEHAVIOR. James L. Gould. W. W. Norton & Co., 1982.

THE BIOLOGY OF LEARNING. Edited by Peter Marler and Herb S. Terrace. Springer-Verlag, 1984.

POWER FROM THE SEA

POWER FROM THE TROPICAL SEAS. Georges Claude in *Mechanical Engineering*, Vol. 52, No. 12, pages 1039–1044; December, 1930.

THE MANAGEMENT OF PACIFIC MARINE RESOURCES: PRESENT PROBLEMS AND FUTURE TRENDS. John P. Craven. Westview Press, 1982.

ENERGY FROM THE OCEAN. R. Cohen in *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 307, No. 1499, pages 405–437; October 20, 1982.

POTENTIAL ENVIRONMENTAL IMPACTS OF OPEN-CYCLE THERMAL ENERGY CONVERSION. M. S. Quinby-Hunt, P. Wilde and A. T. Dengler in *Environmental Impact Assessment Review*, Vol. 6, pages 77–93; 1986.

FRACTAL GROWTH

DIFFUSION-LIMITED AGGREGATION, A KINETIC CRITICAL PHENOMENON. T. A. Witten, Jr., and L. M. Sander in *Physical Review Letters*, Vol. 47, No. 19, pages 1400–1403; November 9, 1981.

THE FRACTAL GEOMETRY OF NATURE. Benoit B. Mandelbrot. W. H. Freeman and Company, 1982.

DIFFUSION-LIMITED AGGREGATION. T. A. Witten and L. M. Sander in *Physical Review B*, Vol. 27, No. 9, pages 5686–5697; May 1, 1983.

TENUOUS STRUCTURES FROM DISORDERLY GROWTH PROCESSES. T. A. Witten and M. E. Cates in *Science*, Vol. 232, No. 4758, pages 1607–1612; June 27, 1986.

LIPOSOMES

LIPOSOMES AND IMMUNOBIOLOGY. Edited by Baldwin H. Tom and Howard R. Six. Elsevier–North Holland, 1980.

LIPOSOMES. Edited by Marc J. Ostro. Marcel Dekker, Inc., 1983.

LIPOSOME TECHNOLOGY. Edited by Gregory Gregoriadis. CRC Press, 1984.

THE EVOLUTION OF THE FLEECE

SHEEP AND MAN. Michael L. Ryder. Duckworth, 1983.

A RE-ASSESSMENT OF THE BRONZE AGE WOOL. M. L. Ryder in *Journal of Archaeological Science*, Vol. 10, No. 4, pages 327–331; July, 1983.

SHEEP REPRESENTATIONS, WRITTEN RECORDS AND WOOL MEASUREMENTS. M. L. Ryder in *Animals and Archaeology, 3: Early Herders and Their Flocks (B.A.R. International Series 202)*, edited by Juliet Clutton-Brock and Caroline Grigson. B.A.R. [British Archaeological Reports], 1984.

