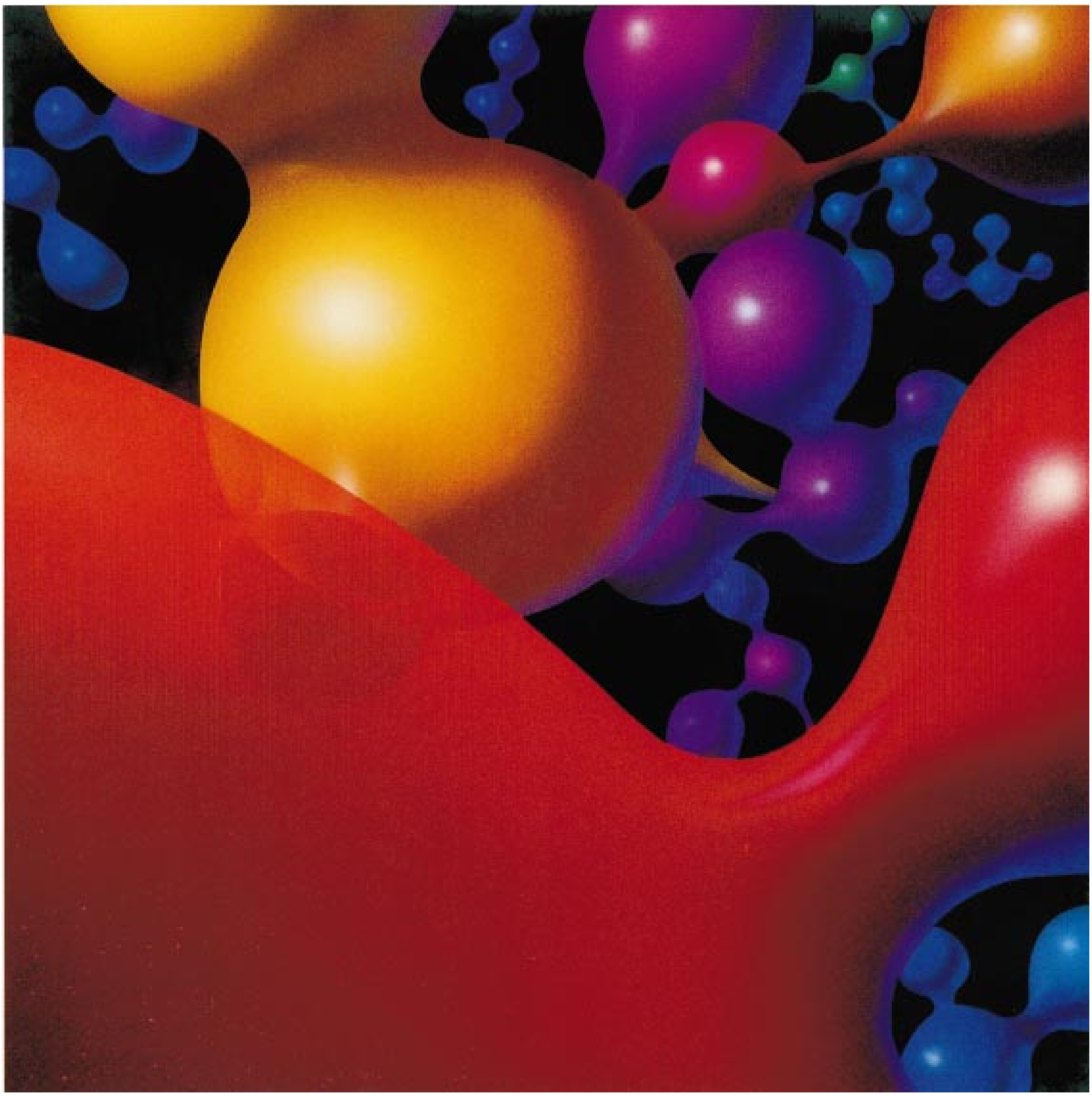
# SCIENTIFIC AMERICAN

NOVEMBER 1994
$3.95

*Safeguarding computer networks.*

*M. C. Escher's visual mathematics.*

*Solving the mystery of meningitis.*



***New universes*** *constantly burst from the old in an inflationary cosmos.*

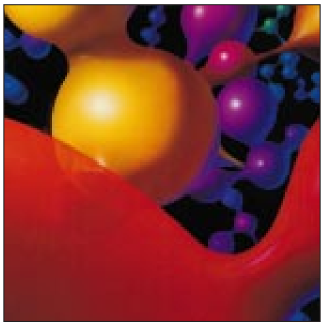
SCIENTIFIC AMERICAN

November 1994 Volume 271 Number 5

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111.  Second-class postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537. Reprints available: write Reprint Department, Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111; fax: (212) 355-0408 or send E-mail to SCAinquiry@aol.com.



# DEPARTMENTS



THE COVER painting depicts an unusual view of the cosmos: a bubbling, branchlike sea of universes. Each bubble represents a universe, which sprouts other universes, ad infinitum. The laws of physics of a particular universe, represented by a color, are not fixed either: a birth may produce a "mutation." Each universe will eventually die, but as a whole, the cosmos is eternal. While perhaps far-fetched, this vision falls naturally from the latest ideas of inflationary cosmology [see "The Self-Reproducing Inflationary Universe," by Andrei Linde, page 48].

## THE ILLUSTRATIONS

Cover painting by Alfred T. Kamajian

| Page | Source |
|---|---|
| 38–39 | Alan Reininger/ Contact Press Images |
| 40 | Dimitry Schidlovsky |
| 41 | Dana Burns-Pizer |
| 42 | Dimitry Schidlovsky |
| 43 | Johnny Johnson |
| 44 | Dimitry Schidlovsky |
| 45 | Patrick S. Moore |
| 49–53 | Visual Arts Service, Stanford University |
| 54 | Ian Worpole |
| 55 | Jared Schneidman/JSD |
| 56–57 | Elliot M. Meyerowitz |
| 60 | Patricia J. Wynne |
| 61 | *Photographs, top to bottom:* G. N. Drews, courtesy of *Cell;* J. L. Bowman; E. M. Meyerowitz; J. L. Bowman; George Retseck (*diagrams*) |
| 62 | J. L. Bowman (*top*), G. N. Drews, courtesy of *Cell* (*bottom*) |
| 63 | J. L. Bowman, courtesy of *Development* (*top*), T. Jack (*bottom*) |
| 64 | J. L. Bowman (*top three*), D. Weigel (*bottom*) |
| 65 | Bruce Hands/ Tony Stone Images |
| 66–71 | M. C. Escher/Cordon Art-Baarn, Holland |
| 72–73 | Mark C. Flannery (*photograph*) |
| 74–75 | Jana Brenning |
| 76 | Sam Ogden |
| 79 | R. Jonathan Rehg |
| 81 | Courtesy of Laura E. Berk |
| 82 | R. Jonathan Rehg |
| 83 | Lillian Holan, reproduced with permission of HarperCollins Publishers, Inc. |
| 85 | Harold E. Edgerton 1992 Trust/Palm Press, Inc. |
| 86 | Patricia J. Wynne |
| 87–88 | Johnny Johnson |
| 89 | University of Oxford Museum of the History of Science |
| 90–91 | Andy Myer, courtesy of *Journal of NIH Research* |
| 92 | Johnny Johnson |
| 92A | Schering-Plough Corporation |
| 92B | Sam Kittner (*left*), Andy Freeberg (*center*), James Holb (*right*) |
| 93 | Andy Freeberg (*left*), Terry Ashe/Gamma Liaison Network (*center*), Chip Anderson (*right*) |
| 94 | Courtesy of University of Washington |
| 96 | Andy Myer, courtesy of *Journal of NIH Research* |
| 108–111 | Johnny Johnson |


**SCIENTIFIC AMERICAN**®

*Established 1845*





**SCIENTIFIC AMERICAN, INC.**

415 Madison Avenue, New York, NY 10017-1111
(212) 754-0550





**PRINTED IN U.S.A.**

# LETTERS TO THE EDITORS

## Quantum Realities

In "Bohm's Alternative to Quantum Mechanics" [Scientific American, May], David Z Albert has given a lucid and accurate account of the subject. One point may lead to some confusion, however. Albert states that the wave-function force field is like a classical force field, which is not true. For example, the strength of the force to which the quantum field gives rise is independent of the amplitude. Thus, it is possible for a wave function of very small amplitude to produce a large effect. One consequence is that the force does not necessarily fall off with distance, which accounts for the nonlocal Einstein-Podolsky-Rosen correlations.

The article also states that Bohm died while in the middle of writing another quantum mechanics book. As his collaborator, I am happy to state that we were actually just putting the finishing touches to that work. The book, *The Undivided Universe: An Ontological Interpretation of Quantum Mechanics,* was published by Routledge, Chapman & Hall in November 1993.

Basil J. Hiley
Department of Physics
Birkbeck College
University of London

Allow me to make one criticism of a generally excellent article. Contrary to Albert's assertion, there is a solution to the quantum-mechanical measurement problem—called "consistent histories"—that yields the same predictions as standard quantum mechanics, does not suffer from nonlocality and is *not* a "many worlds/many minds" interpretation. I described it originally in the *Journal of Statistical Physics* in 1984. From the consistent-histories perspective, the appearance of peculiar nonlocalities in Bohm's approach is a result of adding unneeded classical variables to the description of Hilbert's space. When the Hilbert space approach is used consistently, these nonlocal effects are absent.

Robert B. Griffiths
Department of Physics
Carnegie Mellon University

***Albert replies:***
Unfortunately, in the available space, I cannot deal adequately with the issue Griffiths raises. Perhaps it is worth merely setting down that I think he is mistaken. The details of his very interesting consistent-histories explanation are largely irrelevant in this instance. The discovery by the late John S. Bell of CERN is proof that no local theory can possibly reproduce the statistical predictions of quantum mechanics while still satisfying certain very weak, natural conditions. The first question to ask anybody who claims to have come up with a local version of quantum mechanics is, "Which of those conditions does your version fail to satisfy?" Griffiths's theory, unless I somehow misunderstand it, satisfies them all.

## Mixed Grades

The subheadline for "Grading the Gene Tests," by John Rennie [Scientific American, June], claims that "ethical problems surrounding [genetic] testing are as ominous as the diseases themselves." This equates potential discrimination by insurers or employers with the slow, painful death of Tay-Sachs.

Fortunately, reality routinely trumps such vaporings. Parents hope for children who are healthy and smart. They always have. The emerging ability of genetic testing and selection to fulfill this hope does not render it malign.

Prospective parents take their chances on the genetic lottery only because they have no choice. Within 10 to 20 years, couples will test embryos before implantation with as few qualms as they now vaccinate their children. Articles that agonize over whether we should or shouldn't serve no purpose when we so obviously will.

C. Owen Paepke
Phoenix, Ariz.

Your recent article on genetic screening should serve as a call to arms. Unless society responds swiftly to the current and future misuse of genetic information by insurers, employers, governments and other institutions, millions of healthy Americans will be deprived of access to insurance, credit (especially mortgages) and career-track employment. They will become members of a new genetic underclass.

The article seriously understated the commercial pressures for widespread multidisease screening. Many universities and leading researchers have intimate financial connections, including equity ownership in biotech companies. Johns Hopkins University, for example, has a stake in Oncor, a company engaged in questionable programs screening for fragile X syndrome.

Progress in genetics will eventually lead to wonderful treatments; in the meantime, it threatens everyone's civil rights. Our response must not be to slow the science but to strengthen our political commitment to true equality.

Hans S. Goerl
Director
Genethics Center
Hagerstown, Md.

## Why Race in Space?

As I read "Was the Race to the Moon Real?" by John M. Logsdon and Alain Dupas [Scientific American, June], about how the Soviet bureaucracy stymied the building of a propulsion system, I felt as though I were reading a description of the current U.S. efforts to build a space station. Plans laid aside for new redesigns, concerns about budgets, political infighting and a complete lack of vision—it sounds like NASA these days. America's space effort needs to focus and intensify. Why does it need a race to accomplish this?

James Gruber
Bound Brook, N.J.

The photograph on page 43 entitled "Neil Armstrong on the moon (July 20, 1969)" is really of Armstrong's comrade Buzz Aldrin. Because Armstrong carried the camera attached to his suit, most of the lunar surface pictures from *Apollo 11* featured Aldrin. Armstrong did appear in some photographs as a reflection on Aldrin's mirrored helmet.

David M. Sawyer
Winston-Salem, N.C.

*Letters selected for publication may be edited for length and clarity. Unsolicited manuscripts and correspondence will not be returned or acknowledged unless accompanied by a stamped, self-addressed envelope.*

# 50 AND 100 YEARS AGO

NOVEMBER 1944

“The chemical industry is becoming more and more conscious that profits are to be found in greater bulk when basic chemicals are turned into new synthetic consumer products or into materials from which these products can be fabricated, than when the same chemicals are sold in carload lots to processors who reap the harvest. This new-type thinking on the part of the large chemical producers—Dow, Monsanto, Du Pont, Union Carbide, and others—is setting a trend in the largest basic industry of the United States.”

“Up to about 10 years ago electronics had not been accepted in large plants such as steel mills, foundries, machine shops, and mines, to any extent at all. The head of a steel mill might point out the rough-and-ready workmen hoisting things about the plant and ask with a laugh: ‘What chance would a glass tube have in such an environment?’ Today, however, electronic tubes are mounted right on huge punch presses and rolling mills, doing the job so satisfactorily that shutting down the electronic controls would create a minor catastrophe among the men.”

“The bactericidal effect of sunlight has been duplicated in an ultra-violet tube called the Sterilamp. The radiations of this lamp speedily kill surface and air-borne bacteria, viruses and mold spores. Hundreds of thousands of these lamps now are on guard in a wide variety of fields.”

“Chest X-rays of industrial workers in factories controlled by the government indicate that about one person in every 1,000 has unsuspected tuberculosis.”

SCIENTIFIC AMERICAN

NOVEMBER 1894

“The nineteenth century can be no better defined as a century of wonders. But the great increase in mechanical appliances and the growth of population in cities has brought about a disagreeable effect, the increase of noise. From the private office, where the rattle of the typewriter has proved the successor to the classic squeaking of the quill pen, to the street, where the traffic of carriages and carts is overtopped by the roar of the elevated railroad, our life is spent in the midst of noise.”

“On September 8, 1894, after 73 years of life, which yielded a record almost unsurpassed of work in physiology, anatomy and physics, Herman Ludwig Ferdinand von Helmholtz died.”

“The latest of Mr. Edison’s inventions is the kinetoscope. The London Times says: This instrument is to the eye what Edison’s phonograph is to the ear, in that it reproduces living movements of the most complex and rapid character. Mr. Edison promises to add the phonograph to the kinetoscope and to reproduce plays. Then by amplifying the phonograph and throwing the pictures on a screen, making them life size, he will give the world a startling reproduction of human life.”

“Mrs. Peary, the only lady to take part in any Arctic expedition, spent a year in Greenland. She has recently published her journal. We quote: ‘The native method of treating the skins of all animals intended for clothing is first to rid them of as much of the fat as can be got off by scraping with a knife; then they are stretched as tight as possible, and allowed to become perfectly dry. After this, they are taken by the women and chewed and sucked all over in order to get as much of the grease out as possible; then they are again dried and scraped with a dull implement so as to break the fibers, making the skins pliable. Chewing the skins is very hard on the women; they are obliged to rest their jaws every other day.’ ”

“The whole world owes a deep debt of gratitude to the young French savant, Dr. Roux, for the discovery of an effectual cure for diphtheria. Diphtheria is produced by microbes which plant themselves in the membrane of the throat, and multiply. There, they secrete a poison of extreme violence, called ‘toxin,’ which quickly penetrates the circulation and infects the whole body. Dr. Roux’s ‘serum therapy’ is produced by first injecting isolated toxin into a horse. The second step is to draw from the animal a judicious quantity of blood. If the blood be allowed to stand for a while, the red corpuscles settle to the bottom, and the operator can draw off the fluid containing the serum, or antitoxin. This, in turn, is injected under the skin of a patient [*see illustration below*]. The distinguished Dr. Marsan says there are toxins and antitoxins for all microbic affections. Serum therapy will eventually discover a remedy for all infectious diseases.”

***The new cure for diphtheria—injecting the serum***

## Talk about the Weather

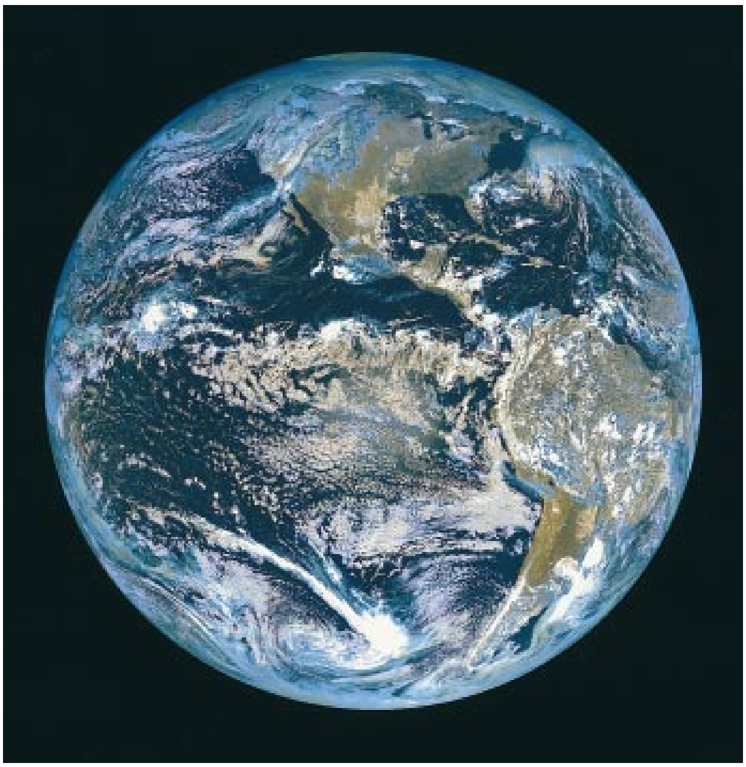
*Insights help to explain solar effects on climate*

One of the brightest gems in the New England weather is the dazzling uncertainty of it," Mark Twain once quipped. That uncertainty is not quite so amusing to scientists attempting to understand and forecast long-term weather patterns in the U.S. and around the globe. Their task is further complicated by a surprisingly poor knowledge of how changes in the sun affect conditions on the earth. After centuries of searching, climatologists are finally finding apparently indisputable links between the 11-year cycle of solar activity and shifts observed in terrestrial weather. But why those links exist is still hotly debated.

Given that the sun provides the energy that drives all weather systems, it seems obvious that solar variations should have environmental consequences. Until recently, however, attempts to find such correlations were often viewed with the kind of skepticism that is usually reserved for ESP and flying saucers. "A lot of the meteorological community thought this wasn't a respectable field," laments Brian A. Tinsley of the University of Texas at Dallas. "Papers have been published that suffered from weak statistics and improbable theories. We've had to work hard to make it respectable."

The turning point came during the late 1980s, when Karin Labitzke of the Free University in Berlin and Harry Van Loon of the National Center for Atmospheric Research in Boulder, Colo., presented convincing evidence that winter storms trace out a distinctive 11-year pattern of low-pressure systems over the North Atlantic Ocean. The pattern matched both the period and phase of the solar cycle, during which the level of solar activity (such as sunspots and flares) rises and falls. Unlike many previously reported sun-weather correlations, this one shows no sign of going away. "The association looks very nice and has continued through all subsequent winters," Labitzke states.

Building on that finding, Labitzke and Van Loon reported this year a more general 10- to 12-year atmospheric oscillation. The two workers deduce that the troposphere, the dense bottom layer of the earth's atmosphere, grows hotter and cooler in step with the solar cycle in regions near the tropics. Labitzke suspects that changes in solar radiation affect conveyor-belt atmospheric motions known as Hadley circulation. These movements carry warm air up over the equator, away toward the poles, down to the surface at subtropical latitudes and back to the equator. During periods of high solar activity, Labitzke speculates that Hadley circulation intensifies, transporting more hot air to the subtropics and accounting for the observed temperature increases.



GEOPIC EARTH SATELLITE CORPORATION

***WEATHER PATTERNS, including the paths of cyclones (such as the one at the bottom in this enhanced-color image), seem to shift in response to tiny changes in the sun. Researchers are struggling to learn why.***

Impressive though Labitzke and Van Loon's statistics are, they do not explain how the sun-earth connection takes place. Measurements from the *Nimbus-7* satellite show that the total luminosity of the sun changed by only about 0.1 percent during the past cycle. How could such a tiny fluctuation in the sun's total output significantly influence the weather? "I don't know how the sun does it," Labitzke confesses genially.

The search for a process that would explain the Labitzke–Van Loon findings has produced two hypotheses built around two very different ways of looking at the earth's atmosphere. Labitzke favors the more conventional of these views: weather shifts respond to variations in the intensity of solar ultraviolet radiation, which are more pronounced than are the changes in visible light. Ultraviolet rays are absorbed by stratospheric ozone and so help to determine the temperature of that layer of the atmosphere. Ultraviolet radiation also creates additional ozone in the stratosphere, which may lead to a complex feedback process. Changes in stratospheric temperature could alter Hadley circulation or other aspects of atmospheric mixing that influence weather.

Several researchers, including David Rind of the Goddard Institute for Space Studies, are examining the plausibility of the ultraviolet hypothesis using elaborate computer models. Rind points out that during times when the sun is relatively active, the elevated intensity of ultraviolet radiation heats up the stratosphere. A hotter stratosphere, he argues, changes the manner in which giant atmospheric waves—those that are 10,000 kilometers or more in length—are generated and propagate between the stratosphere and the troposphere.

Such changes could affect cloud cover, winds and temperatures at the surface, perhaps by as much as five degrees Celsius locally. Moreover, Rind believes these effects could accumulate from one solar cycle to the next. Small variations in solar activity could thus bring about long-lived climate changes such as the Little Ice Age—a period of abnormally cold weather that persisted in Europe from the 15th to 18th centuries. "This sort of explanation is fairly subtle," Rind concedes. But he thinks it provides the most plausible way to amplify solar twitches into shudders in the earth's climate.

Tinsley disagrees. For years he has championed the intriguing but unorthodox alternative hypothesis that charged particles, not ultraviolet light, constitute the primary mechanism by which solar variability stirs up weather. Tinsley notes that the solar wind—a stream of charged particles that continuously blows outward from the sun, past the earth—affects the electric currents that flow in the atmosphere. A slight buildup of electrical charge could promote the formation of ice crystals, effectively "seeding" clouds. The heat released by freezing, and by reduced reevaporation, would intensify vertical motions in the atmosphere and facilitate the development of winter cyclones; changes in the amount of cloud cover could alter climate over longer periods.

Tinsley freely admits that his concepts are "all still hypothetical." He observes, however, that the distribution of current in the global atmospheric electrical circuit varies in step with the level of solar activity and with the intensity of cyclones and related atmospheric instabilities. More significantly, he finds that some atmospheric phenomena correlate with magnetic storms and other solar wind effects but clearly are not associated with changes in solar ultraviolet radiation. Nevertheless, he continues to face dubious reactions even from some of his close colleagues. "He says he has a mechanism, but I still don't see how it works," Labitzke remarks. Rind is a bit more equivocal. "It is not out of the question that charged particles could affect clouds," he says cautiously, "but it would have to be proved through observations."

There Tinsley finds himself in a bit of a catch-22. Because of doubts within the community, "I haven't had a peer-reviewed proposal funded in the past five years," he reports. Tinsley hopes he or other researchers will be able to carry out laboratory tests to help nail down the validity of his ideas. And studies of day-to-day effects of the ever changing atmospheric electrical circuit "should show if the physics I've outlined works," he says doggedly.

Resolving the debate will not be easy. "We are trying to unscramble a very scrambled area," Labitzke says. Rind points out that so little is known about mechanisms linking the sun and weather that both hypotheses may be right—and that there may be others not yet

## Global Aid Wars

Poverty, it seems, is not foremost among the criteria by which wealthy nations choose to disburse their aid. The *Human Development Report 1994,* published by the United Nations Development Program, notes that two thirds of the world's poor get less than one third of the total development aid. And donor nations routinely tie assistance to military spending. In 1992 countries that spent more than 4 percent of their GDP on their military received $83 per capita in aid, whereas nations that spent less than 2 percent got $32.

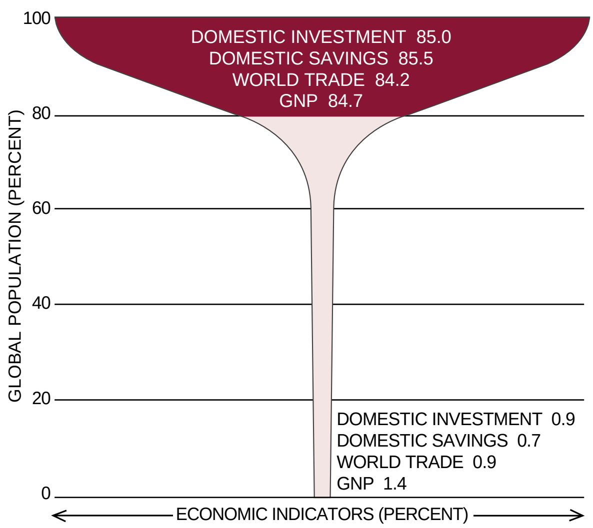
A large part of this imbalance is brought about by bilateral donors, who offer not just military but economic aid to strategic allies. For instance, Israel and Egypt will receive more than $2 billion of the $7.4 billion of bilateral assistance the U.S. plans to give in 1994. (The two nations receive an additional $3.1 billion in military assistance from the U.S. every year.) The U.S., Russia, China, France and the U.K.—the five permanent members of the U.N. Security Council—continue to supply the most weapons to developing countries.

Although multilateral institutions are more evenhanded—the World Bank gives about half its aid to two thirds of the world's poor—they do not redress the imbalance. As a result, a Brazilian woman living below the poverty line receives $3 in such support a year, whereas her Egyptian counterpart receives $280.

These days far more foreign capital flows to developing countries in the form of private investment instead of aid. In 1992 more than $100 billion was invested—as opposed to the $60 billion donated. Unfortunately for the poorest of the poor, this form of cash flow misses them, too. In the late 1980s sub-Saharan Africa received only 6 percent of foreign direct investment.

Trade, another means by which developing countries earn foreign capital, also benefits the more developed—and illustrates the ambivalence of wealthy states toward the world's poor. Although poverty wins a measure of sympathy, the cheap workforce of poor nations makes them an economic threat. By one estimate, if developed countries lifted all trade barriers to Third World goods, the latter would gain in exports twice what they now receive in aid.

Another constraint on the development of the Third World—foreign debt—keeps growing. In 1970 total debt was $100 billion; in 1992 it stood at $1.5 trillion, including service charges. During the decade preceding 1992, net financial transfers related to loans amounted to $125 billion—from the developing to the developed world. —*Madhusree Mukerjee*



*SOURCE:* Human Development Report 1994, *U.N. Development Program*

***WORLD'S RICHES are unevenly distributed: one fifth of the population has four fifths of the wealth.***

dreamed up. But the mere existence of clear-cut connections between tiny variations in the sun and measurable changes on the earth demonstrates that amazingly delicate balances are at work in the atmosphere. "The climate system has extreme points of sensitivity that were not previously appreciated," Rind observes—a sensitivity that could turn out to be relevant to changes wrought by humans in addition to those doled out by the sun. —*Corey S. Powell*

## Sex, Death and Sugar

*Researchers try to "grow" societies on a computer*

In the trendy field of artificial life, researchers seek the rules underlying nature by mimicking it on a computer. Although most artificial lifers focus on colonies of bacteria or flocks of birds, Joshua M. Epstein and Robert L. Axtell are more ambitious. These two social scientists are trying to simulate and thereby understand the most complex of all biological phenomena: human societies.

The simulation shown here may look like red and blue dots moving around on a yellow background, but it actually shows the evolution of two human societies, complete with birth, sex, death, tribal conflict and other constants of nature. The blue and red dots are people, or "agents," to use the term favored by economists. The yellow regions represent food. Epstein and Axtell refer to this sustenance as sugar and to their artificial world as the Sugarscape.

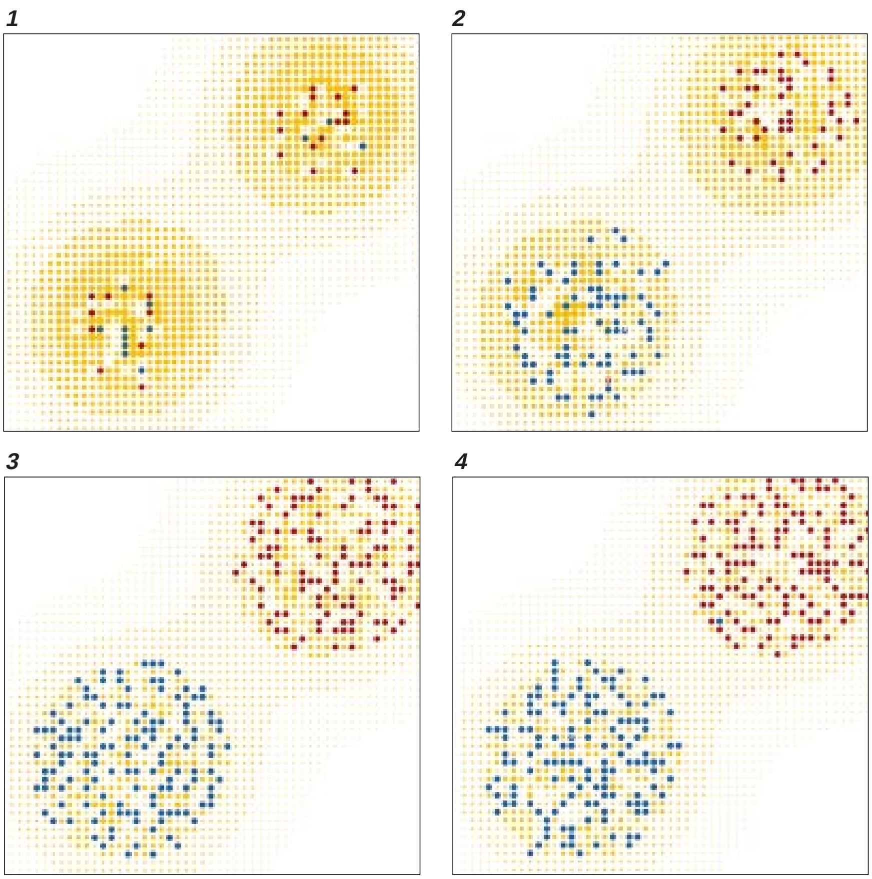
Epstein and Axtell, who hold joint appointments at the Brookings Institution in Washington, D.C., and the Santa Fe Institute in New Mexico (the latter is a hotbed of artificial life), consider the Sugarscape to be a laboratory in which they can test ideas about social evolution. Whereas most economists and social scientists build large-scale demographic trends into their models, Epstein and Axtell take a more bottom-up approach. They want to show how such trends may emerge, or "grow," from the interactions of individual agents. Conventional models, if they employ such agents at all, usually bless them with attributes rarely seen in the real world, such as immortality and a perfect knowledge of their economic environment. Epstein and Axtell have sought to make their agents more, well, human.

For example, not all agents are born equal in the Sugarscape. Some can spot sugar at greater distances than can others, and some have metabolisms that allow them to survive on a given amount of food for longer periods. Natural selection thus comes into play. Agents are either male or female, and each one belongs to one of two tribes: red or blue. When a red agent moves next to a blue agent (or vice versa), the red agent has a better than random chance of converting the stranger to his or her tribe. If a male and female of either tribe meet, they may have children if both are of childbearing age and have enough food stored up. The children inherit the vision, metabolism and tribal affiliation of their parents according to a simple Mendelian scheme. If the agents do not starve, they eventually die of old age.

In the first picture of the sequence, agents are scattered at random across the Sugarscape. They soon migrate toward the two sugar-rich mountains, where they begin to reproduce more rapidly than they die; the population of each mountain also becomes ethnically homogeneous. As the populations soar, the tribes consume the sugar faster than it can be replenished, and some agents venture away from their mountains in search of new sources of food. In the final picture, a blue "forager" enters red territory, where he or she can try to convert blues to red or be converted.

In more complicated simulations, Epstein and Axtell have investigated the effects of combat (one agent can kill another and steal his or her sugar), trade (agents can exchange sugar for another resource, "spice"), infectious diseases, pollution and the inheritance of wealth. The researchers claim that their agent-based simulations generate many of the same results—such as the tendency of inheritance rules to suppress natural selection and make populations more susceptible to disease—that scientists have observed in the real world.

Epstein concedes that the simulations are still merely "cartoons" compared with the intricacies of modern societies, but he thinks they may offer insights into the evolution of relatively simple cultures. He and Axtell are now collaborating with archaeologists affiliated with the Santa Fe Institute. The group is trying to understand the rise and sudden fall of the Anasazi, a civilization that thrived in the southwest U.S. from A.D. 1000 to 1300. One archaeologist, George J. Gumerman of Southern Illinois University, hopes the Sugarscape may illuminate links between maize production and population fluctuations




***"SUGARSCAPE" simulation shows agents from different tribes*** **(red and blue dots)** ***migrating toward mountains rich in sugar*** **(yellow),** ***where populations soar. In the final panel*** **(bottom right),** ***one blue agent has infiltrated the red tribe.***

JOSHUA M. EPSTEIN AND ROBERT L. AXTELL

among the Anasazi. Gumerman does not expect the simulations to provide specific answers but to serve as a "prosthesis for the mind." —*John Horgan*

## Brain Storm

*Controlling chaos could help treat epilepsy*

Chaos once seemed less a new frontier of science than an absolute limit. Take weather, the archetypal chaotic system. Weather exhibits cyclic behavior of a sort, and it conforms to certain boundary conditions. Yet the meteorologist Edward N. Lorenz pointed out decades ago that the fluttering of a butterfly's wings in Iowa could, in principle, trigger a typhoon in Bangladesh. This "butterfly effect" makes weather unpredictable and, by implication, uncontrollable.

Over the past several years, however, researchers have learned how to master chaos in systems as diverse as lasers, electronic circuits and heart tissue by exploiting their sensitivity to minute influences. Now experiments reported in *Nature* have raised hopes that similar chaos-control techniques can arrest the neural storms that trigger epilepsy. Some epilepsy patients can be treated only with surgery, which can permanently impair cognitive functions.

The research—done by a group that included Steven J. Schiff, a neurosurgeon at the George Washington University School of Medicine, and two physicists, William L. Ditto of the Georgia Institute of Technology and Mark L. Spano of the Naval Surface Warfare Center—involved a slice of a rat's hippocampus. This region of the brain is thought to be a primary source of epileptic seizures. When placed in a solution containing potassium, hippocampal neurons emit electrical pulses resembling those observed in human epileptics before the occurrence of a seizure. These firing patterns, in which clusters of 1,000 or so neurons discharge simultaneously, are known as interictal spikes. On plotting the timing of the spikes, the investigators saw a familiar sight. The spikes exhibited the same quasiperiodic patterns that chaotic lasers and heart muscles do.

The workers then delivered electrical pulses to this in vitro "brain." By varying the timing of pulses, the researchers were able to nudge the hippocampal neurons toward either more periodic or, conversely, more chaotic firing. These two methods are called control and anticontrol, respectively. The investigators suspect that anticontrol may be the most promising method for preventing epileptic seizures. Indeed, previous studies have indicated that highly periodic neural stimulation may be more likely to induce seizures than prevent them, Schiff says.

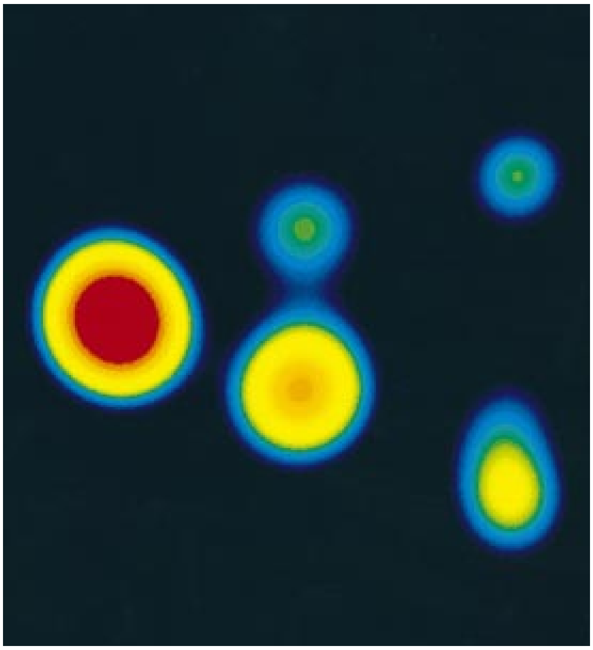
"More chaos may be better than less," agrees Walter J. Freeman, a neuroscientist at the University of California at Berkeley. Freeman's own research has suggested that chaos plays a vital role in perception and other brain functions: the chaotic firing of neurons may form a kind of carrier wave that can respond rapidly to the most subtle of signals. Studies have also indicated that mental disorders such as Alzheimer's disease may be associated with excessive periodicity, Freeman says.

Schiff emphasizes that many questions remain about the causes of—let alone the possible treatments for—epileptic attacks. Within the next year he and his colleagues plan to address these issues in trials with human subjects—epileptics who have already had electrodes implanted in their brains to monitor their seizures.

The group hopes that many years from now its work may lead to an implantable device that can both foresee and forestall the onset of a seizure. Such a device could be programmed to learn from experience and adopt the best possible strategy for each patient. To prevent adverse side effects, Schiff says, "you want minimal intervention." Ideally, the neural cyclones ravaging the brains of epileptics may be quelled by electrical pulses as gentle as the gust from a butterfly's wing. —*John Horgan*

MARTIN H. SIMON *SABA*

***SLICE OF A RAT'S HIPPOCAMPUS, monitored by electrodes, served as a model of an epileptic's brain in recent experiments on chaos-control techniques.***




*"MICROJETS" flee an invisible central object in this sequence of radio-spectrum images. An optical illusion makes the bottom clump seem to move faster than light.*

## Microquasars

*Giant blobs fly faster than light (sort of) in our own Milky Way*

Discovered several decades ago, quasars remain among the most mysterious of all denizens of the cosmic deep. They blaze with much greater intensity than do ordinary galaxies, and they often spout superluminal jets, plumes of matter so fast-moving that an optical illusion makes them seem to exceed the speed of light. Theorists suspect that buried within each quasar is a gargantuan, spinning black hole spewing matter from its poles, but quasars are so distant that their inner workings cannot be discerned.

Astronomers are therefore thrilled to find that the Milky Way, our celestial backyard, also harbors an object powerful enough to generate one of the hallmarks of quasars: superluminal jets. I. Felix Mirabel of the Saclay Research Center in France and Luis F. Rodriguez of the National Autonomous University in Mexico City discovered the "microquasar" using the Very Large Array at the National Radio Astronomy Observatory in New Mexico. The Very Large Array had just undergone an upgrade that improved its resolution. Miller Goss, an official at the radio observatory, says the finding "offers the best opportunity yet" to understand how quasars generate superluminal jets.

Mirabel and Rodriguez were monitoring an intense x-ray source known as GRS 1915+105, which lies some 40,000 light-years away from the earth, when it expelled two blobs of matter. The microjets are minute—only about as massive as the moon—in comparison to the galaxy-size plumes generated by true quasars, but they are just as speedy. One microjet, which is directed obliquely toward the earth, seems to be traveling at 1.25 times the speed of light as a result of an effect predicted by Einstein's theory of special relativity. After correcting for this phenomenon, Mirabel and Rodriguez calculated that the jets are actually hurtling away from their launchpad at 92 percent of the speed of light.

The researchers believe GRS 1915+105 is a spinning black hole or neutron star dragging matter from some unseen companion (probably an ordinary star) and flinging it back into space along its poles. Although Mirabel might have spotted the companion's infrared glow, other efforts to learn more about the microquasar have been stymied by its position near the Milky Way's crowded, dusty center. In a report in *Nature,* however, Mirabel and Rodriguez suggest that where there is one microquasar, there must be more. That prophecy may have already been fulfilled by another group using the Very Large Array. In late August workers led by Robert M. Hjellming spotted plumes of matter, one of which may be superluminal, hurtling from an x-ray source a mere 11,000 light-years away. *—John Horgan*

---

## A Nova Burns Out

*A premature death poses questions for astronomers*

Nova V1974 Cygni had a short, violent and public life. Exploding in February 1992, it was the brightest nova in 17 years and had by far the largest and best-equipped audience. The glowing gases it blew off evolved just as Sumner Starrfield of Arizona State University and his collaborators had predicted 20 years earlier.

Their model fit beautifully until a group led by Joachim Krautter of Heidelberg Observatory checked up on the nova with the *ROSAT* satellite in December 1993. The team could no longer see the x-rays coming from the underlying hot core. Inexplicably, the nova had turned off. This summer Steven N. Shore of Indiana University at South Bend and his co-workers confirmed the demise. "Honestly, I thought it would live another 10 years," Starrfield remarks.

The sudden end of V1974 Cygni has put long-standing models of nova explosions in turmoil. Novae occur in binary systems in which a white dwarf and a star about the size of the sun orbit each other. The dense white dwarf, having an intense gravitational field, pulls gases off its neighbor and onto itself. The deepest layers of the accreted gas are compressed and therefore heated, until a thermonuclear runaway reaction like that in a hydrogen bomb starts up. The surface of the white dwarf explodes, shedding its outer layers.

A massive white dwarf—such as the one that hosted V1974 Cygni—compresses its accumulated gases more intensely than does a smaller white dwarf. The higher pressure causes the nuclear explosion to occur before much material has collected. Because the fuel runs out faster, the explosion is short-lived. The brief existence of V1974 Cygni implies that its accreted layers had no more than $10^{-5}$ of the mass of the sun. But much more material—some $10^{-4}$ solar mass—appears to have been expelled in the explosion.

Shore has a possible explanation for this discrepancy of a factor of 10. He believes the key to the problem lies in another intriguing feature of V1974 Cygni: it threw out knots of slow-moving material along with the gases. (The resolving power of the *Hubble Space Telescope* made these structures apparent to researchers for the first time.) If these clumps turn out to contain large quantities of heavy elements such as neon, they could be coming from deep within the core. In that case, the explosion must have hurled out not just outer layers but chunks of the interior of the white dwarf itself.

The first indications from the spectra of the knots, taken with the *International Ultraviolet Explorer* satellite, support this view. But current models cannot describe the complex clumps and filaments in the ejecta. "We're going to start the calculations all over again," says Shore, who, along with Starrfield, will describe the nova in the December issue. The death of nova V1974 Cygni may prove even more illuminating than its birth. *—Madhusree Mukerjee*


***FRACTAL BRANCHING** of a drainage network, such as this one in the mudflats of San Pablo Bay near San Francisco, suggests self-organized criticality at work. Such systems may form because they are minimizing their energy expenditure.*

## Branching Out

*Rivers suggest a new feature of self-organized criticality*

It is easy to be critical about the complexity of life. Trickier, though, is using complexity to explain critical behavior. That is what some physicists recently accomplished after examining the fractal branching of a river drainage network. They uncovered a mechanism that may govern a variety of unpredictable phenomena, from the rumbling of an earthquake to a crash of the Dow Jones.

Such catastrophes are often attributed to some arbitrary, random event—a sudden slip at a fault, say, or extraordinarily bad investment advice. In actuality, these occurrences may just be following the principles of self-organized criticality. This idea propounds that complicated interactive systems can evolve toward a wobbly condition in which the slightest disturbance may elicit a major disaster. The pedagogic example typically invoked is the building of a sandpile grain by grain. Once the pile rises to a certain height—a critical state—it avalanches. Besides its utility in modeling earthquake and economic activity, the concept has found its way into evolutionary biology, solid-state physics and astronomy.

Now Ignacio Rodriguez-Iturbe of Texas A&M University, Andrea Rinaldo of the University of Padua in Italy and Rafael L. Bras of the Massachusetts Institute of Technology and their co-workers think they may have uncovered another mechanism through which self-organized criticality operates. Rather than emerging from events happening nearby or taking place immediately before, critical catastrophes may occur because of a global, long-term mechanism. Specifically, the systems may be minimizing the amount of energy they expend in maintaining themselves, thereby optimizing the way in which they develop. "It's a unique application," comments Per Bak of Brookhaven National Laboratory and a founder of self-organized criticality studies. "It's a specific model that's different from any other model."

The physicists' conclusion comes from comparing two different network models: one derived from a real river drainage system and the other from a mathematical model. To study actual drainage networks, Rodriguez-Iturbe and his colleagues used digital maps from the U.S. Geological Survey. These maps enabled them to look at how runoff and erosion locally modify the landscape. They also could calculate the energy expended in creating the river network. Using this information, the group members created models of drainage basins. The subsequent rearrangement of the landscape by erosion is analogous to the sandpile avalanches.

The researchers then compared this finding with a purely statistical model. Called an optimal channel network, this simulation was based on global rules about how a network minimizes its expenditure of energy. Generally speaking, the workers constructed a complicated network based on a few simple optimization rules. The strategy is similar to techniques designed to study the "traveling salesman" problem, in which the goal is to find the shortest travel distance between several cities.