THE AMATEUR SCIENTIST

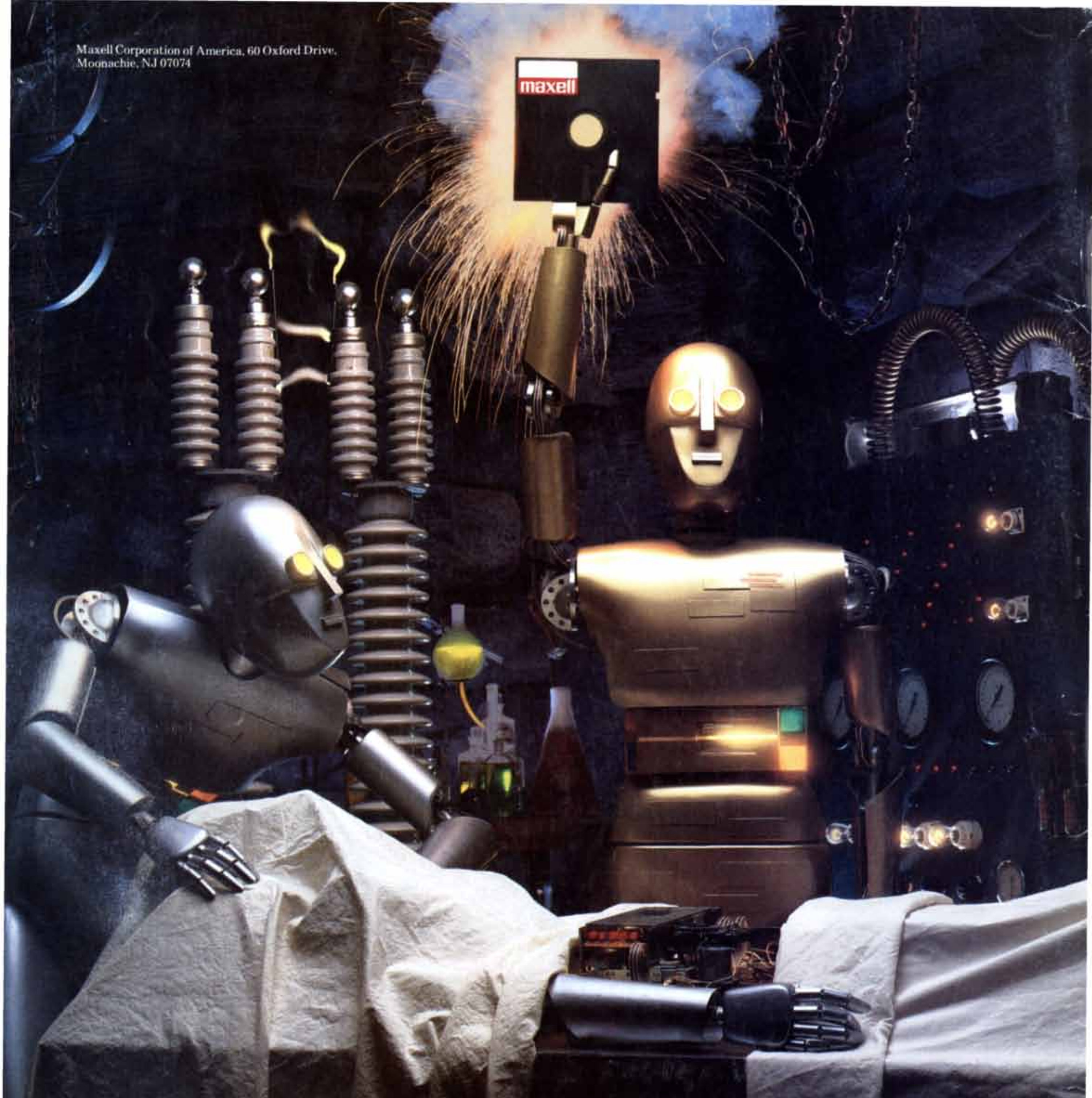
THE NATURE OF LIGHT & COLOR IN THE OPEN AIR. M. Minnaert. Dover Publications, Inc., 1954.

REFLECTIONS ON CLOSED LOOPS. David K. Lynch in *Nature*, Vol. 316, No. 6025, pages 216–217; July 18, 1985.

How we pushed America's premier road sedan to the ultimate braking point. *When driving conditions are less than ideal, Pontiac 6000 STE helps keep you out of a bind with advanced anti-lock braking. This new system delivers improved braking capability on almost any surface, wet or dry. Together with superb handling and fuel-injected V6 power, it gives you a machine with some of the finest performance credentials going. Or stopping.*

PONTIAC 6000

WE BUILD EXCITEMENT



Maxell Corporation of America, 60 Oxford Drive, Moonachie, NJ 07074
maxell

maxell
MINI-FLOPPY DISK
MD2-D
FREE
AI SOFTWARE
$200