The team discovered that the statistical properties of both models were exactly the same. Indeed, the computer produced images indistinguishable from patterns formed by real river basins. A key point of the research, however, was that both types of networks obeyed so-called power laws. Such rules are inherent in all self-organized critical processes. The number of earthquakes exceeding a given magnitude, for instance, depends on the size of previous earthquakes. In the drainage system the length of the stream channels and the distribution of branches and of the energy at any point all obey these laws.

The finding raises the question of whether all self-organized critical systems evolve through some global principle of energy minimization. Rodriguez-Iturbe speculates that in earthquake models, tectonic plates may be organizing themselves locally so as to minimize stress for the entire system. Optimization rules might apply to other fractal structures, such as the shapes of tree branches. “Maybe the trees are optimizing their leaf distribution to take advantage of the light they receive,” Rodriguez-Iturbe ponders.

The results are still too new to enable investigators to draw definite conclusions about widespread applicability. “It’s an intriguing idea, to get self-organized criticality out of a minimization principle, but it’s not clear how deep that is,” Bak warns. “There’s a feeling of tautology—someone can always take some local dynamical rule and write it as a minimization.”

Even if the rules are not broadly applicable, other uses may be possible. “It may improve our ability to forecast the behavior of river basins that have not been observed for long periods,” says Peter S. Eagleson, a civil engineer at M.I.T. The power-law distribution, for instance, may indicate where a river network has the most energy to yield, marking a good spot for the construction of a hydroelectric plant. Most important, Eagleson observes, is that the work helps us “to understand why the world is the way it is.” *—Philip Yam*

## Some Like It Hot—and Cold

An epically vast analysis of plant diversity at 94 locations scattered across the globe has produced answers to a long-standing biological riddle. The question is what controls the number of species in a region. In addition to its scholarly importance, the new work could put the study of biodiversity on surer scientific footing and so benefit efforts to slow the rapid pace at which animals and plants are going extinct.

Samuel M. Scheiner of Northern Illinois University, together with Jose M. Rey-Benayas, now at the Council for Scientific Research in Madrid, tabulated data on all the plants counted by dozens of ecologists around the world. The locations surveyed range from the Russian Arctic to Chile and from Alaska to the Australian outback. Many separate sites were studied in each place, so the researchers could assess how species are distributed in different areas. By pooling results from diverse sources, Scheiner and Rey-Benayas could assemble a database much larger than any one worker could have hoped to create. They published their results this past August in *Evolutionary Ecology.*

When all the numbers were crunched, the principal conclusion was that more plant species are found in well-lit, well-watered places where a lot of photosynthesis is taking place than in less favored spots. The result is not as tautological as it sounds. There is no reason that, a priori, a well-provided environment should encourage diversity. It might instead support a particular species or a small range of them. After all, corporations in an established industry tend to grow bigger and squeeze one another out.

What was more surprising was Scheiner and Rey-Benayas’s further finding that places where the temperature between summer and winter varies strongly have more species than do those areas blessed with relatively even seasons. This discovery clearly contradicts one of the dominant hypotheses ecologists have entertained about species numbers: the idea that equable climes make it easier for new species to evolve and persist.

Warm locations with large seasonal temperature fluctuations not only had more species, they also were more likely to have different varieties at different sampling sites. That result, too, is consistent with the idea that a large seasonality in temperature boosts the variety of species in a region. The tropics, in this view, are varied and ecologically rich despite having low seasonality, not because of it.

GARY BUSS FPG

***HOH RIVER VALLEY in Olympic National Park in Washington State is a haven for plant diversity. Recent research suggests that strong seasonal temperature swings—as opposed to stable climates—favor large numbers of species.***

Scheiner points out that the results suggest a way to estimate how global climatic change might affect biodiversity. Computer models of the atmosphere and oceans can project how the climate in different regions might change. The connections between climate and biological diversity that Scheiner and Rey-Benayas have found could then in principle be used to estimate how the numbers of plant species in different regions would be affected by the expected changes.

First, though, other workers will have to confirm and refine the approach, a task that will certainly take many years. In the meantime, both plants and animals are vanishing at a rate unprecedented in human history. Edward O. Wilson of Harvard University, the prominent biologist, has estimated that from 4,000 to 6,000 species are lost annually because of the destruction of tropical rain forests. *—Tim Beardsley*

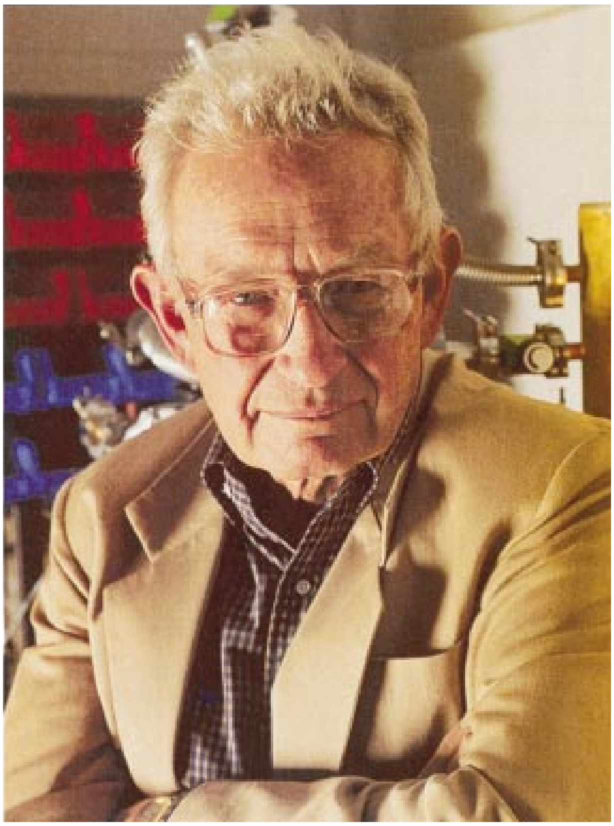
## Gruff Guru of Condensed-Matter Physics

Philip W. Anderson speaks in a ponderous growl, pausing between sentences to plot his next move. His basal expression, too, is deadpan. But like some exotic alloy in an unstable state, his mood can flip in an instant between different modes.

Anderson, a condensed-matter theorist at Princeton University, has just returned from an interdisciplinary conference held in Colorado, and his face brightens as he recounts a session on the genetics of cancer. He marvels at the "incredible profusion of reproduction that has to go exactly right" for cells to avoid becoming cancerous. Researchers will have to discover "God knows how many new principles" to account for this phenomenon, he exults.

But then Anderson frowns, remembering a "horrible" session on one of his specialties, high-temperature superconductors. Discovered eight years ago, the materials still resist theoretical explanations. Anderson accuses his fellow workers of being "naive" and of "looking under the streetlight" instead of venturing away from known territory for solutions to their problems. The 70-year-old Nobel laureate bemoans his relation to the rest of the field. "Sometimes I feel like a large, slow target," he confesses. "To say something a little cynical, sometimes I think the aspiring young man dreams of showing up the prestigious old character. It certainly would do his reputation no good simply to show that I was right."

Anderson is definitely a large, if not slow, target. He is in some respects an ideal leader for American physics in the post–Superconducting Super Collider world. His preternatural ability to intuit the meaning of experiments has made him a "commanding presence" in condensed-matter physics for more than 40 years, in the words of one colleague. He is a potent proselytizer, writing with lyrical fervor on how the interplay of order and disorder found in condensed-matter systems can serve as a metaphor for life itself. Anderson also began challenging the hegemony of particle physics and of the entire reductionist paradigm of modern science long before it became fashionable to do so. His 1972 article "More Is Different" remains a rallying cry for antireductionist approaches to science, notably chaos theory and complexity studies.



ROBERT PROCHNOW

***ANDERSON was an early champion of antireductionism.***

Readers of Anderson's essays in *Physics Today* and elsewhere—many of which were just published by World Scientific in a book entitled *A Career in Theoretical Physics*—know he can be a gruff guru. He complained four years ago that young scientists "don't seem to realize that the possession of a Ph.D. never guaranteed a career in basic research: it is and should be the privilege of a small elite." He has scolded physicists, including a colleague at Princeton, for reporting evidence on psychic phenomena. Such claims should be handled not by other physicists, Anderson argued, but by "those who are more used to dealing in flimflam, such as magicians and policemen."

Anderson insists that he never tries to stir up trouble. "I just call 'em as I see 'em," he remarks. He recognizes that he is sometimes perceived as being "excessively dogmatic and authoritarian." That perception distresses him. "I've been a rebel so often," he says in an uncharacteristically pleading tone.

Anderson's contrarianism emerged while he was seeking a doctorate in physics at Harvard University after World War II. Other students were flocking to the lectures of Julian Schwinger, whose quantum theory of electromagnetism became a milestone in particle physics. Anderson was put off by Schwinger's highly abstract style—and his popularity. "There was this tremendous excitement and gang of people around Schwinger, and I didn't want that," Anderson explains.

Drawn to topics he felt had more practical relevance, Anderson took a job at Bell Laboratories, whose researchers were "head and shoulders above the rest of the industry." He is not one of those who deplores the cutbacks in pure research brought about by the divestment of AT&T. During the previous two decades, he contends, such research had "spread like a contagion" and was carried out "by people without the same level of competence" as in the 1950s.

At Bell Labs, theorists such as Anderson were encouraged to work closely with an experimentalist. In the 1950s his partner, George Feher, found that once impurities in semiconductors reached a certain density, they suddenly became barriers to the conductance of electrons. The results disagreed with current theory, and so most theorists ignored them. Not Anderson. He showed that such results could be a generic property of a lattice whose atoms are randomly arranged.

Experimentalists later showed that Anderson's so-called localization theory could explain effects not only in condensed matter but also in plasmas and trapped electromagnetic radiation. Ultimately, it was this work on localization that earned him the Nobel Prize in 1977, along with John H. Van Vleck, his adviser at Harvard, and Sir Nevill F. Mott of the University of Cambridge, with whom Anderson had collaborated.

Anderson went on to show how magnetic resonance, ferromagnetism, superfluidity and other antics of condensed matter could be understood in the light of a concept called symmetry breaking. For example, a liquid crystal, which consists of molecules that act like minute magnets, is at its most symmetrical when the molecules are randomly aligned. That symmetry is "broken" and replaced by a new, more restrictive symmetry when a current is applied to the crystal, forcing the molecules to line up in the same direction.

Anderson began expounding these ideas during a teaching stint at Cambridge in the early 1960s. He inspired one of his students, Brian D. Josephson, to propose that electrons in a superconducting circuit might "tunnel" through an insulating barrier. Anderson and a colleague at Bell Labs experimentally validated Josephson's prediction, and in 1973 Josephson received the Nobel Prize for his discovery, now known as the Josephson effect.

To Anderson's dismay, Josephson later proclaimed his belief in psychic phenomena and stopped doing conventional physics. Two years ago, after *Physics Today* published Anderson's attack on psychic research, the journal printed a letter in which Josephson accused Anderson of being "caught in a paradigm that may be outliving its relevance." Anderson retorts that "the Brian Josephson who wrote that letter is not the same Brian Josephson who discovered the Josephson effect."

Particle physicists, too, have borrowed Anderson's ideas. In the mid-1960s Peter W. Higgs suggested that a symmetry-breaking mechanism similar to one proposed by Anderson might have caused particles to acquire their masses when the universe was still very young and hot. The Higgs boson, the particle that supposedly precipitates the symmetry breaking, has become the most prized quarry of particle physics; it was the raison d'être of the SSC.

In spite of his contribution to the field, Anderson had long been piqued by what he felt was the arrogance of particle physicists. By the late 1960s, he recalls, "the particle physicists were claiming on all sides that they were doing the fundamental science, and what the rest of us were doing was just engineering." Anderson challenged this contention in his "More Is Different" article, pointing out that reality has a hierarchical structure, with each level independent, to some degree, of the levels above and below. "At each stage entirely new laws, concepts, and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one," Anderson argued. "Psychology is not applied biology, nor is biology applied chemistry."

*Anderson's only regret about the SSC's death is that it took so long to happen.*

As a result of the essay, Anderson found himself invited to different kinds of conferences than those to which he was accustomed. In the late 1970s he attended a meeting on neurophysiology, where he discussed how symmetry breaking might help illuminate mental processes. His fellow speakers included an authority on psychedelic drugs and religious experiences and a Marxist who espoused a physics of social evolution. "I won't call them weirdos, because I was one of them," Anderson muses.

In the late 1980s Anderson was asked to testify at congressional hearings on the SSC. He asserted that—contrary to the claims of particle physicists—the collider would not address issues of uniquely fundamental importance, nor would it yield anything of practical value. "I confined myself to saying things that were manifestly true," Anderson recounts dryly. Asked if the decision of Congress to kill the accelerator once and for all last year left him with any regrets, Anderson replies, "I'm sorry that Congress let them go on so long." He does not mind that some particle physicists now refuse to speak to him, but he cannot forgive one Nobel laureate for being rude to his wife.

Anderson rejects the contention of some of these researchers that the SSC's death portends a growing antiscience movement in the U.S. "There have been yahoos in every period in American history. I don't think there is an effective antiscience movement." He is frightened by the religious right, but for political reasons: "If they get into power, we'll have to defend a hell of a lot besides science."

Particle physicist Murray Gell-Mann of the California Institute of Technology suggests in *The Quark and the Jaguar,* a book published this year, that Anderson might have supported the SSC if the theorists had named their quarry the Anderson-Higgs boson. Asked to respond to Gell-Mann's quip, Anderson glowers; he says only, "I didn't like that comment." Later he sniffs that Gell-Mann's book has a "very unsatisfying" explanation of how the relatively simple laws of physics could generate so much complexity. Gell-Mann's problem, Anderson notes, may be that he "has been away from the real nuts and bolts of physics too long."

Anderson and Gell-Mann have both lent their prestige to the decade-old Santa Fe Institute in New Mexico, which is a center for complexity studies. Although Anderson has touted the center in his writings, he is not wildly enthusiastic about the attempts of some researchers there to glean insights about nature from computer simulations. "That worries me," he acknowledges. "Since I know a little bit about global economic models, I know they don't work." He adds, "I always wonder whether global climate models and oceanic circulation models or things like that are as full of phony statistics and measurements." Anderson has urged the institute to devote more resources to real as opposed to artificial biology.

Anderson still teaches at Princeton, whose faculty he joined in 1975. (He retired from Bell Labs in 1984.) Teaching condensed-matter physics represents a special challenge, he remarks, one that is generally not well met. Most courses on condensed-matter physics make it sound "almost as bad as chemistry—just phenomenon after phenomenon after phenomenon." Anderson has sought to make symmetry breaking a unifying theme.

Last year he taught a course on "origins and beginnings," which probed ultimate questions in cosmology and biology. Anderson's faith that science can answer such questions seems tempered by an even deeper skepticism. He looks askance at the claims of some scientists—including several eminences at the Santa Fe Institute—that science may one day achieve a "theory of everything."

To be sure, there are certain scientific principles and mechanisms with wide applicability, such as evolution in biology and symmetry breaking in physics. "But you mustn't give in to the temptation to believe that a principle that works at one level will work at all levels," Anderson declares. Abruptly shifting to another mode, he throws up his hands and shouts, "I've finally seen the light! I understand everything!" With a rueful smile, he lowers his hands again. "You never understand everything. When one understands everything, one has gone crazy." —*John Horgan*

SCIENTIFIC AMERICAN

# Cerebrospinal Meningitis Epidemics

*A debilitating and often deadly disease, meningitis remains common in many developing countries. New insights may soon enable us to predict and control outbreaks*

by Patrick S. Moore and Claire V. Broome

By the middle of April 1988 the meningitis epidemic in N'Djamena, the capital of Chad, was in full swing. The outbreak had begun with a few isolated cases in mid-February; within four weeks nearly 150 patients were being admitted to the city's Central Hospital every day. As the facility ran out of bed space, people were treated in huge army tents scattered throughout the inner courtyards. Despite the best efforts of the Ministry of Health and foreign volunteer agencies, the epidemic spread. A shortage of medicines burdened health workers straining under the fatigue of seemingly endless days. Although a massive vaccination campaign was being implemented that would eventually stem the epidemic, each day threatened to paralyze further the country's fragile health care system.

By the time the scourge ended, 4,500 people had acquired meningitis, according to official statistics. Hundreds, or even thousands more, however, were uncounted. In Chad, as in many African countries, medical care is generally not available to people outside major cities. Meningitis sufferers who lived more than a day's walk from the nearest health station generally did not receive antibiotic treatment. Many died or were left with permanent brain damage.

The chaos and misery caused by the epidemic in Chad characterize most outbreaks of meningococcal meningitis, commonly known as spinal meningitis. The hallmark of the disease is its extremely rapid onset, which has earned this illness an uncommon degree of respect among medical experts. A healthy person first develops fever and malaise similar to that associated with influenza. Within hours these symptoms evolve into severe headache, neck rigidity and aversion to bright lights. If untreated, the patient can lapse into coma and eventually a fatal form of shock. Although it is now rare in the U.S., intense epidemics still affect most of the developing world; within weeks an entire country can be stricken.

Why do such epidemics occur? What causes a disease like meningitis to simmer within a population for years and then suddenly erupt? While many mysteries continue to surround the potentially lethal illness, its peculiar epidemiology offers some clues about the causes of meningitis epidemics and how to prevent them. The disease includes cycles of incidence that may correspond to environmental changes, to unusual patterns of immunity as well as to an association with still other infectious diseases. Medical detective work and the application of new biological techniques have begun to unveil some of these deadly secrets.

The bacterium causing meningococcal meningitis, called *Neisseria meningitidis* or meningococcus, is a close relative of the bacterium responsible for gonorrhea. Unlike gonorrhea, meningococci often colonize the lining of the throat and spread easily through respiratory secretions. The organism is so common that it can arguably be considered part of the normal human oral flora—at any given time, between 2 and 10 percent of healthy people carry meningococci. In all probability, most of us have been carriers at some point in our lives.

The throat's epithelial lining normally serves as a barrier to the bacteria, but occasionally the balance between colonization and invasion is disturbed. This imbalance results in life-threatening illness. Meningococcal meningitis begins when the organism invades the bloodstream and crosses the meninges—the membranes that surround the brain and spinal cord—to gain access to the cerebrospinal fluid that bathes the central nervous system. This fluid acts as a culture medium for the rapid growth of bacteria, which subsequently inflame the meningeal lining.

Typical symptoms, including fever, neck stiffness, headache and coma, result from the inflammation. In up to 30 percent of patients, profound septic shock also ensues as meningococci disseminate throughout the circulatory system. Such shock is characterized by a loss of blood pressure, particularly in the extremities. This reaction is probably caused by the release of an endotoxin from the bacterium, which stimulates the production of proteins, such as tumor necrosis factor-a. These factors, in turn, increase the permeability of blood vessels. Such a change can precipitate an often lethal drop in blood pressure. Patients who survive meningococcal septic shock may suffer a disfiguring loss of skin and parts of limbs. Meningococcal disease is generally fatal if left untreated; prompt intervention with antibiotics reduces its mortality rate to about 10 percent. Survivors may have residual neurological problems, such as deafness, paralysis and mental retardation.

Epidemics of meningitis have not taken place in the U.S. or most other industrial countries since World War II. Although carrying the bacterium is common, fewer than three cases of endemic meningococcal meningitis appear per 100,000 people in the U.S. every year.

**MENINGITIS PATIENT rests with his father in an African hospital. Treatment with antibiotics can reduce the mortality rate of this disease to 10 percent. But most of the people who become ill live in rural areas in developing countries and generally do not have access to medical care or to hospitals, as this boy does.**

PATRICK S. MOORE and CLAIRE V. BROOME specialize in public health and the prevention of infectious diseases. Currently a professor at Columbia University's School of Public Health, Moore has had a long-standing interest in the health of refugees and in understanding the process of epidemics. He is now using molecular techniques to study new and emerging pathogens. Broome serves as deputy director of the Centers for Disease Control and Prevention and as deputy administrator of the Agency for Toxic Substances and Disease Registry. Until 1990 she directed the Meningitis and Special Pathogens Branch at the National Center for Infectious Diseases.

Especially virulent epidemics, however, still happen in developing countries. More than 40,000 cases were documented during a 1989 epidemic in Ethiopia, and as many as three million cases may have occurred in China during the 1960s. The sudden influx of hundreds or even thousands of cases can overburden the often rudimentary health care systems in developing nations.

Because outbreaks are generally unpredictable and infrequent, the epidemic process itself is difficult to study. A wide band of countries in Africa, lying south of the Sahara Desert, form the "meningitis belt." This region consists of broad, grassy, savanna plains extending from the Gambia in West Africa across the continent to Ethiopia. Epidemiologists have known for decades that epidemics are particularly common in this region and tend to recur every five to 12 years. Each wave usually lasts for several years.

Within any given year meningitis rates follow a second, annual cycle: cases are highest during the dry season and disappear with the onset of rains. Even during the peak of an epidemic, incidence declines to baseline levels during the rainy season and rises again with the next dry season. Thus, it appears that when the bacterium is circulating in a susceptible population, some event during the dry season determines whether an epidemic will take hold.

These mysterious features of meningococcal scourges have intrigued epidemiologists for years. Unlike endemic disease, risk factors for an epidemic affect an entire population, not just scattered individuals. Because the risk of an epidemic varies over time, the factors responsible for initiating it must also vary. For example, in the U.S., people born with a rare genetic deficiency in their complement system (a series of blood proteins activated by antibodies to destroy bacteria) are unusually susceptible to meningococcal meningitis. Because the disease is rare in America, a substantial proportion of cases may result from this genetic deficiency. Yet the number of people with such a predisposition tends to be constant, so it is unlikely to be a significant cause of an epidemic. Indeed, studies conducted in Nigeria and in the Gambia have confirmed that patients with complement deficiencies were not commonly encountered during these outbreaks.

Unlike complement deficiencies, other host factors such as antibody levels against meningococci might change in a population over time. The general level of immunity in a population against an organism is called herd immunity, a term taken from early studies in livestock. Decline in the herd immunity of a population could be partially responsible for the cyclic patterns of meningitis in Africa.


**MENINGITIS BELT cuts across central Africa, from Eritrea in the East to the Gambia in the West. People living in this region appear to be uniquely susceptible to repeated epidemics of meningitis. Outbreaks in the belt are most often the result of infection by one particular strain of bacteria: serogroup A meningococci.**

In the late 1960s Irving Goldschneider and Emil C. Gotschlich and their colleagues at the Walter Reed Army Institute of Research performed elegant studies showing the importance of host defenses against meningococci. Before the advent of vaccines in the 1970s, military recruits were particularly susceptible to meningitis. The Walter Reed group drew blood from thousands of soldiers as they entered basic training and then followed them through boot camp. As men became ill, their stored sera were tested for the ability to kill meningococci and compared with the antimeningococcal activity in sera from their healthy counterparts.

The researchers found that the disease occurred primarily in recruits who had low antimeningococcal activity in their sera before they became ill. Because most adults have protective antibody activity against meningococci, this finding seemed unusual. Clearly, a person has to be exposed to the organism to develop antibodies. Yet the recruit studies indicated that those individuals without antibodies who were exposed to meningococci stood a high chance of becoming ill. It was unclear why most people developed protective antibodies after their first exposure to meningococci rather than becoming ill.

The answer to this paradox may lie in nonpathogenic bacteria that are part of the normal oral flora in humans. *Neisseria lactamica,* a relative of *N. meningitidis,* is one such organism. Ronald Gold and Martha L. Lepow of the University of Connecticut, along with the Walter Reed researchers, showed that young children who acquire *N. lactamica* throat infections frequently develop antibodies that are also protective against meningococci. Infection by one kind of nonpathogenic bacteria appears to protect against invasion by another, more virulent strain. This pattern provides an explanation for the experience of the recruits. Any soldier lacking cross-protective antibodies from a childhood infection of *Neisseria* would be especially susceptible because meningococcal strains from all over the country are rapidly spread among barracks mates.

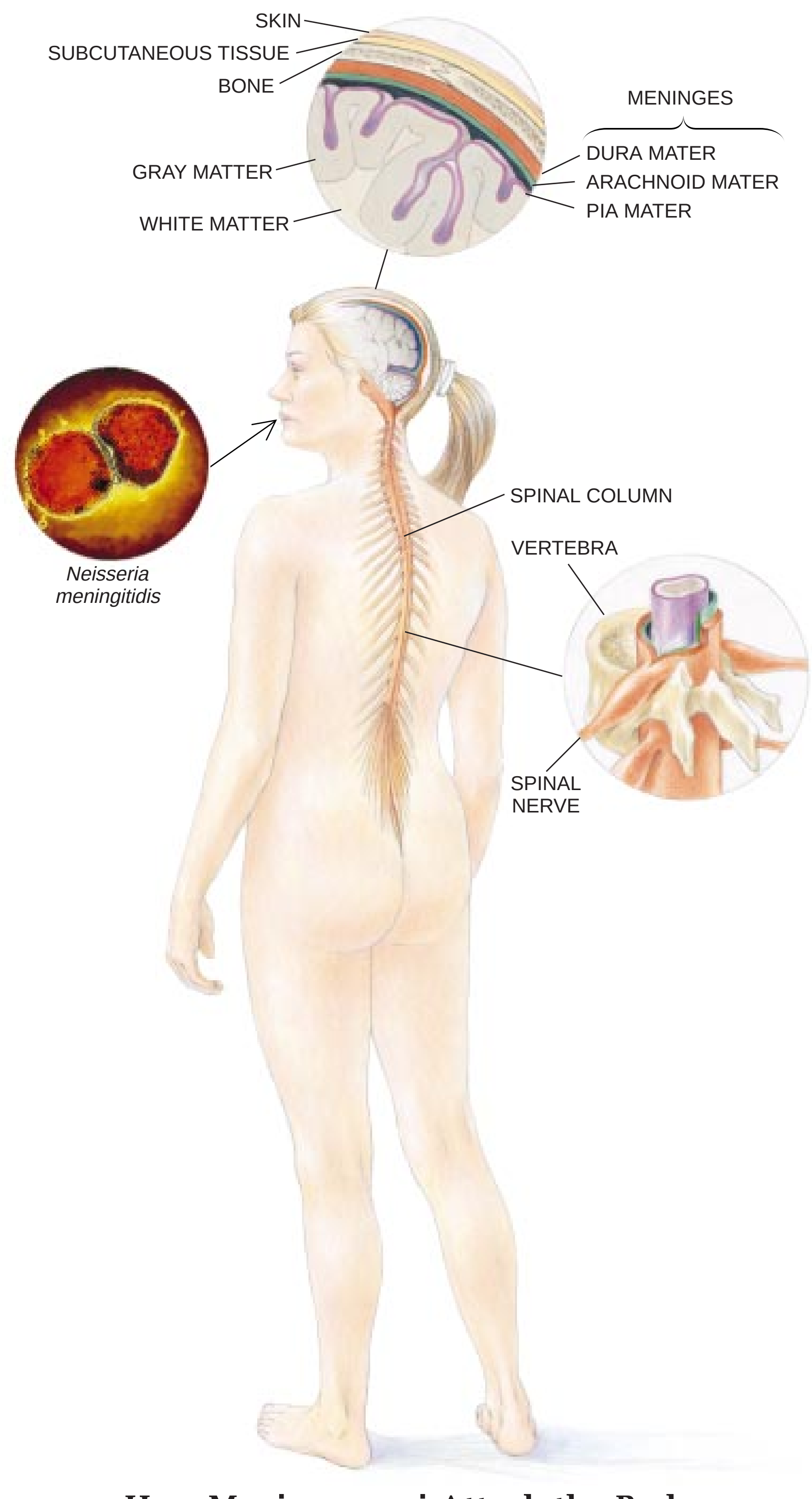
Research efforts are under way to determine which components of the meningococci actually bring about the protective immune response. One meningococcal antigen—that is, a molecule that stimulates an immune response—is the polysaccharide capsule that surrounds the organism. Meningococcal strains possess different polysaccharide antigens, and at least 13 polysaccharide types, called serogroups, have been found. Serogroup A *N. meningitidis* is responsible for the massive epidemics that periodically affect Africa, China and Latin America.

Other serogroups are less likely to cause epidemics, although they account for most of the endemic disease in the U.S. Vaccines made from a particular polysaccharide are very effective against the corresponding serogroup, yet they do not offer cross-protection. Unfortunately, the serogroup B polysaccharide, which is the most common serogroup in the U.S., does not elicit a persistent immune response.

Research by J. McLeod Griffiss and Robert E. Mandrell and others at the University of California at San Francisco, as well as Wendell D. Zollinger and his group at Walter Reed, suggests that antigens besides polysaccharides also play an important role in immunity. Because protection against the polysaccharide is not cross-protective, broad immunity from *N. lactamica* infection is likely caused by other antigens. In fact, *N. lactamica* probably does not possess a capsular polysaccharide at all. Other cellular components—such as outer membrane proteins and membrane-bound lipooligosaccharides common to *N. lactamica* and *N. meningitidis*—may confer immunity.

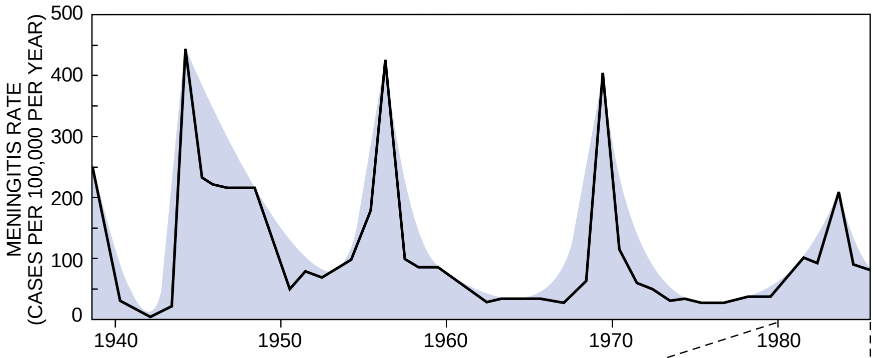
This mechanism of immunity could explain the five- to 12-year intervals between meningitis epidemics in Africa. High rates of meningococcal infection during an epidemic may provoke widespread natural immunity, which subsequently protects the population for several years. As immunity declines with the birth of susceptible children and with the natural loss of antibodies, a population would again become vulnerable to an epidemic.

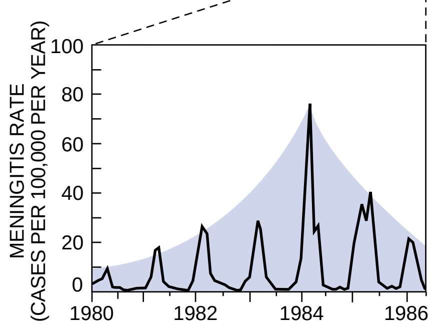
Nevertheless, the loss of immunity does not entirely explain the intriguing patterns of disease in the meningitis belt. The seasonality of disease indicates that environmental factors are also pivotal. Much of what is known about these factors comes from studies conducted in the Gambia and in




## How Meningococci Attack the Body

The bacterium that gives rise to cerebrospinal meningitis, *Neisseria meningitidis,* often thrives in the lining of the throat. Illness results when the organism enters the bloodstream and gains access to the meninges—the membranes that cover the brain and spinal cord. The pathogen grows rapidly in this environment, inflaming the meninges and causing fever, neck stiffness, headache and, often, coma. In up to 30 percent of patients the bacterium releases an endotoxin that increases the permeability of blood vessels. As a result, blood pressure falls; the ensuing septic shock can bring about the loss of skin and parts of limbs. The illness is generally fatal if untreated.




**ANNUAL MENINGITIS INCIDENCE in Burkina Faso follows patterns typical of countries in the meningitis belt. Epidemics tend to last for several years and follow a crescendo-decrescendo pattern (*above*). But even during an epidemic, meningitis rates are highly seasonal. As seen in this more detailed chart (*right*), outbreaks occur only during the dry season, from January until June.**

Nigeria by a team headed by Brian M. Greenwood of the U.K. Medical Research Council Laboratories. This group has been actively involved in uncovering the epidemiology of this disease, as well as designing new vaccination and control programs. Greenwood and his colleagues found that meningococcal transmission happens year-round, although meningococcal disease arises only during the dry season.

Furthermore, high antibody levels could be detected in some villages after the rainy season (during which no cases occurred), suggesting that meningococci had swept through the population, raising immunity without causing disease. The seasonality of meningitis in Africa, therefore, is not the result of increased transmission during the dry season. Instead it appears that high temperatures and low humidity make people more prone to disease once infected. Desiccation of the mucosal lining of the throat during the dry season might increase meningococcal colonization of the underlying tissues.

In addition to climate, viral upper respiratory infections may also affect the pharyngeal mucosae, making them more vulnerable to invasion. Bacterial pneumonia, for instance, can take hold after a viral infection. Our group at the Meningitis and Special Pathogens Branch of the Centers for Disease Control and Prevention (CDC) became interested in this possibility during a series of epidemics that began in the mid-1980s.

In August 1987 we were contacted by public health authorities in New York City and told that two people had contracted meningitis during separate airplane flights returning from Saudi Arabia. During the previous week, the CDC had received reports that meningococcal meningitis was breaking out among participants at the annual Muslim pilgrimage to Mecca, but it was not clear that an epidemic was under way. (The 1987 outbreak in Mecca eventually resulted in at least 10,000 cases of disease.) Lee H. Harrison, another epidemiologist, Gloria W. Ajello, a microbiologist, and one of us (Moore) left for John F. Kennedy International Airport to meet arriving planeloads of pilgrims.

We set up an assembly-line dispensary on the airport concourse, administering prophylactic antibiotics. Of the 550 passengers we examined, 11 percent of those returning from Mecca were carrying serogroup A meningococci—an exceedingly rare strain in the U.S. Persons who were carriers were more likely to be suffering from common cold symptoms, such as fever and sore throat, than were noncarriers. But because we were able to examine only meningococcal carriers rather than meningitis patients, the evidence was circumstantial. The next step needed was a direct search for upper respiratory infections in meningococcal meningitis patients.

We had a chance to investigate this question during the epidemic in Chad. One morning in April 1988 we received a call from Theo Lippeveld of the Harvard Institute for International Development. Lippeveld reported that N'Djamena, a city of 500,000 people in the center of the meningitis belt, was suffering from a major serogroup A meningococcal epidemic. The Ministry of Health, under the leadership of P. Matchock Yankalbé, was organizing control efforts in conjunction with the French Bioforce, a group of public health physicians formed by the French government, and the Merieux Institute in Marseilles. The Chad government gave approval for an investigation. So a CDC team traveled to N'Djamena for a collaborative study with physicians from the Central Hospital there.

In order to look for respiratory infections, we carefully matched persons with meningitis to a control group consisting of healthy persons of the same age, sex and neighborhood. Nasal washes were collected and shipped to John Hierholzer of the CDC, who began the arduous task of processing and culturing the hundreds of nasal washings. The results were surprising. Overall, meningitis patients were 23 times more likely to have an upper respiratory pathogen than were their matched controls. Not only were they more likely to have such viruses, but a large proportion was also infected with a small intracellular bacterium called *Mycoplasma hominis.*

These findings suggest another reason for the seasonality of meningitis epidemics. Perhaps a combination of low humidity and respiratory infections places a population at risk. Further studies are needed to clarify the mechanism by which respiratory infections interact with meningococci. Nevertheless, David S. Stephens of Emory University and Zell A. McGee of the University of Utah have shown in the laboratory that under similar circumstances meningococci are taken up by the cells of the pharyngeal lining. Respiratory infections could conceivably stimulate this uptake process. Alternatively, respiratory infections might directly damage the mucosae or inhibit immune cells there.

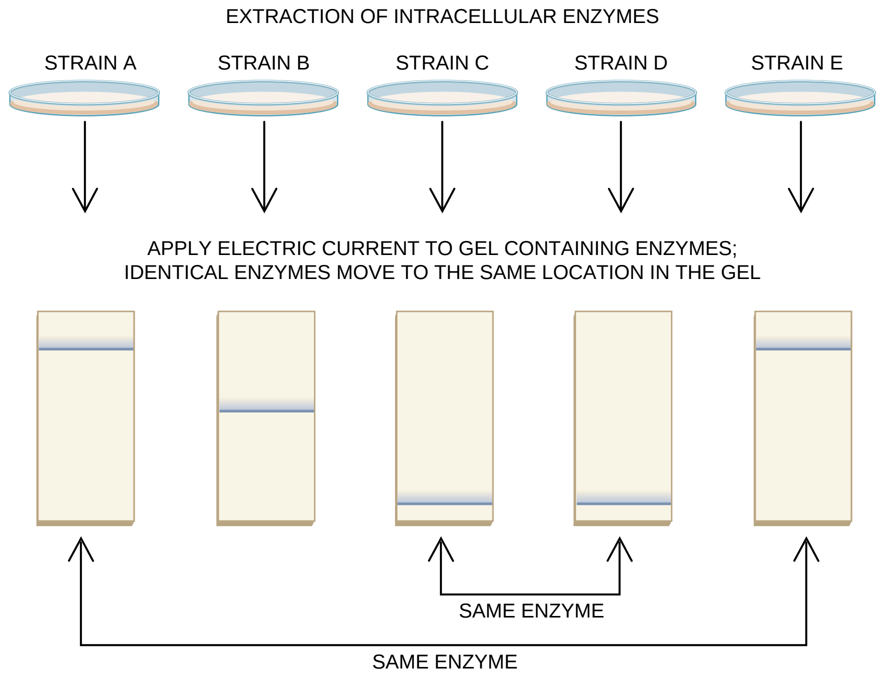
Respiratory infections have been associated with meningococcal disease in industrial countries as well, consistent with the fact that the illness is most prevalent during the midwinter months when cold viruses are common. Keith A. V. Cartwright, Dennis M. Jones and James M. Stuart and their colleagues at the Department of Public Health Medicine in Gloucester, England, recently noted a similar association between meningococcal disease and influenza infections. The same relation has been found by Bruno Hubert and his co-workers at the Direction Générale de la Santé in France. This research may lead to new ways to predict the occurrence of outbreaks in these countries. Ironically, the reasons upper respiratory infections are so seasonal remain elusive.

While we conducted our study in Chad, the Ministry of Health controlled the epidemic. In conjunction with the Harvard group, physicians from the Central Hospital and foreign volunteers, the epidemic was eventually stemmed. Numerous governments, including the U.S. and the French, provided aid during the epidemic. Nearly 1 percent of the entire population of N'Djamena contracted meningitis (for some groups, such as schoolchildren and soldiers, the rate may have been as high as 10 percent). Incidence would have been even higher without proper control measures. It was a clear example of the international mobilization needed when an epidemic strikes.

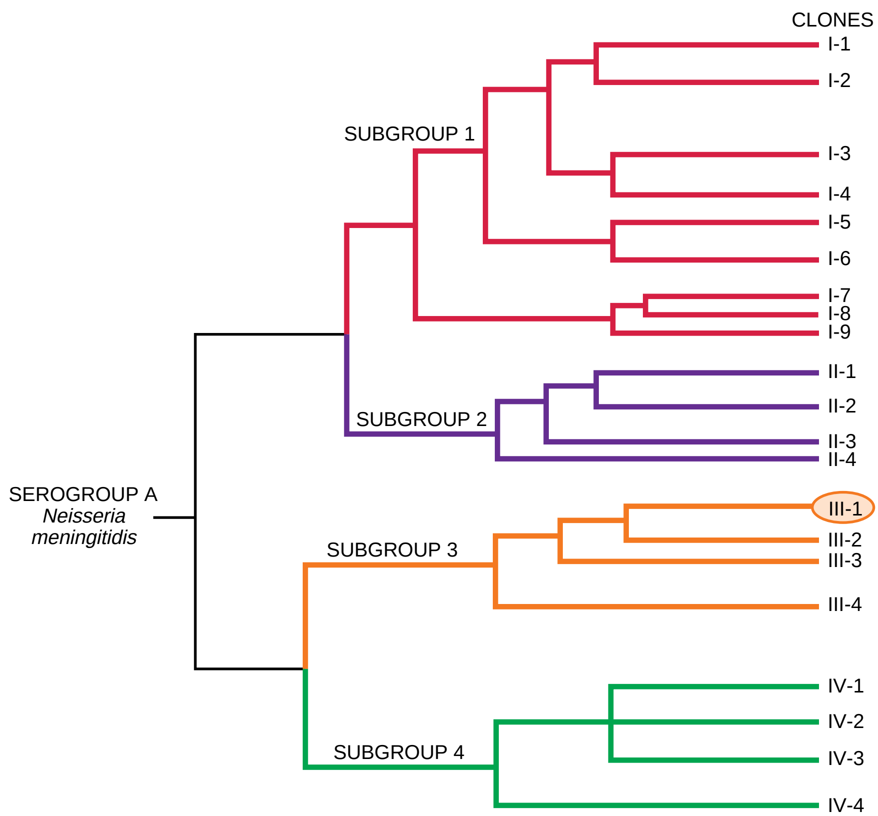
If antibodies, climate and respiratory infections are important for an epidemic to take shape, what role does the organism itself play? The new field of molecular epidemiology can help us answer that question. By borrowing techniques from molecular biology—such as DNA sequencing and enzyme electrophoresis—epidemiologists can now unravel the skein of outbreaks caused by the descendants of a single strain, or clone. These techniques have already been able to trace a number of bacterial and viral pathogens and proved particularly useful for documenting a case of human immunodeficiency virus transmission from a dentist in Florida to several of his patients.

Just as biologists can use the accumulation of mutations over time to track evolutionary divergence between two species, epidemiologists can analyze DNA, looking for mutations, to distinguish two strains of the same microorganism as they pass through human populations. To track serogroup A meningococcus, an indirect method of examining the genetic relatedness between different strains, called multilocus enzyme electrophoresis, has been extensively employed.

This intuitively simple but powerful application relies on detecting mutations that alter the amino acid sequence of bacterial enzymes. Generally, these mutations do not affect the enzyme's chemical activity. If they did, the strain would quickly die out. Still, minor mutations may cause differently charged amino acids to be incorporated into the enzyme, which can then be detected using electrophoresis. If the cytoplasm from two different strains is placed in a gel and an electric current is applied, enzymes will migrate through the gel at different speeds if there are differences in amino acid sequences.




**ENZYME ELECTROPHORESIS can be used to detect the divergence between strains of bacteria. Intracellular enzymes from various strains are placed in a gel and are separated when an electric current is applied to the mixture. Each enzyme is identified by individual mutations—which mark their divergence. The enzymes from different strains will migrate to different places in the gel. Multiple enzymes can be tested in this way to identify whether they are identical or not.**




*SOURCE: Mark Achtman, Max Planck Institute for Molecular Genetics, Berlin*

**GROUP A MENINGOCOCCI were once thought to be homogeneous. But enzyme electrophoresis studies have shown that group A meningococci are composed of at least 21 lineages, or clones. One clone, III-1, has recently caused epidemics in Asia, the Middle East and Africa.**

If two serogroup A meningococcal strains have recently diverged from each other, the likelihood is low that mutations will have accumulated in any given enzyme. An electrophoretic comparison of enzymes from the two strains will be similar. The more highly divergent the two strains are, the larger the number of enzymes that will be electrophoretically different between the strains. Enzymes from a number of strains can be compared using a statistical technique known as cluster analysis. This process can reveal the relative genetic divergence between strains. Each group of similar strains represents a single clone in which all the individual isolates are closely related and presumably derived from a single, recent ancestral cell. Researchers can examine strains and compose a family tree of different clones.

In an ambitious project, Mark Achtman, Tom Olyhoek and Brian A. Crowe of the Max Planck Institute for Molecular Genetics in Berlin used the technique to study 423 serogroup A meningococcal strains. Their analysis allowed them to outline the population genetics for this serogroup: the strains could be divided into four subgroups, which in turn were classified into a total of 21 different clones. Although this family tree is far from complete (many strains from outbreaks in developing countries have not been saved, and a fifth subgroup has recently been found), it provides a framework for comparing the epidemic potential of different strains.

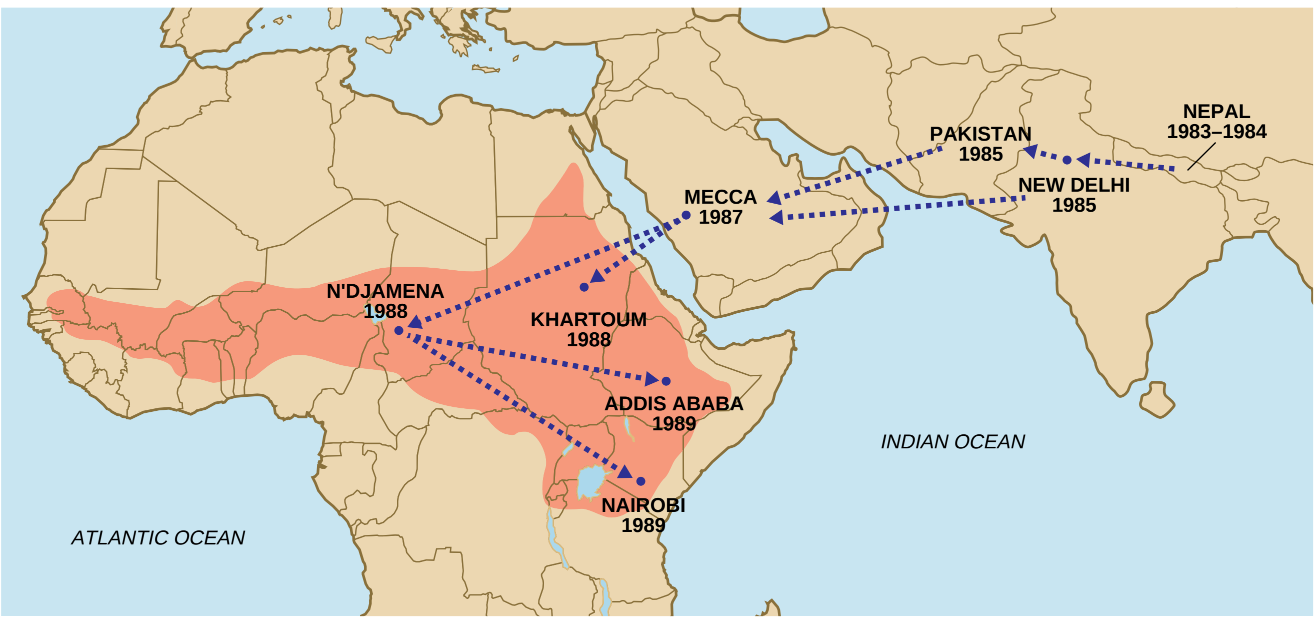
It is reasonable to assume that meningococcal strains diverge slowly over time as they pass through any given human population. Studies have shown that a variety of strains are present under endemic conditions. If all strains are equally virulent, and epidemics occur solely because of changes in host and environmental factors, then epidemics would tend to be polyclonal because each strain would be equally likely to cause disease. This, however, is not the case. The Berlin group, as well as Musa Hassan-King and Greenwood in the Gambia, has discovered that epidemics are generally the result of a single clone. Thus, clonal virulence may also play an important role.

With the help of Michael Reeves in the laboratory section of the Meningitis and Special Pathogens Branch, strains stored at the CDC from epidemics in South Asia, Africa and the Middle East were studied for evidence of clonal virulence. A striking correlation was found. Each epidemic had previously been considered an isolated event. Once the strains were compared, the connections became clear. Achtman's group had earlier determined that one clone, III-1, caused epidemics in China and Nepal in the early 1980s. The electrophoretic patterns of the strains stored at the CDC revealed that a series of epidemics in China, Nepal, Saudi Arabia and Chad were all caused by the same clone.

The III-1 clone first appeared in China in the 1960s. A second serious epidemic struck the Kathmandu Valley in Nepal shortly after roadways between Nepal and Tibet were opened in 1984. The strain spread to northern India and Pakistan in 1985, causing additional epidemics. Apparently it remained quiescent in South Asia until the summer of 1987. At that time, III-1 traveled with South Asian pilgrims to Mecca. Saudi Arabian and CDC epidemiologists who investigated the Mecca epidemic confirmed that it started among those pilgrims, who also had the highest attack rates. When the pilgrimage ended, III-1 carriers returned home. It was their return to the U.S. that triggered our investigation at Kennedy Airport.

Unfortunately, many Muslim pilgrims from meningitis-belt countries were also III-1 carriers. Not surprisingly, during the 1988 season, outbreaks sprouted simultaneously in Chad and Sudan. Subsequently, III-1 epidemics spread over East Africa, affecting Ethiopia, Tanzania, Kenya and Uganda. Public health officials are concerned about the potential for III-1 epidemics in other countries in the meningitis belt as well. Last year an epidemic of meningitis occurred in Togo, but it remains to be seen whether this epidemic was brought on by the III-1 strain.

Although III-1 has caused hundreds of thousands of meningitis cases, it does not appear to be uniquely virulent. Now that it is possible to perform clonal analysis of meningococcal strains, it is clear that other clones have caused similar epidemics in Africa and Asia. These findings do suggest, however, that the introduction of a potentially epidemic clone under the right circumstances can be devastating. Two explanations have




**VOYAGING BACTERIUM, serogroup A III-1 meningococci, has recently been responsible for an international pandemic. The pattern and timing of epidemics suggest that travelers carried the infection with them from Asia—where it originated—to Mecca and then to countries throughout the meningitis belt, including Ethiopia and Kenya.**

been given for this process: epidemic clones randomly expand as they progress through a population, or they survive by escaping herd immunity. As an analogy to influenza outbreaks, it has been proposed that epidemics might result from what are called antigenic shifts. Although all serogroup A meningococci share the same polysaccharide, individual clones differ in the other antigens exposed on the cell surface. Once immunity to the shared antigens wanes, a new clone with sufficiently different surface antigens might escape immune surveillance and start an epidemic. Epidemiologists following disease patterns will then see an "antigenic shift" as new clones supersede older clones.

If antigenic shifts do occur in meningococci, the cycles of disease seen in Africa would result from a combination of the time required for loss of immunity and the average time it takes for a new clone to enter the population. The environment would then also contribute because the introduction of the new clone alone is insufficient to start an outbreak. If the organism enters the population during the rainy season, it may boost immunity without causing disease. Although the exact conditions needed for an epidemic remain unclear, it appears that if the strain enters a population whose immunity is low during the dry season, there is a great risk of an epidemic.

Thus, a combination of host-, environmental- and organism-related factors seem to be responsible for the unique epidemiology of this disease. These characteristics are just beginning to be deciphered. Epidemics are probably not uniform, and perhaps other mechanisms may account for some, or all, of the features of meningitis outbreaks. For instance, why did not the III-1 strain cause an epidemic in the U.S.? The country has been free from sweeping epidemics since the 1940s, so it would seem that immunity against III-1 would be low. As any sufferer of the common cold knows, there is certainly no shortage of upper respiratory infections in the U.S. Furthermore, no one knows why childhood infections with *Neisseria* are so successful in protecting people in industrial countries but do not seem to have the same effect in African countries. Poorly defined socioeconomic factors seem to make industrial nations resistant. Although the antigenic-shift hypothesis is appealing, long-term studies in Africa are needed to determine its validity.

There is hope that new developments could reduce the threat of meningococcal epidemics. Available meningococcal vaccines are based on the polysaccharide capsule surrounding the bacterium and are not effective in infants who are vaccinated during routine vaccination programs. The protection generated by these vaccines is short-lived in children, and vaccinating them during nonepidemic periods does not protect them from the next wave.

**AIRPORT CLINIC was quickly set up at John F. Kennedy International in New York City in August 1987 to treat infected passengers. Pilgrims returning from Mecca were given antibiotics, and a potential epidemic was prevented.**

A novel technique of chemically linking capsular polysaccharides to a protein carrier and making a conjugate vaccine may overcome this problem. A protein-polysaccharide conjugate vaccine against *Hemophilus influenzae,* another bacterium causing meningitis, has been quite successful in infants. The World Health Organization has commissioned research to produce and test similar protein-polysaccharide conjugate vaccines against the group A meningococcus. Another exciting characteristic of these vaccines is their ability to decrease carriage of the organism by healthy persons. This feature could interrupt transmission of the organism and protect individuals from disease.

These insights, and others, may provide us with tools needed to vaccinate and protect the populations at risk for epidemics. But the hurdles facing the creation of new vaccines are political and economic as well as scientific. The entire annual health budget for some developing countries is less than $5 per capita, so creative approaches to vaccine production and purchasing are urgently needed. It is hoped that collaboration among manufacturers, international aid agencies and developing nations will be established to overcome these problems.

For now, early detection of an impending epidemic is crucial. Although the current vaccine does not confer long-term immunity, it can be used in emergency campaigns during an outbreak. Methods are being devised to guide vaccination campaigns based on the number of cases within a given population. New clones might also serve as an early-warning system. Jan T. Poolman of the Dutch National Institute for Public Health has developed monoclonal antibodies to detect serogroup A meningococcal clones. His work offers a simpler and quicker way to track clones.

The spread of III-1 shows how interconnected the global village has become. We were able to eliminate carriage of the III-1 strain from a mere fraction of U.S. pilgrims to Mecca. Fortunately, socioeconomic factors probably prevented this strain from causing outbreaks in the U.S. and Europe. We have not been so lucky with other diseases, such as AIDS. Testing and quarantining travelers has never worked, and it clearly will not work in the future. We will be able to protect people in developed countries only by making a commitment to monitoring and to public health, particularly in nations that have scarce medical resources.

FURTHER READING


THE EPIDEMIOLOGY OF ACUTE BACTERIAL MENINGITIS IN TROPICAL AFRICA. B. M. Greenwood in *Bacterial Meningitis.* Edited by J. D. Williams and J. Burnie. Academic Press, 1987.

GLOBAL EPIDEMIOLOGY OF MENINGOCOCCAL DISEASE. Benjamin Schwartz, Patrick S. Moore and Claire V. Broome in *Clinical Microbiology Reviews,* Vol. 2, Supplement, pages S118–S124; April 1989.

MENINGOCOCCAL MENINGITIS IN SUB-SAHARAN AFRICA: A MODEL FOR THE EPIDEMIC PROCESS. Patrick S. Moore in *Clinical Infectious Diseases,* Vol. 14, No. 2, pages 515–525; February 1992.

# The Self-Reproducing Inflationary Universe

*Recent versions of the inflationary scenario describe the universe as a self-generating fractal that sprouts other inflationary universes*

by Andrei Linde

If my colleagues and I are right, we may soon be saying good-bye to the idea that our universe was a single fireball created in the big bang. We are exploring a new theory based on a 15-year-old notion that the universe went through a stage of inflation. During that time, the theory holds, the cosmos became exponentially large within an infinitesimal fraction of a second. At the end of this period, the universe continued its evolution according to the big bang model. As workers refined this inflationary scenario, they uncovered some surprising consequences. One of them constitutes a fundamental change in how the cosmos is seen. Recent versions of inflationary theory assert that instead of being an expanding ball of fire the universe is a huge, growing fractal. It consists of many inflating balls that produce new balls, which in turn produce more balls, ad infinitum.

Cosmologists did not arbitrarily invent this rather peculiar vision of the universe. Several workers, first in Russia and later in the U.S., proposed the inflationary hypothesis that is the basis of its foundation. We did so to solve some of the complications left by the old big bang idea. In its standard form, the big bang theory maintains that the universe was born about 15 billion years ago from a cosmological singularity—a state in which the temperature and density are infinitely high. Of course, one cannot really speak in physical terms about these quantities as being infinite. One usually assumes that the current laws of physics did not apply then. They took hold only after the density of the universe dropped below the so-called Planck density, which equals about $10^{94}$ grams per cubic centimeter.

As the universe expanded, it gradually cooled. Remnants of the primordial cosmic fire still surround us in the form of the microwave background radiation. This radiation indicates that the temperature of the universe has dropped to 2.7 kelvins. The 1965 discovery of this background radiation by Arno A. Penzias and Robert W. Wilson of Bell Laboratories proved to be the crucial evidence in establishing the big bang theory as the preeminent theory of cosmology. The big bang theory also explained the abundances of hydrogen, helium and other elements in the universe.

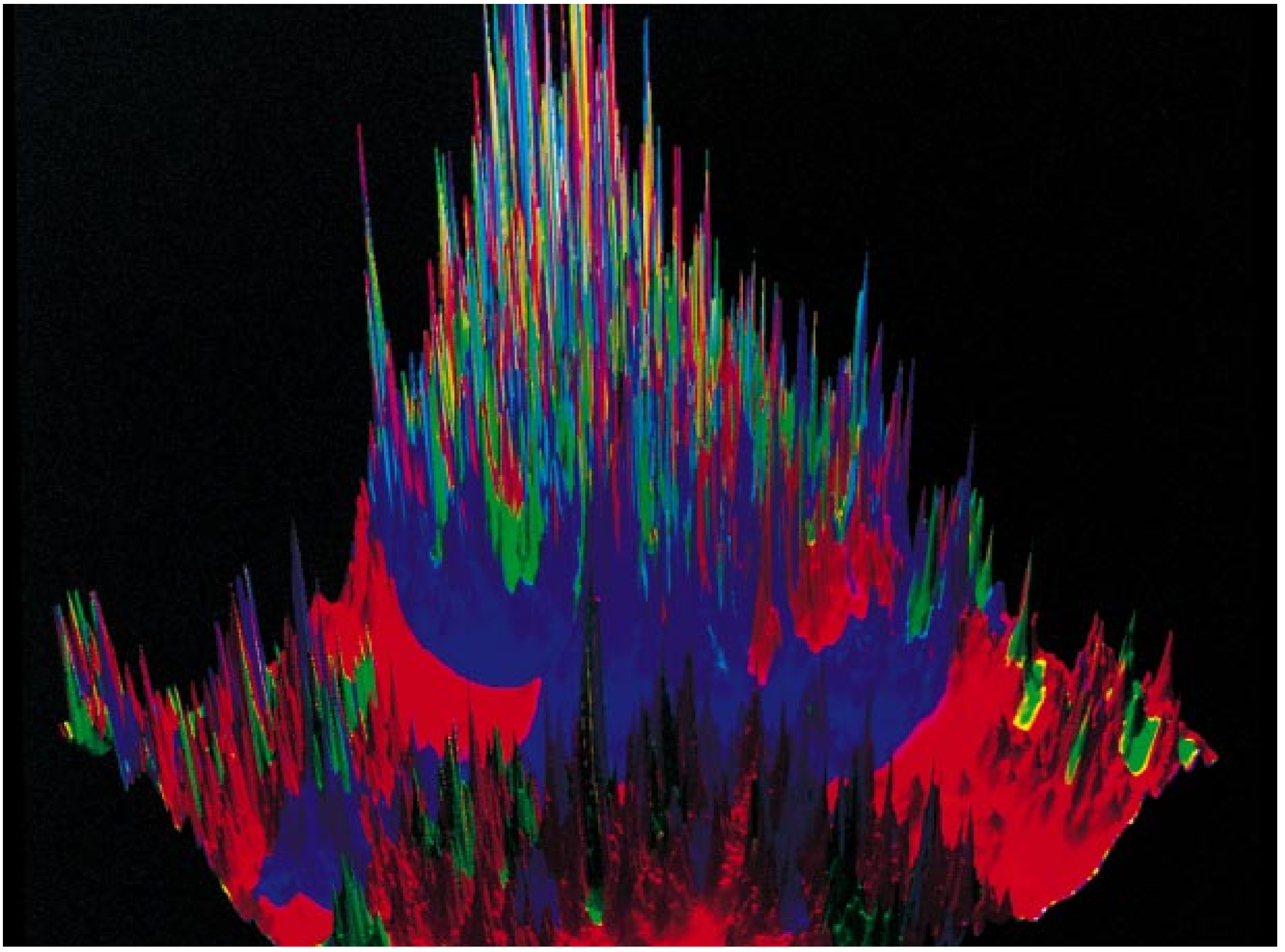
As investigators developed the theory, they uncovered complicated problems. For example, the standard big bang theory, coupled with the modern theory of elementary particles, predicts the existence of many superheavy particles carrying magnetic charge—that is, objects that have only one magnetic pole. These magnetic monopoles would have a typical mass $10^{16}$ times that of the proton, or about 0.00001 milligram. According to the standard big bang theory, monopoles should have emerged very early in the evolution of the universe and should now be as abundant as protons. In that case, the mean density of matter in the universe would be about 15 orders of magnitude greater than its present value, which is about $10^{-29}$ gram per cubic centimeter.

This and other puzzles forced physicists to look more attentively at the basic assumptions underlying the standard cosmological theory. And we found many to be highly suspicious. I will review six of the most difficult. The first, and main, problem is the very existence of the big bang. One may wonder, What came before? If space-time did not exist then, how could everything appear from nothing? What arose first: the universe or the laws determining its evolution? Explaining this initial singularity—where and when it all began—still remains the most intractable problem of modern cosmology.

A second trouble spot is the flatness of space. General relativity suggests that space may be very curved, with a typical radius on the order of the Planck length, or $10^{-33}$ centimeter. We see, however, that our universe is just about flat on a scale of $10^{28}$ centimeters, the radius of the observable part of the universe. This result of our observation differs from theoretical expectations by more than 60 orders of magnitude.

A similar discrepancy between theory and observations concerns the size of the universe. Cosmological examinations show that our part of the universe contains at least $10^{88}$ elementary particles. But why is the universe so big? If one takes a universe of a typical initial size given by the Planck length and a typical initial density equal to the Planck density, then, using the standard big bang theory, one can calculate how many elementary particles such a universe might encompass. The answer is rather unexpected: the entire universe should only be large enough to accommodate just one elementary particle—or at most 10 of them. It would be unable to house even a single reader of *Scientific American,* who consists of about

ANDREI LINDE is one of the originators of inflationary theory. After graduating from Moscow University, he received his Ph.D. at the P. N. Lebedev Physics Institute in Moscow, where he began probing the connections between particle physics and cosmology. He became a professor of physics at Stanford University in 1990. He lives at Stanford with his wife, Renata Kallosh (also a professor of physics at Stanford), and his sons, Dmitri and Alex. Besides theorizing about the birth of the cosmos, Linde also dabbles in stage stunts such as sleight-of-hand, acrobatics and hypnosis.



**SELF-REPRODUCING UNIVERSE in a computer simulation consists of exponentially large domains, each of which has different laws of physics (represented by colors). Sharp peaks are new "big bangs"; their heights correspond to the energy density of the universe there. At the top of the peaks, the colors rapidly fluctuate, indicating that the laws of physics there are not yet settled. They become fixed only in the valleys, one of which corresponds to the kind of universe we live in now.**

$10^{29}$ elementary particles. Obviously, something is wrong with this theory.

The fourth problem deals with the timing of the expansion. In its standard form, the big bang theory assumes that all parts of the universe began expanding simultaneously. But how could all the different parts of the universe synchronize the beginning of their expansion? Who gave the command?

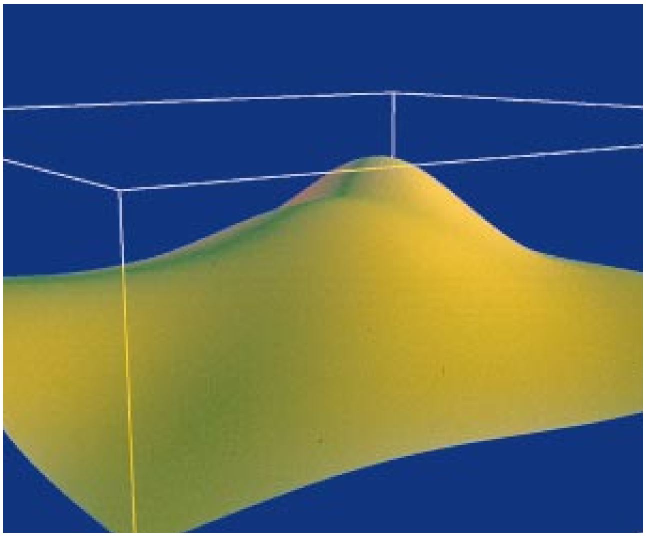
Fifth, there is the question about the distribution of matter in the universe. On the very large scale, matter has spread out with remarkable uniformity. Across more than 10 billion light-years, its distribution departs from perfect homogeneity by less than one part in 10,000. For a long time, nobody had any idea why the universe was so homogeneous. But those who do not have ideas sometimes have principles. One of the cornerstones of the standard cosmology was the "cosmological principle," which asserts that the universe must be homogeneous. This assumption, however, does not help much, because the universe incorporates important deviations from homogeneity, namely, stars, galaxies and other agglomerations of matter. Hence, we must explain why the universe is so uniform on large scales and at the same time suggest some mechanism that produces galaxies.

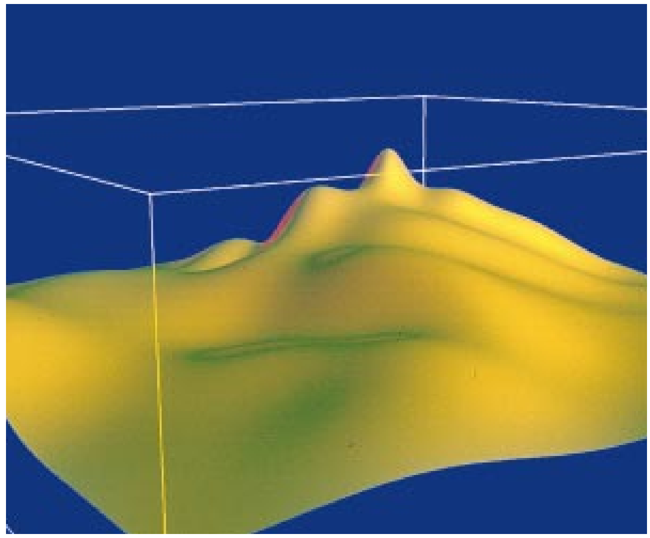
Finally, there is what I call the uniqueness problem. Albert Einstein captured its essence when he said: "What really interests me is whether God had any choice in the creation of the world." Indeed, slight changes in the physical constants of nature could have made the universe unfold in a completely different manner. For example, many popular theories of elementary particles assume that space-time originally had considerably more than four dimensions (three spatial and one temporal). In order to square theoretical calculations with the physical world in which we live, these models state that the extra dimensions have been "compactified," or shrunk to a small size and tucked away. But one may wonder why compactification stopped with four dimensions, not two or five.

Moreover, the manner in which the other dimensions become rolled up is significant, for it determines the values of the constants of nature and the masses of particles. In some theories, compactification can occur in billions of different ways. A few years ago it would have seemed rather meaningless to ask why space-time has four dimensions, why the gravitational constant is so small or why the proton is almost 2,000 times heavier than the electron. Now developments in elementary particle physics make answering these questions crucial to understanding the construction of our world.

All these problems (and others I have not mentioned) are extremely perplexing. That is why it is encouraging that many of these puzzles can be resolved in the context of the theory of the self-reproducing, inflationary universe.

The basic features of the inflationary scenario are rooted in the physics of el-





**EVOLUTION OF A SCALAR FIELD leads to many inflationary domains, as revealed in this sequence of computer-generated images. In most parts of the universe, the scalar field decreases (represented as depressions and valleys). In other places,**

ementary particles. So I would like to take you on a brief excursion into this realm—in particular, to the unified theory of weak and electromagnetic interactions. Both these forces exert themselves through particles. Photons mediate the electromagnetic force; the *W* and *Z* particles are responsible for the weak force. But whereas photons are massless, the *W* and *Z* particles are extremely heavy. To unify the weak and electromagnetic interactions despite the obvious differences between photons and the *W* and *Z* particles, physicists introduced so-called scalar fields.

Although scalar fields are not the stuff of everyday life, a familiar analogue exists. That is the electrostatic potential—the voltage in a circuit is an example. Electrical fields appear only if this potential is uneven, as it is between the poles of a battery or if the potential changes in time. If the entire universe had the same electrostatic potential, say, 110 volts, then nobody would notice it; the potential would seem to be just another vacuum state. Similarly, a constant scalar field looks like a vacuum: we do not see it even if we are surrounded by it.

These scalar fields fill the universe and mark their presence by affecting properties of elementary particles. If a scalar field interacts with the *W* and *Z* particles, they become heavy. Particles that do not interact with the scalar field, such as photons, remain light.

To describe elementary particle physics, therefore, physicists begin with a theory in which all particles initially are light and in which no fundamental difference between weak and electromagnetic interactions exists. This difference arises only later, when the universe expands and becomes filled by various scalar fields. The process by which the fundamental forces separate is called symmetry breaking. The particular value of the scalar field that appears in the universe is determined by the position of the minimum of its potential energy.

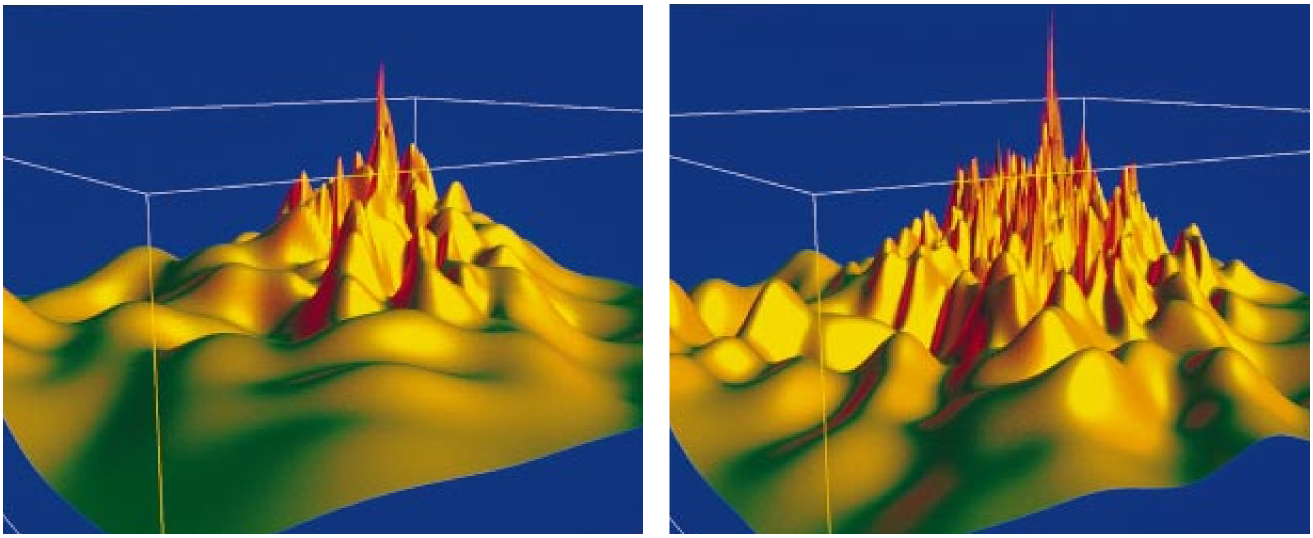
Scalar fields play a crucial role in cosmology as well as in particle physics. They provide the mechanism that generates the rapid inflation of the universe. Indeed, according to general relativity, the universe expands at a rate (approximately) proportional to the square root of its density. If the universe were filled by ordinary matter, then the density would rapidly decrease as the universe expanded. Therefore, the expansion of the universe would rapidly slow down as its density decreased. But because of the equivalence of mass and energy established by Einstein, the potential energy of the scalar field also contributes to the expansion. In certain cases, this energy decreases much more slowly than does the density of ordinary matter.

The persistence of this energy may lead to a stage of extremely rapid expansion, or inflation, of the universe. This possibility emerges even if one considers the very simplest version of the theory of a scalar field. In this version the potential energy reaches a minimum at the point where the scalar field vanishes. In this case, the larger the scalar field, the greater the potential energy. According to Einstein's theory of gravity, the energy of the scalar field must have caused the universe to expand very rapidly. The expansion slowed down when the scalar field reached the minimum of its potential energy.

One way to imagine the situation is to picture a ball rolling down the side of a large bowl [*see upper illustration on page 54*]. The bottom of the bowl represents the energy minimum. The position of the ball corresponds to the value of the scalar field. Of course, the equations describing the motion of the scalar field in an expanding universe are somewhat more complicated than the equations for the ball in an empty bowl. They contain an extra term corresponding to friction, or viscosity. This friction is akin to having molasses in the bowl. The viscosity of this liquid depends on the energy of the field: the higher the ball in the bowl is, the thicker the liquid will be. Therefore, if the field initially was very large, the energy dropped extremely slowly.

The sluggishness of the energy drop in the scalar field has a crucial implication in the expansion rate. The decline was so gradual that the potential energy of the scalar field remained almost constant as the universe expanded. This behavior contrasts sharply with that of ordinary matter, whose density rapidly decreases in an expanding universe. Thanks to the large energy of the scalar field, the universe continued to expand at a speed much greater than that predicted by preinflation cosmological theories. The size of the universe in this regime grew exponentially.

This stage of self-sustained, exponentially rapid inflation did not last long. Its duration could have been as short as $10^{-35}$ second. Once the energy of the field declined, the viscosity nearly disappeared, and inflation ended. Like the ball as it reaches the bottom of the bowl, the scalar field began to oscillate near the minimum of its potential ener-



**quantum fluctuations cause the scalar field to grow. In those places, represented as peaks, the universe rapidly expands, leading to the creation of inflationary regions. We live in one of the valleys, where space is no longer inflating.**

gy. As the scalar field oscillated, it lost energy, giving it up in the form of elementary particles. These particles interacted with one another and eventually settled down to some equilibrium temperature. From this time on, the standard big bang theory can describe the evolution of the universe.

The main difference between inflationary theory and the old cosmology becomes clear when one calculates the size of the universe at the end of inflation. Even if the universe at the beginning of inflation was as small as $10^{-33}$ centimeter, after $10^{-35}$ second of inflation this domain acquires an unbelievable size. According to some inflationary models, this size in centimeters can equal $10^{10^{12}}$—that is, a 1 followed by a trillion zeros. These numbers depend on the models used, but in most versions this size is many orders of magnitude greater than the size of the observable universe, or $10^{28}$ centimeters.

This tremendous spurt immediately solves most of the problems of the old cosmological theory. Our universe appears smooth and uniform because all inhomogeneities were stretched $10^{10^{12}}$ times. The density of primordial monopoles and other undesirable "defects" becomes exponentially diluted. (Recently we have found that monopoles may inflate themselves and thus effectively push themselves out of the observable universe.) The universe has become so large that we can now see just a tiny fraction of it. That is why, just like a small area on a surface of a huge inflated balloon, our part looks flat. That is why we do not need to demand that all parts of the universe began expanding simultaneously. One domain of a smallest possible size of $10^{-33}$ centimeter is more than enough to produce everything we see now.

Inflationary theory did not always look so conceptually simple. Attempts to obtain the stage of exponential expansion of the universe have a long history. Unfortunately, because of political barriers, this history is only partially known to American readers.

The first realistic version of the inflationary theory came in 1979 from Alexei A. Starobinsky of the L. D. Landau Institute of Theoretical Physics in Moscow. The Starobinsky model created a sensation among Russian astrophysicists, and for two years it remained the main topic of discussion at all conferences on cosmology in the Soviet Union. His model, however, was rather complicated (it was based on the theory of anomalies in quantum gravity) and did not say much about how inflation could actually start.

In 1981 Alan H. Guth of the Massachusetts Institute of Technology suggested that the hot universe at some intermediate stage could expand exponentially. His model derived from a theory that interpreted the development of the early universe as a series of phase transitions. This theory was proposed in 1972 by David A. Kirzhnits and me at the P. N. Lebedev Physics Institute in Moscow. According to this idea, as the universe expanded and cooled, it condensed into different forms. Water vapor undergoes such phase transitions. As it becomes cooler, the vapor condenses into water, which, if cooling continues, becomes ice.

Guth's idea called for inflation to occur when the universe was in an unstable, supercooled state. Supercooling is common during phase transitions; for example, water under the right circumstances remains liquid below zero degrees Celsius. Of course, supercooled water eventually freezes. That event would correspond to the end of the inflationary period. The idea to use supercooling for solving many problems of the big bang theory was exceptionally attractive. Unfortunately, as Guth himself pointed out, the postinflation universe of his scenario becomes extremely inhomogeneous. After investigating his model for a year, he finally renounced it in a paper he co-authored with Erick J. Weinberg of Columbia University.

In 1982 I introduced the so-called new inflationary universe scenario, which Andreas Albrecht and Paul J. Steinhardt of the University of Pennsylvania also later discovered [see "The Inflationary Universe," by Alan H. Guth and Paul J. Steinhardt; SCIENTIFIC AMERICAN, May 1984]. This scenario shrugged off the main problems of Guth's model. But it was still rather complicated and not very realistic.

Only a year later did I realize that inflation is a naturally emerging feature in many theories of elementary particles, including the simplest model of the scalar field discussed above. There is no need for quantum gravity effects, phase transitions, supercooling or even the standard assumption that the universe originally was hot. One just considers all possible kinds and values of scalar fields in the early universe and then checks to see if any of them leads to inflation. Those places where inflation does not occur remain small. Those domains where inflation takes place become exponentially large and dominate the total volume of the universe. Because the scalar fields can take arbitrary values in the early universe, I called this scenario chaotic inflation.

In many ways, chaotic inflation is so simple that it is hard to understand why the idea was not discovered sooner. I think the reason was purely psychological. The glorious successes of the big bang theory hypnotized cosmologists. We assumed that the entire universe was created at the same moment, that initially it was hot and that the scalar field from the beginning resided close to the minimum of its potential energy. Once we began relaxing these assumptions, we immediately found that inflation is not an exotic phenomenon invoked by theorists for solving their problems. It is a general regime that occurs in a wide class of theories of elementary particles.

That a rapid stretching of the universe can simultaneously resolve many difficult cosmological problems may seem too good to be true. Indeed, if all inhomogeneities were stretched away, how did galaxies form? The answer is that while removing previously existing inhomogeneities, inflation at the same time made new ones.

These inhomogeneities arise from quantum effects. According to quantum mechanics, empty space is not entirely

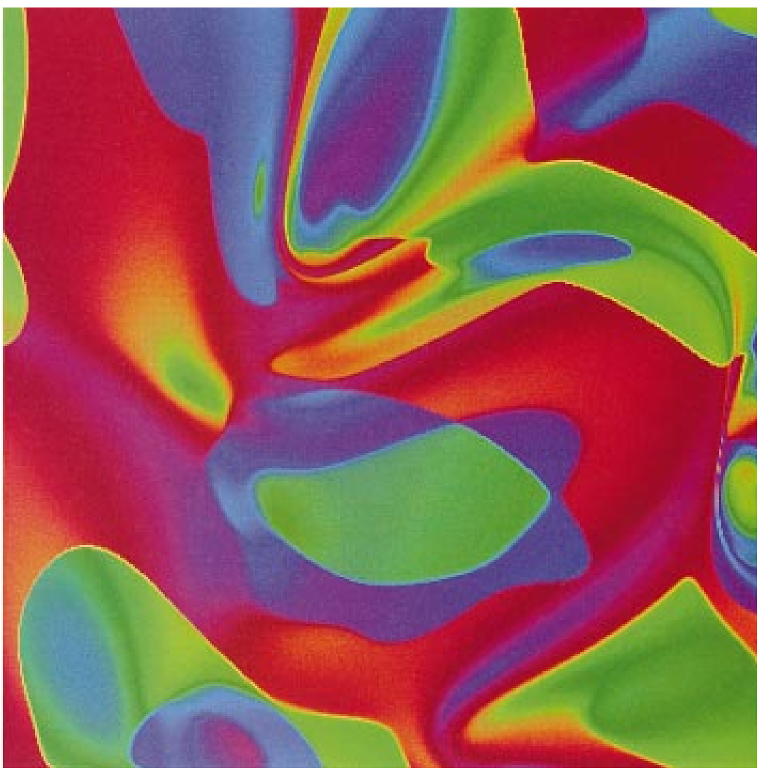
## On the Eighth Day...

The new cosmological theory is highly unusual and, understandably, may be difficult to picture. One of the main reasons for the popularity of the old big bang scenario is that imagining the universe as a balloon expanding out in all directions is relatively easy. It is much harder to grasp the structure of an eternally self-replicating fractal universe. Computer simulations can help to some extent. Here I will describe some of these simulations, which I performed with my son Dmitri, now a student at the California Institute of Technology.

We began our simulations with a two-dimensional slice of the universe filled by an almost homogeneous scalar field. We calculated how the scalar field changed in each point of our domain after the beginning of inflation. Then we added to this result sinusoidal waves, corresponding to the quantum fluctuations that freeze.

By continually applying this procedure, we obtained a sequence of figures that shows the distribution of the scalar field in the inflationary universe. (For viewing purposes, the computer shrank down the original image, rather than expanding the inflating domains.) The images revealed that in the main part of the original domain the scalar field slowly decreases [*see illustrations on pages 50 and 51*]. We live in such a part of the universe. Small waves frozen on top of an almost homogeneous field eventually give rise to the perturbations in temperature of the background radiation the *Cosmic Background Explorer* satellite discovered. Other parts of the picture show growing mountains, which correspond to huge energy densities that lead to extremely rapid inflation. Hence, one can interpret each peak as a new "big bang" that creates an inflationary "universe."

The fractal nature of the universe became even more apparent after we added in another scalar field. To render things even more interesting, we considered a theory in which the potential energy of this field has three different minima, represented as different colors [*see illustration on page 49*]. In a two-dimensional slice of the universe, the colors near the peaks of the mountains change all the time, indicating that the scalar field is rapidly jumping from one energy minimum to another. The laws of physics there are not yet fixed. But in the valleys, where the rate of expansion is slow, the colors no longer fluctuate. We live in one of such domains. Other domains are extremely far away from us. Properties of elementary particles and the laws of their interaction change as one crosses from one domain to another—one should think twice before doing so.



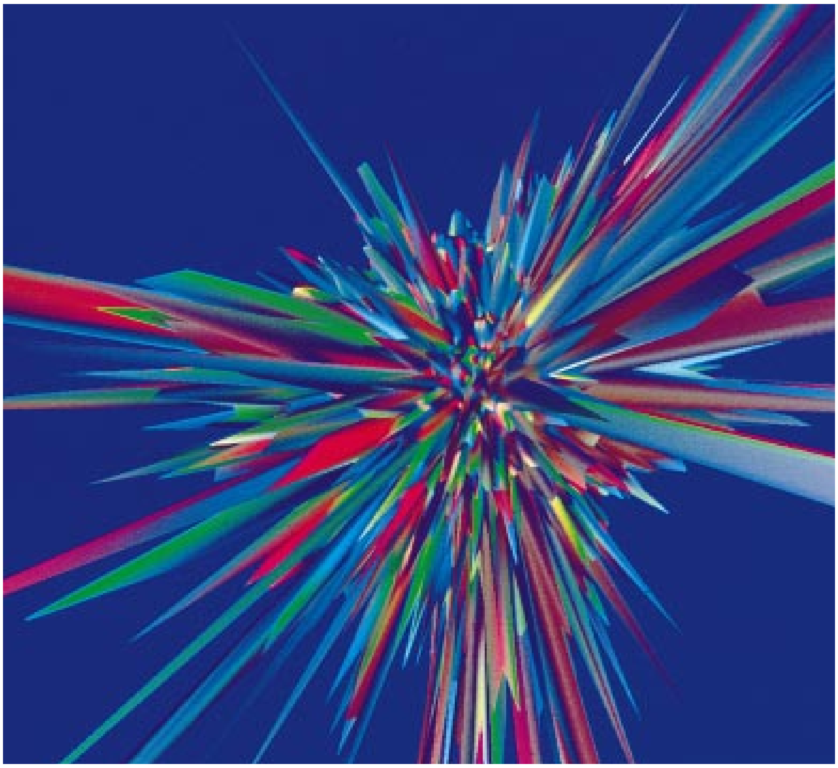
***A "Kandinsky" universe***

In another set of figures, we explored the fractallike nature of the inflationary universe along the lines of a different theory of particle physics. Describing the physical meaning of these images is harder. The strange color pattern (*left*) corresponds to the distribution of energy in the theory of axions (a kind of scalar field). We called it a Kandinsky universe, after the famous Russian abstractionist. Seen from a different perspective, the results of our simulations sometimes appear as exploding stars (*opposite page*).

We conducted the first series of our simulations several years ago after we persuaded Silicon Graphics in Los Angeles to loan us one of their most powerful computers for a week. Setting up the simulations was hard work, and only on the seventh day did we finish the first series of our calculations empty. The vacuum is filled with small quantum fluctuations. These fluctuations can be regarded as waves, or undulations in physical fields. The waves have all possible wavelengths and move in all directions. We cannot detect these waves, because they live only briefly and are microscopic.

In the inflationary universe the vacuum structure becomes even more complicated. Inflation rapidly stretches the waves. Once their wavelengths become sufficiently large, the undulations begin to "feel" the curvature of the universe. At this moment, they stop moving because of the viscosity of the scalar field (recall that the equations describing the field contain a friction term).

The first fluctuations to freeze are those that have large wavelengths. As the universe continues to expand, new fluctuations become stretched and freeze on top of other frozen waves. At this stage one cannot call these waves quantum fluctuations anymore. Most of them have extremely large wavelengths. Because these waves do not move and do not disappear, they enhance the value of the scalar field in some areas and depress it in others, thus creating inhomogeneities. These disturbances in the scalar field cause the density perturbations in the universe that are crucial for the subsequent formation of galaxies.



***An "explosion" of the scalar field***

and saw for the first time all these growing mountains that represent inflationary domains. We were able to fly between them and to enjoy a view of our universe at the first moments of creation. We looked at the shining screen, and we were happy—we saw that the universe is good! But our work did not last long. On the eighth day we returned the computer, and the machine's gigabyte hard drive crashed, taking with it the universe that we had created.

Now we continue our studies using different methods (and a different Silicon Graphics computer). But one can play an even more interesting game. Instead of watching the universe at the screen of a computer, one may try to create the universe in a laboratory. Such a notion is highly speculative, to say the least. But some people (including Alan H. Guth and me) do not want to discard this possibility completely out of hand. One would have to compress some matter in such a way as to allow quantum fluctuations to trigger inflation. Simple estimates in the context of the chaotic inflation scenario suggest that less than one milligram of matter may initiate an eternal, self-reproducing universe.

We still do not know whether this process is possible. The theory of quantum fluctuations that could lead to a new universe is extremely complicated. And even if it is possible to "bake" new universes, what shall we do with them? Can we send any message to their inhabitants, who would perceive their microscopic universe to be as big as we see ours? Is it conceivable that our own universe was created by a physicist-hacker? Someday we may find the answers.

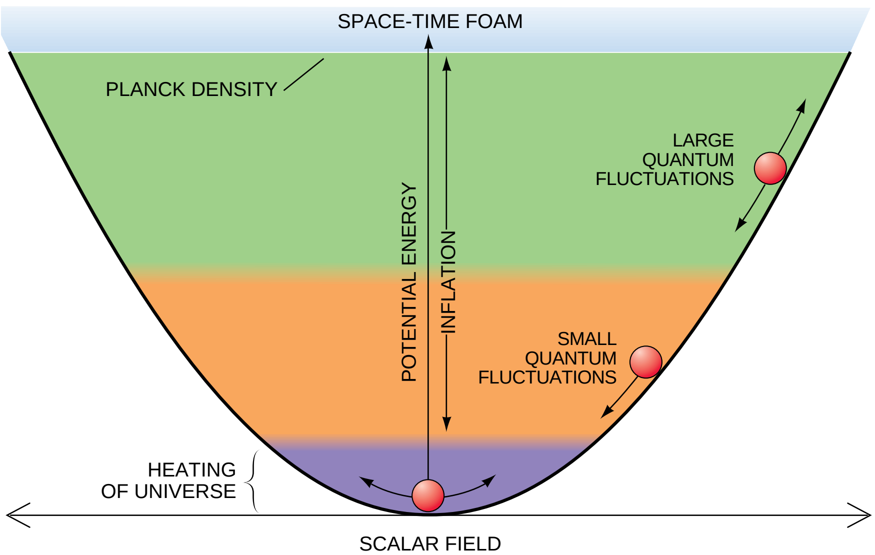
In addition to explaining many features of our world, inflationary theory makes several important and testable predictions. First, inflation predicts that the universe should be extremely flat. Flatness of the universe can be experimentally verified, because the density of a flat universe is related in a simple way to the speed of its expansion. So far observational data are consistent with this prediction.

Another testable prediction is related to density perturbations produced during inflation. These density perturbations affect the distribution of matter in the universe. Furthermore, they may be accompanied by gravitational waves. Both density perturbations and gravitational waves make their imprint on the microwave background radiation. They render the temperature of this radiation slightly different in various places in the sky. This nonuniformity is exactly what was found two years ago by the *Cosmic Background Explorer* (*COBE*) satellite, a finding later confirmed by several other experiments.

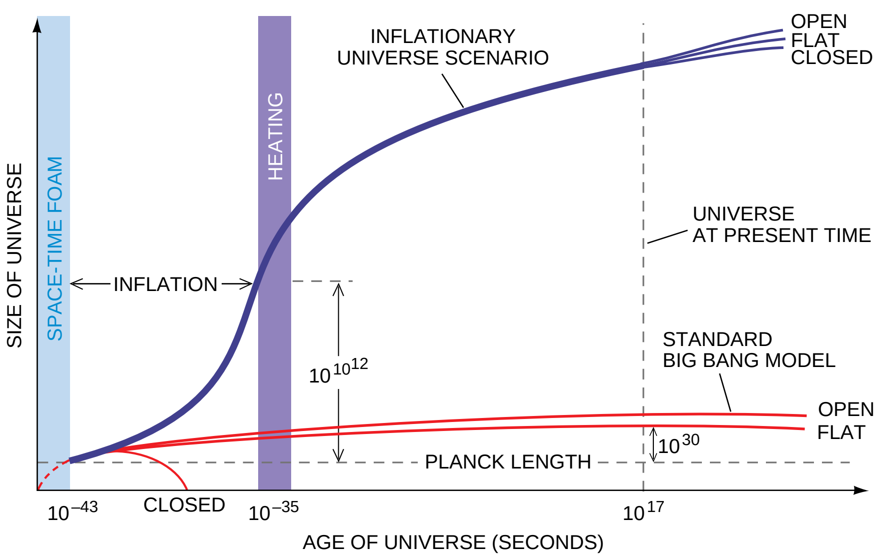
Although the *COBE* results agree with the predictions of inflation, it would be premature to claim that *COBE* has confirmed the inflationary theory. But it is certainly true that the results obtained by the satellite at their current level of precision could have definitively disproved most inflationary models, and it did not happen. At present, no other theory can simultaneously explain why the universe is so homogeneous and still predict the "ripples in space" discovered by *COBE*.

Nevertheless, we should keep an open mind. The possibility exists that some new observational data may contradict inflationary cosmology. For example, if observations tell us that the density of the universe is considerably different from the critical density, which corresponds to a flat universe, inflationary cosmology will face a real challenge. (It may be possible to resolve this problem if it appears, but it is fairly complex.)

Another complication has a purely theoretical origin. Inflationary models are based on the theory of elementary particles, and this theory by itself is not completely established. Some versions (most notably, superstring theory) do not automatically lead to inflation. Pulling inflation out of the superstring model may require radically new ideas. We should certainly continue the search for alternative cosmological theories. Many cosmologists, however, believe inflation, or something very similar to it, is absolutely essential for constructing a consistent cosmological theory. The inflationary theory itself changes as particle physics theory rapidly evolves. The list of new models includes extended inflation, natural inflation, hybrid inflation and many others. Each model has unique features that can be tested through observation or experiment. Most, however, are based on the idea of chaotic inflation.




**SCALAR FIELD in an inflationary universe can be modeled as a ball rolling down the side of a bowl. The rim corresponds to the Planck density of the universe, above which lies a space-time "foam," a region of strong quantum fluctuations. Below the rim (*green*), the fluctuations are weaker but may still ensure the self-reproduction of the universe. If the ball stays in the bowl, it moves into a less energetic region (*orange*), where it slides down very slowly. Inflation ends once the ball nears the energy minimum (*purple*), where it wobbles around and heats the universe.**




**EVOLUTION OF THE UNIVERSE differs in the chaotic inflation scenario and the standard big bang theory. Inflation increases the size of the universe by $10^{10^{12}}$, so that even parts as small as $10^{-33}$ centimeter (the Planck length) exceed the radius of the observable universe, or $10^{28}$ centimeters. Inflation also predicts space to be mostly flat, in which parallel lines remain "parallel." (Parallel lines in a closed universe intersect; in an open one, they ultimately diverge.) In contrast, the original hot big bang expansion would have increased a Planck-size universe to only 0.001 centimeter and would lead to different predictions about the geometry of space.**

Here we come to the most interesting part of our story, to the theory of an eternally existing, self-reproducing inflationary universe. This theory is rather general, but it looks especially promising and leads to the most dramatic consequences in the context of the chaotic inflation scenario.

As I already mentioned, one can visualize quantum fluctuations of the scalar field in an inflationary universe as waves. They first moved in all possible directions and then froze on top of one another. Each frozen wave slightly increased the scalar field in some parts of the universe and decreased it in others.

Now consider those places of the universe where these newly frozen waves persistently increased the scalar field. Such regions are extremely rare, but still they do exist. And they can be extremely important. Those rare domains of the universe where the field jumps high enough begin exponentially expanding with ever increasing speed. The higher the scalar field jumps, the faster the universe expands. Very soon those rare domains will acquire a much greater volume than other domains.

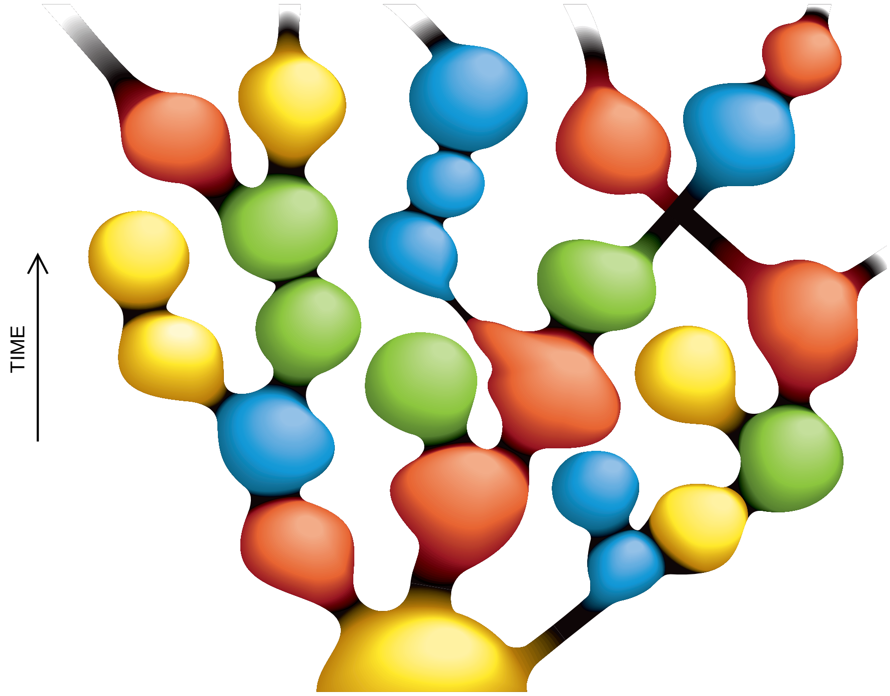
From this theory it follows that if the universe contains at least one inflationary domain of a sufficiently large size, it begins unceasingly producing new inflationary domains. Inflation in each particular point may end quickly, but many other places will continue to expand. The total volume of all these domains will grow without end. In essence, one inflationary universe sprouts other inflationary bubbles, which in turn produce other inflationary bubbles [*see illustration on opposite page*].

This process, which I have called eternal inflation, keeps going as a chain reaction, producing a fractallike pattern of universes. In this scenario the universe as a whole is immortal. Each particular part of the universe may stem from a singularity somewhere in the past, and it may end up in a singularity somewhere in the future. There is, however, no end for the evolution of the entire universe.

The situation with the very beginning is less certain. There is a chance that all parts of the universe were created simultaneously in an initial, big bang singularity. The necessity of this assumption, however, is no longer obvious. Furthermore, the total number of inflationary bubbles on our "cosmic tree" grows exponentially in time. Therefore, most bubbles (including our own part of the universe) grow indefinitely far away from the trunk of this tree. Although this scenario makes the existence of the initial big bang almost irrelevant, for all practical purposes, one can consider the moment of formation of each inflationary bubble as a new "big bang." From this perspective, inflation is not a part of the big bang theory, as we thought 15 years ago. On the contrary, the big bang is a part of the inflationary model.

In thinking about the process of self-reproduction of the universe, one cannot avoid drawing analogies, however superficial they may be. One may wonder, Is not this process similar to what happens with all of us? Some time ago we were born. Eventually we will die, and the entire world of our thoughts, feelings and memories will disappear. But there were those who lived before us, there will be those who will live after, and humanity as a whole, if it is clever enough, may live for a long time.

Inflationary theory suggests that a similar process may occur with the universe. One can draw some optimism from knowing that even if our civilization dies, there will be other places in the universe where life will emerge again and again, in all its possible forms.




**SELF-REPRODUCING COSMOS appears as an extended branching of inflationary bubbles. Changes in color represent "mutations" in the laws of physics from parent universes. The properties of space in each bubble do not depend on the time when the bubble formed. In this sense, the universe as a whole may be stationary, even though the interior of each bubble is described by the big bang theory.**

Could matters become even more curious? The answer is yes. Until now, we have considered the simplest inflationary model with only one scalar field, which has only one minimum of its potential energy. Meanwhile realistic models of elementary particles propound many kinds of scalar fields. For example, in the unified theories of weak, strong and electromagnetic interactions, at least two other scalar fields exist. The potential energy of these scalar fields may have several different minima. This condition means that the same theory may have different "vacuum states," corresponding to different types of symmetry breaking between fundamental interactions and, as a result, to different laws of low-energy physics. (Interactions of particles at extremely large energies do not depend on symmetry breaking.)

Such complexities in the scalar field mean that after inflation the universe may become divided into exponentially large domains that have different laws of low-energy physics. Note that this division occurs even if the entire universe originally began in the same state, corresponding to one particular minimum of potential energy. Indeed, large quantum fluctuations can cause scalar fields to jump out of their minima. That is, they jiggle some of the balls out of their bowls and into other ones. Each bowl corresponds to alternative laws of particle interactions. In some inflationary models, quantum fluctuations are so strong that even the number of dimensions of space and time can change.

If this model is correct, then physics alone cannot provide a complete explanation for all properties of our allotment of the universe. The same physical theory may yield large parts of the universe that have diverse properties. According to this scenario, we find ourselves inside a four-dimensional domain with our kind of physical laws, not because domains with different dimensionality and with alternative properties are impossible or improbable but simply because our kind of life cannot exist in other domains.

Does this mean that understanding all the properties of our region of the universe will require, besides a knowledge of physics, a deep investigation of our own nature, perhaps even including the nature of our consciousness? This conclusion would certainly be one of the most unexpected that one could draw from the recent developments in inflationary cosmology.

The evolution of inflationary theory has given rise to a completely new cosmological paradigm, which differs considerably from the old big bang theory and even from the first versions of the inflationary scenario. In it the universe appears to be both chaotic and homogeneous, expanding and stationary. Our cosmic home grows, fluctuates and eternally reproduces itself in all possible forms, as if adjusting itself for all possible types of life that it can support.

Some parts of the new theory, we hope, will stay with us for years to come. Many others will have to be considerably modified to fit with new observational data and with the ever changing theory of elementary particles. It seems, however, that the past 15 years of development of cosmology have irreversibly changed our understanding of the structure and fate of our universe and of our own place in it.

FURTHER READING

PARTICLE PHYSICS AND INFLATIONARY COSMOLOGY. Andrei Linde in *Physics Today,* Vol. 40, No. 9, pages 61–68; September 1987.

THE FRACTAL DIMENSION OF THE INFLATIONARY UNIVERSE. M. Aryal and A. Vilenkin in *Physics Letters B,* Vol. 199, No. 3, pages 351–357; December 24, 1987.

INFLATION AND QUANTUM COSMOLOGY. Andrei Linde. Academic Press, 1990.

PARTICLE PHYSICS AND INFLATIONARY COSMOLOGY. Andrei Linde. Harwood Academic Publishers, 1990.

FROM THE BIG BANG THEORY TO THE THEORY OF A STATIONARY UNIVERSE. A. Linde, D. Linde and A. Mezhlumian in *Physical Review D,* Vol. 49, No. 4, pages 1783–1826; February 1994.

# The Genetics of Flower Development

*Flower cells learn which organs to become from genes that convey positional information. A model based on just half a dozen such genes can predict how mutations will affect floral structure*

by Elliot M. Meyerowitz

All flowers are assemblages of several distinct types of organs. They may have sepals (modified leaflike structures), petals, stamens (which make pollen) and carpels (which are subunits of a plant's ovaries). In almost all wild flowers that have a full complement of these organs, their order from the periphery to the center of the flower is sepals-petals-stamens-carpels. Individual species and varieties, however, differ in their characteristic shapes and arrangements of these components.

Because the order of floral organs in a plant is one of its inherited traits, the information describing that pattern must be carried in its genetic material. Yet how can the one-dimensional order of base pairs in chromosomal DNA, the repository of inherited information, encode the three-dimensional structure of a flower—or of any other organism, for that matter?

The mechanisms of developmental pattern formation are still only partly understood. It is clear, however, that the relative position of a cell within an organism is what instructs that cell about what kinds of structures it and its progeny should form. This positional information is conveyed through combinations or gradients of regulatory proteins that are present in specific regions of the organism.

My colleagues and I have chosen to study the genesis of pattern in flowers because of its simplicity. Flowers generally consist of only the four easily distinguished types of organs, each of which is composed of a small number of cell types, and plants can be maintained and manipulated easily in the laboratory. We have taken a genetic approach to the problem: that is, we alter the DNA of the plants and observe the effects on the final structure of the flowers, then draw inferences about the mechanisms at work. Those inferences can be tested by both genetic and biochemical experiments, some of which involve designing and producing new types of flowers.

The ability to reorder and transform flowers according to our desires cannot be considered trivial in a world where most human food consists of flower parts or products, such as grains and fruits. Someday, when understanding of pattern development in flowers is more advanced, it should be possible to create plants that produce more of those structures that are most precious to us. Our laboratory has already taken some steps in that direction.

Even before my colleagues and I started our experiments on floral pattern formation, we knew that a genetic approach was bound to succeed. Breeders have genetically altered many horticultural varieties of plants to display abnormal orders or numbers of organs in their flowers. Indeed, most prized ornamental varieties of flowers are developmental monstrosities in which genetic mutations have disrupted the normal processes of pattern formation. For instance, wild roses have only five petals, whereas the hybrid tea roses in my backyard have 35 or 40. Wild camellias have stamens and carpels (as they must, to produce seeds). In my front yard, however, is a horticultural camellia (Pink Perfection) that has neither—instead it has many extra petals in the positions that stamens and carpels would normally occupy.

From such genetic variants one can infer that alterations in the DNA sequence of plants can change the usual organ order of flowers in consistent ways. We therefore knew we would learn much about the genetics of pattern formation in flowers by looking at mutant varieties. Yet genetic analysis in roses and camellias is difficult because rigorous studies need large numbers of them. The lengthy generations and large size of those plants make them too expensive for a basic research laboratory.

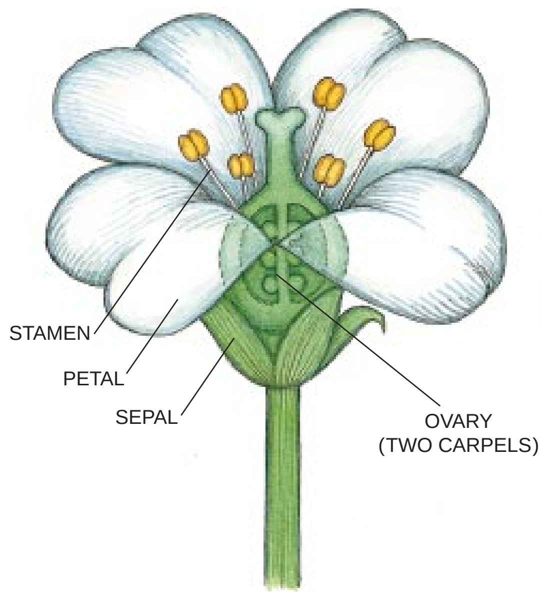
We thus began our studies by choosing as an experimental subject a rapidly breeding, tiny weed in the mustard family called *Arabidopsis thaliana,* or mouse ear cress. It grows well indoors under ordinary fluorescent lights and is small enough that several thousand

**TINY FLOWERS of *Arabidopsis thaliana,* or mouse ear cress (*right*), are excellent subjects for the genetic study of floral organ development. Because these plants are small, thousands can be cultivated on a laboratory bench (*above*).**

ELLIOT M. MEYEROWITZ, professor of biology at the California Institute of Technology, studies the cellular and molecular mechanisms of pattern formation in flower development. Meyerowitz graduated summa cum laude in biology from Columbia College in 1973. In his doctoral research at Yale University and postdoctoral work at Stanford University, he used *Drosophila* fruit flies to study development in animals. After joining the Caltech faculty in 1980, however, he became fascinated by plant development and refocused his research interests on the flowering plant *Arabidopsis.*

plants can grow to maturity on a tabletop. In much of Europe and the U.S., it is a frequently encountered meadow or lawn plant, with sprays of small, white flowers that mature as elongated fruits filled with tiny seeds. Despite its size, *Arabidopsis* is in all fundamental aspects a typical flowering plant.

Normal *Arabidopsis* flowers consist of four concentric regions, or whorls, each occupied by a different type of floral organ. At the periphery are four green sepals, which surround a second whorl of four white petals. Within the petals are the third whorl, containing six stamens (four long, two short) and the innermost whorl, which is an ovary made of two fused carpels.




**FLORAL ORGANS are of four basic types: the leaflike sepals, the petals, the stamens (the male reproductive organs) and the carpels (the female reproductive organs). In many varieties of flowers, these organs are arranged in four concentric whorls.**

Each flower begins its development as a primordium, or outgrowth of undifferentiated cells near the tip of a flowering stem. Through a combination of cell division, cellular differentiation and cellular elongation—all occurring in precise positions—this small primordium develops into a flower. The question, therefore, is how the cells of the primordium learn their positions.

Our initial genetic experiments looked at mutant strains in which the order of floral organs was abnormal. We hoped to identify single genes whose protein products are essential for pattern formation: without them, undifferentiated cells in developing flowers will misinterpret their position and become normal organs in inappropriate locations. We obtained many such strains, both by applying the mutagenic chemical ethyl methane sulfonate to seeds and by soliciting *Arabidopsis* strains with abnormal flowers from our colleagues, such as Maarten Koornneef of the Agricultural University of Wageningen in the Netherlands.

Initially, all the mutant flowers we saw fell into three classes. One class, caused by mutations that inactivate the gene *APETALA2,* has carpels instead of sepals in the first whorl and stamens instead of petals in the second whorl. (By scientific convention, the names of *Arabidopsis* genes are written in all capital letters; plants that have mutations in those genes have the same names, written in lowercase.)

A second class is caused by mutations in either of the genes *APETALA3* or *PISTILLATA*. These flowers have sepals in their first and second whorls and carpels in their third and fourth.

In the third class of mutants, in which the gene *AGAMOUS* is inactive, the abnormal radial order of organs is sepals-petals-petals-sepals. These *agamous* mutants also have additional internal whorls of organs that repeat the pattern in the outer whorls. The presence of these extra parts suggests that the *AGAMOUS* gene product, in addition to establishing the identity of the organs in the third and fourth whorls, stops flower development after the fourth whorl has been established.

Each of these three classes of mutants is missing a different genetic activity essential to normal floral pattern formation. For simplicity, I will call the activity absent in *apetala2* mutants A, the activity missing in *apetala3* and *pistillata* mutants B, and the activity absent in *agamous* mutants C.

Even with just that little information, we were able to build a simple provisional hypothesis, or working model, about how these four genes might normally contribute to the specification of organ identity in developing flowers. The model has three parts. First, the protein encoded by each of these genes is active early in the region affected by mutations in its genes. The A activity gene *APETALA2* therefore acts in the first and second whorls, the B activity genes *APETALA3* and *PISTILLATA* act in the second and third whorls, and the C activity gene *AGAMOUS* acts in the third and fourth whorls.

Second, combinations of these gene products determine the identity of floral organs developing from the primordium. Any organ that forms in a region where only the A activity is present will develop into a sepal. Any organ forming where both the A and B activities are present becomes a petal. The combined presence of the B and C activities creates stamens. If only the C activity is present, the emerging organ is a carpel.

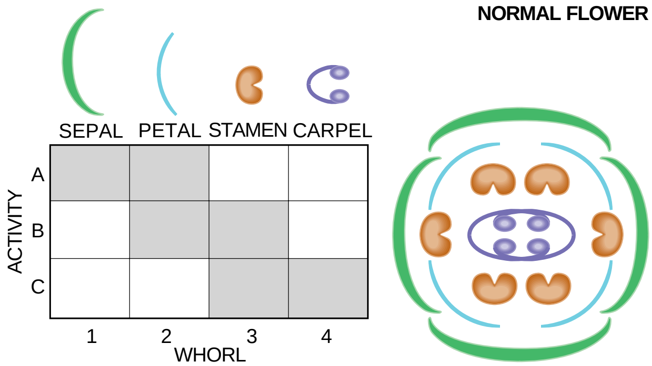
The third part of the model is that the A and C activities are mutually exclusive: any region where A is present will lack C activity, and vice versa. If A activity is eliminated (as in the *apetala2* mutants), C will be abnormally present in the first and second whorls. Conversely, if C is absent (as in *agamous* mutants), A will be abnormally present in the third and fourth whorls.

These three rules are sufficient to explain the flowers produced by the mutants. Of course, the fact that the model is consistent with the evidence is not a demonstration that it is correct—the model must be tested.

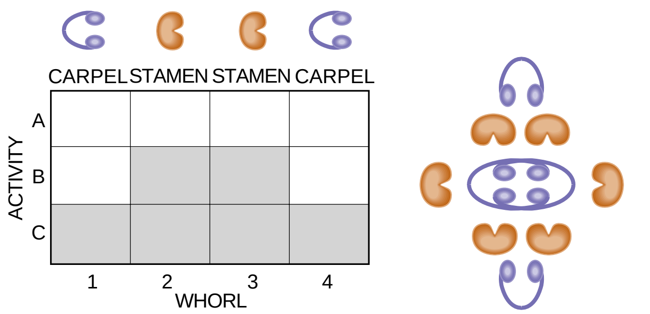
Simple genetic experiments provide one type of test. The model makes specific predictions about what should happen if we eliminate pairs of activities. If both the A and B activities are eliminated—by breeding plants with mutations in both the *APETALA2* and *APETALA3* genes, for example—the flower primordium should have only C activity (which will extend to the first and second whorls because A activity is absent). The model therefore predicts that all the floral organs should be carpels. When we performed the experiment, it confirmed this prediction. By similar rea-

**GENETIC STUDIES OF FLOWERS from normal and mutant varieties of *Arabidopsis* have produced a simple model that explains the pattern of organs. Combinations of three genetic activities—A, B and C—are sufficient to specify the identity of the organs in each whorl. Schematic diagrams of the flowers and charts of the genetic activities in each whorl are shown for several varieties of *Arabidopsis.* In normal flowers, A is active in whorls 1 and 2, B in whorls 2 and 3, and C in whorls 3 and 4. The A and C activities are complementary: if one is missing, the other manifests itself in all the whorls. Sepals grow from tissues in which only A is present; petals result from a combination of A and B. Stamens develop where both B and C are present, and carpels are produced where only C is active.**

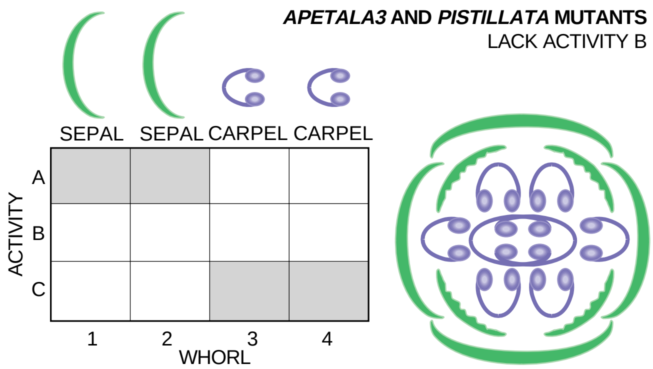
**NORMAL FLOWER**



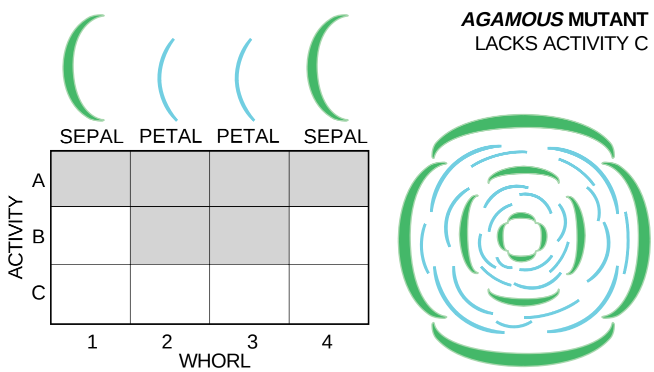

***APETALA2* MUTANT**
LACKS ACTIVITY A

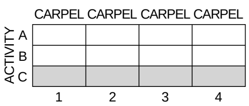



***APETALA3* AND *PISTILLATA* MUTANTS**
LACK ACTIVITY B

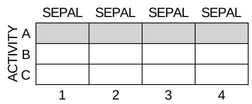



***AGAMOUS* MUTANT**
LACKS ACTIVITY C




***APETALA2 APETALA3*** **double mutants lack both A and B activities. The C activity expands into all the whorls, so all the organs are carpels.**




***APETALA3 AGAMOUS*** **double mutants lack both B and C activities. The A activity expands into all the whorls, so all the organs are sepals.**

soning, if both B and C are eliminated, as in the double mutant combination *apetala3 agamous,* all floral organs should develop as sepals. They do.

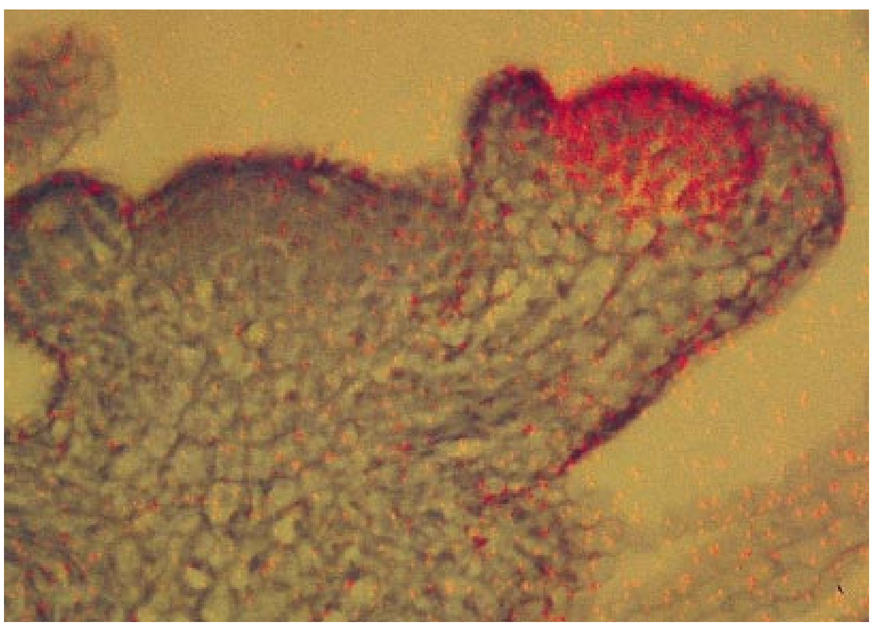
From the types of flowers we observed, we had no way of knowing in advance what types of organs would arise in mutant plants that had only the B activity. Nevertheless, the model did still tell us that the organs found in the first and fourth whorls should be of the same type and different from those found in the second and third whorls. When we bred *apetala2 agamous* double mutants (which do lack both the A and C activities), we found that they have leaves in the first and fourth whorls and organs intermediate between petals and stamens in whorls two and three.

Moreover, in triple mutant flowers that lack the activity of all three functional classes of genes, all the floral organs differentiate as leaves. These results further confirm the model and indicate that the A, B and C activities seem sufficient for specifying not only the differences among the floral organs but also between the floral organs and the vegetative organs, leaves.

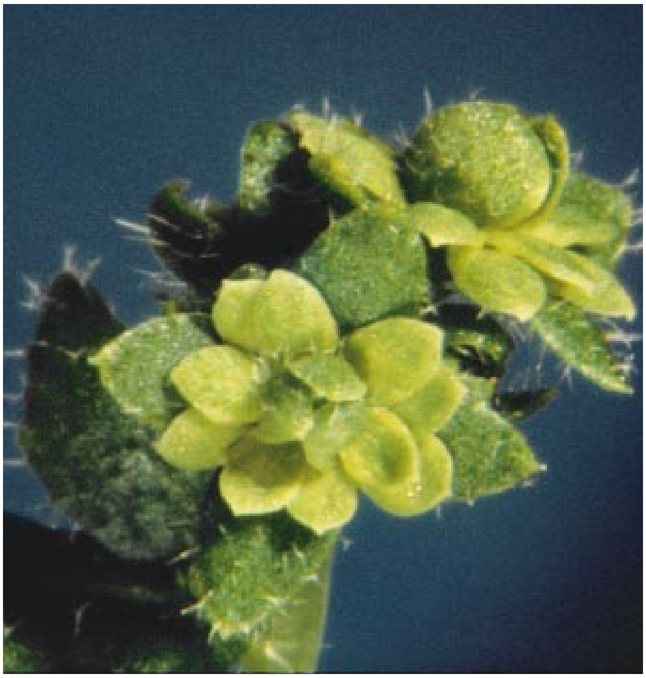
Further evidence of the validity of our radial pattern model came after we and others cloned some of the genes that specify the identity of flower organs. The availability of those clones allowed us to perform experiments based on the technique known as in situ hybridization. Using the clones as a template, we created “antisense” RNA molecules complementary to the RNA products of the genes. These antisense molecules contained small quantities of radioactive elements.

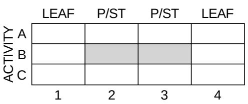
When applied to sections of floral tissue, the antisense probes bound to any complementary RNA from active genes in the cells. By laying a film emulsion over the tissue and exposing it to the probes’ radiation, we could find where the probes bound and therefore where the genes of interest were active.

The first gene we were able to clone was that for C activity, *AGAMOUS.* The in situ hybridization of *AGAMOUS* probes to sections of developing flow-

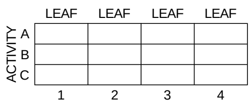


**IN SITU HYBRIDIZATION experiments confirm the predictions of the genetic model by showing where the organ-identity genes are active. Radioactive probe molecules bind to the RNA produced by specific active genes. At the left, probes for**





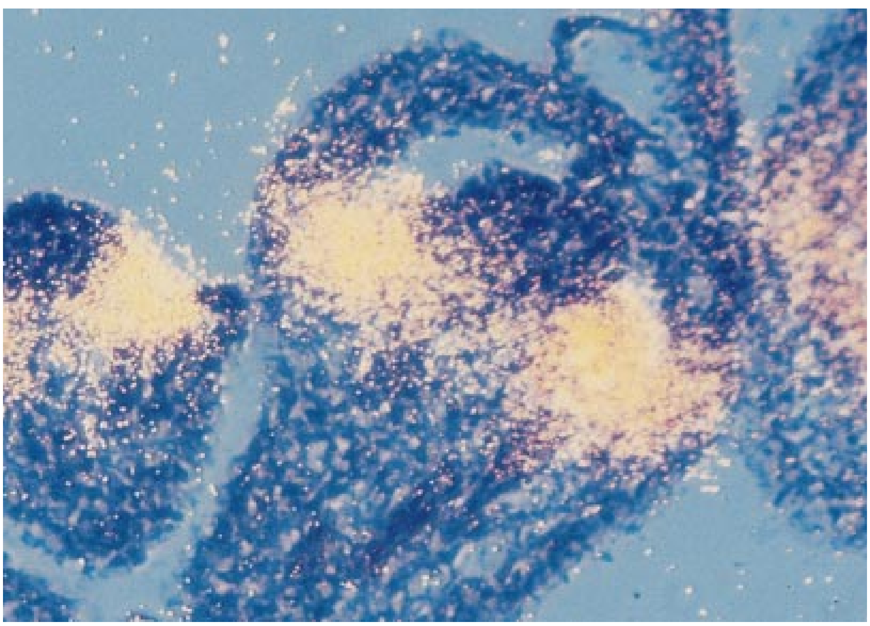

***APETALA2 AGAMOUS*** **double mutants have leaves in whorls 1 and 4 and organs intermediate between petals and stamens in whorls 2 and 3.**




**TRIPLE MUTANTS that lack all three categories of genetic activities have flowers that consist entirely of leaves arranged in whorls.**

ers verified a prediction of the model: RNA from *AGAMOUS* appears early in flower development and only in those regions of the primordium that later become the third and fourth whorls.

Our genetic model also states that when A activity is absent, the C activity should extend abnormally into the first and second whorls. Once again, in situ hybridization confirmed a prediction of the model. When we tested *apetala2* mutant flowers, we found *AGAMOUS* RNA in all four whorls of the developing flowers. Later, after cloning the *APETALA3* and *PISTILLATA* genes, we conducted similar probes for B activity. As anticipated, those genes were active in the second and third whorls.



***AGAMOUS*** **RNA (*red*) localize within the primordium of a normal flower that will later become parts of the stamens and carpels. At the right, probes for *APETALA3* RNA (*yellow*) bind to the regions that will differentiate as petals and stamens.**

Cloning these organ-identity genes made the in situ hybridization experiments possible, but it also brought another, more far-reaching benefit. By analyzing the DNA of the genes, we could begin to understand the proteins they produced and see how they acted at the subcellular level.

For instance, a sizable part of the protein that *AGAMOUS* encodes is strongly similar, in terms of its constituent amino acid sequence, to certain proteins found in yeast, humans and other organisms. This class of proteins is known to bind to DNA and to regulate the transcription of genes into RNA. In other words, these proteins act as molecular switches, turning other genes on or off by binding to specific DNA sequences found near them.

A. Nordheim and his co-workers at the Hanover Medical School in Germany have shown that the *AGAMOUS* protein does indeed bind to DNA. It is therefore highly likely that it, like other proteins of its class, serves a regulatory function.

We still do not know which genes the *AGAMOUS* protein binds and regulates in the *Arabidopsis* flower, but we can postulate that it controls the activity of a series of downstream genes. Those genes may themselves be regulators of other genes, or they may make the final protein products that differentiate one floral cell type from another. (Incidentally, a human relative of *AGAMOUS* makes a protein called serum response factor that regulates the oncogene *c-fos,* which may be involved in human cancers. This remarkable parallel between human and plant gene control underlines the evolutionary unity of life.)

While we were cloning *AGAMOUS* in *Arabidopsis,* Hans Sommer, Zsuzsanna Schwarz-Sommer and Heinz Saedler and their colleagues at the Max Planck Institute for Plant Breeding in Cologne were busy cloning a gene named *DEFICIENS* in snapdragons. *DEFICIENS* provides the equivalent of B activity in those plants (which are only distantly related to *Arabidopsis*). Mutants in *DEFICIENS* have been known for many decades. Notably, *DEFICIENS* is a member of the same class of DNA-binding proteins as *AGAMOUS,* which suggested that other floral pattern genes might be as well.

The German group has subsequently cloned a series of snapdragon regulatory genes. So far the snapdragon genes correspond one for one with the *Arabidopsis* genes, both in the related protein sequences that they encode and in the types of pattern abnormalities that result from their loss. We and our German colleagues have also found in each plant many members of this gene family that do not correlate with any known floral mutants but that are nonetheless active in parts of developing flowers. This gene family may hold additional clues to flower development.

We thus arrive at a picture of a small set of regulatory genes that determine organ identity by becoming active in a highly specific pattern and controlling other genes through a regulatory cascade. Yet how is it that these regulatory genes know to come on in the right places early in flower development?

The answer is that these genes are themselves only in the middle, not at the beginning, of the regulatory cascade. The cells that express organ-identity genes must be different from their neighbors even before those genes turn on. The concentric regions that will later represent domains of organ fate must already be established. The organ-identity genes therefore represent an essential step in the intracellular *interpretation* of position, although they are not themselves involved in the initial *establishment* of positional differences.

***SUPERMAN* AND *LEAFY* genes act early in flower development and influence which cells will express the A, B and C activities. *SUPERMAN* regulates the B activity genes. Thus, *superman* mutants (*top*) have stamens in whorl 4 and a corresponding reduction in carpels. The *superman agamous* double mutants (*second from top*) have petals in whorls 2, 3 and 4. *Superman agamous apetala2* mutants (*third from top*) have leaves surrounding organs intermediate between stamens and petals. The *leafy* mutant flowers (*bottom*) have only sepallike organs in a spiral array like that of the leaves on the stems.**

Another set of regulatory genes, acting even earlier in development, seems to establish the positional domains within flowers. One of these genes is called *SUPERMAN.* (It has also been studied under the name *FLO10* by George W. Haughn, now at the University of Saskatchewan, and his colleagues.) In the flowers of plants with defects in this gene, the fourth whorl does not contain carpels, it contains extra stamens (which is why the mutant was jokingly named *superman*). According to the genetic model, this pattern implies that the B activity is present in the fourth whorl of these mutant flowers, which in turn suggests that the *SUPERMAN* gene normally prevents B activity in that whorl.

If so, then we can hypothesize that flowers with mutations in both their *SUPERMAN* and *APETALA3* genes will look like *apetala3* mutants: if *APETALA3* is nonfunctional, it does not matter whether it is regulated improperly. Breeding experiments have confirmed that prediction. Similarly, they have borne out two others: that mutants lacking both *SUPERMAN* and *AGAMOUS* activities will have sepals surrounding three whorls of petals and that triple mutants of the *superman agamous apetala2* variety will have leaves surrounding three whorls of organs intermediate between stamens and petals.

In situ hybridization experiments also support the idea that *SUPERMAN* regulates the B activity genes. RNA from *APETALA3* and *PISTILLATA* appears in the fourth whorl of *superman* mutants, as well as in the usual positions in the second and third whorls.

The *SUPERMAN* gene's function explains why B activity genes do not turn on in the fourth whorl. Still, it does not explain why those genes do turn on in the second and third whorls at the right time in development. The answer seems to involve a gene that acts even earlier, *LEAFY.* Flowers of plants lacking an active *LEAFY* gene have neither petals nor stamens. In their place is a series of organs similar to sepals. But these organs do not grow in concentric rings, as the normal second and third whorl organs do. Instead they appear in a spiral array—the same array that characterizes the position of leaves on the stem of an *Arabidopsis* plant.

The *LEAFY* gene ordinarily seems to tell the cells of the second and third whorl that they are in flowers. Without this information, the cells divide in a pattern usually seen outside flowers, on stems. Because these organs resemble sepals, one can infer that A activity is present in them and that *LEAFY* is not required for the initial activation of A genes. Because the center of *leafy* mutant flowers can have carpels, *LEAFY* does not seem to be required for the initiation of C activity, either.

But *LEAFY* does appear to be crucial to the initial expression of B activity, because without it, petals and stamens do not form. It must also be the key to even more fundamental functions that specify the relative position and the radial order of organs.

The results of genetic and molecular tests are consistent with that conclusion. Mutants that lack function in both their *LEAFY* and *APETALA3* genes have the same appearance as those lacking just *LEAFY:* none have B activity. Also, in *leafy* mutants, the amount of RNA from *APETALA3* is much lower than in normal flowers. The protein product of *LEAFY* is therefore probably an activator of *APETALA3.*

Experiments to find when the *LEAFY* gene becomes active have also had results consistent with this addition to our original genetic model. Enrico S. Coen of the John Innes Institute in Norwich, England, and his co-workers had cloned a gene with a function related to that of *LEAFY* from a snapdragon. Using that gene as a probe, we were then able to clone *LEAFY* from *Arabidopsis.* On the basis of RNA studies, *LEAFY* and the related snapdragon gene seem to be the earliest genes known to be active in flower development. Their RNA products appear when the initial clusters of cells that become flower primordia first form, long before the organ-identity genes become active.

Mutations in other genes indicate that they too serve early functions in flower development, and no doubt additional organ-identity genes will be discovered as well. We nonetheless have what appears to be at least an outline of the genetic hierarchies that control flower development in *Arabidopsis.* As the groups at the John Innes Institute and the Max Planck Institute have revealed, the *Arabidopsis* genes are comparable to those known in snapdragons. Related genes direct pattern formation in the development of other flowers, according to recent research. The genetic model of flower development described here may therefore hold true for many flowers that have a radial pattern—perhaps even all of them.

**GRAINS, FRUITS and other flower products make up most of the food consumed by humans. The ability to manipulate the development and form of flowers may someday help create more productive crops.**

Equipped with a predictive genetic model and a knowledge of some of the genes that specify floral organs, researchers now have a considerable degree of control over the development of flowers. It is possible to take the cloned genes, to alter either the proteins they encode or their patterns of action, and to reintroduce them to plants to alter flower development in desirable ways.

A collaborative project between my former postdoctoral fellow Martin F. Yanofsky, now at the University of California at San Diego, and our laboratory may be a forerunner of the type of experiment that could someday be routine. Working with clones of the *AGAMOUS* gene, we altered the adjacent control sequences that regulate when and where it will be active. We replaced these sequences with ones from a different gene expressed in all cells. We then inserted this recombinant gene into tobacco plants. As we expected, the resulting plants had carpels (and developing ovules) in the positions usually occupied by sepals and had stamens where petals would ordinarily be.

There is no direct use for tobacco plants of this kind. Nevertheless, they are living testaments to the power that our growing understanding of the molecular basis of flower development already grants us over the structure and function of some plants.

FURTHER READING

*ARABIDOPSIS THALIANA* AND PLANT MOLECULAR GENETICS. E. M. Meyerowitz and R. E. Pruitt in *Science,* Vol. 229, pages 1214–1218; September 20, 1985.

ABNORMAL FLOWERS AND PATTERN FORMATION IN FLORAL DEVELOPMENT. E. M. Meyerowitz, D. R. Smyth and J. L. Bowman in *Development,* Vol. 106, No. 2, pages 209–217; June 1989.

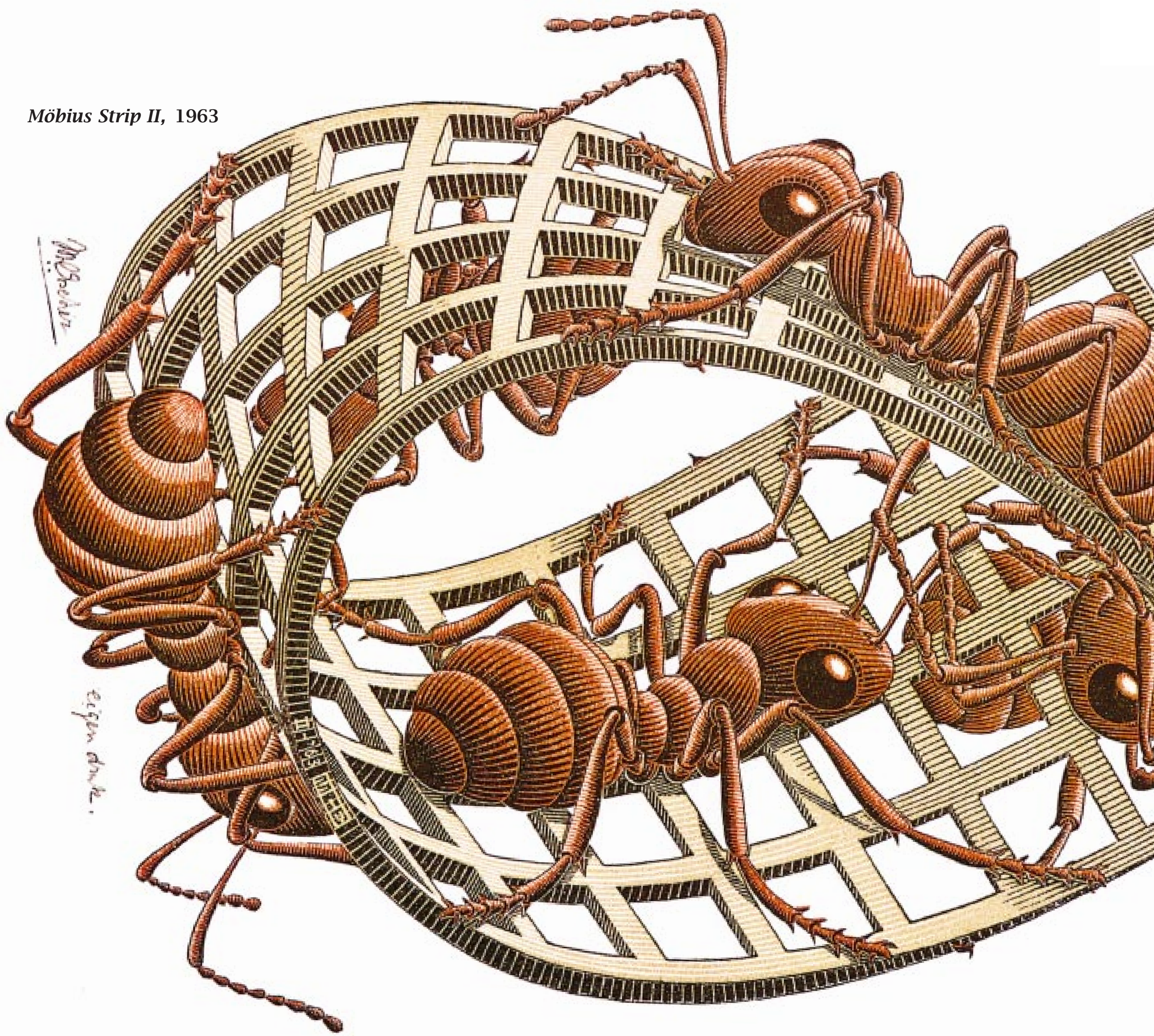
GENETIC CONTROL OF FLOWER DEVELOPMENT BY HOMEOTIC GENES IN *ANTIRRHINUM MAJUS.* Z. Schwarz-Sommer, P. Huijser, W. Nacken, H. Saedler and H. Sommer in *Science,* Vol. 250, pages 931–936; November 16, 1990.

THE WAR OF THE WHORLS: GENETIC INTERACTIONS CONTROLLING FLOWER DEVELOPMENT. Enrico S. Coen and E. M. Meyerowitz in *Nature,* Vol. 353, No. 6339, pages 31–37; September 5, 1991.

# Escher’s Metaphors

*The prints and drawings of M.C. Escher give expression to abstract concepts of mathematics and science*

by Doris Schattschneider



***Möbius Strip II,*** **1963**

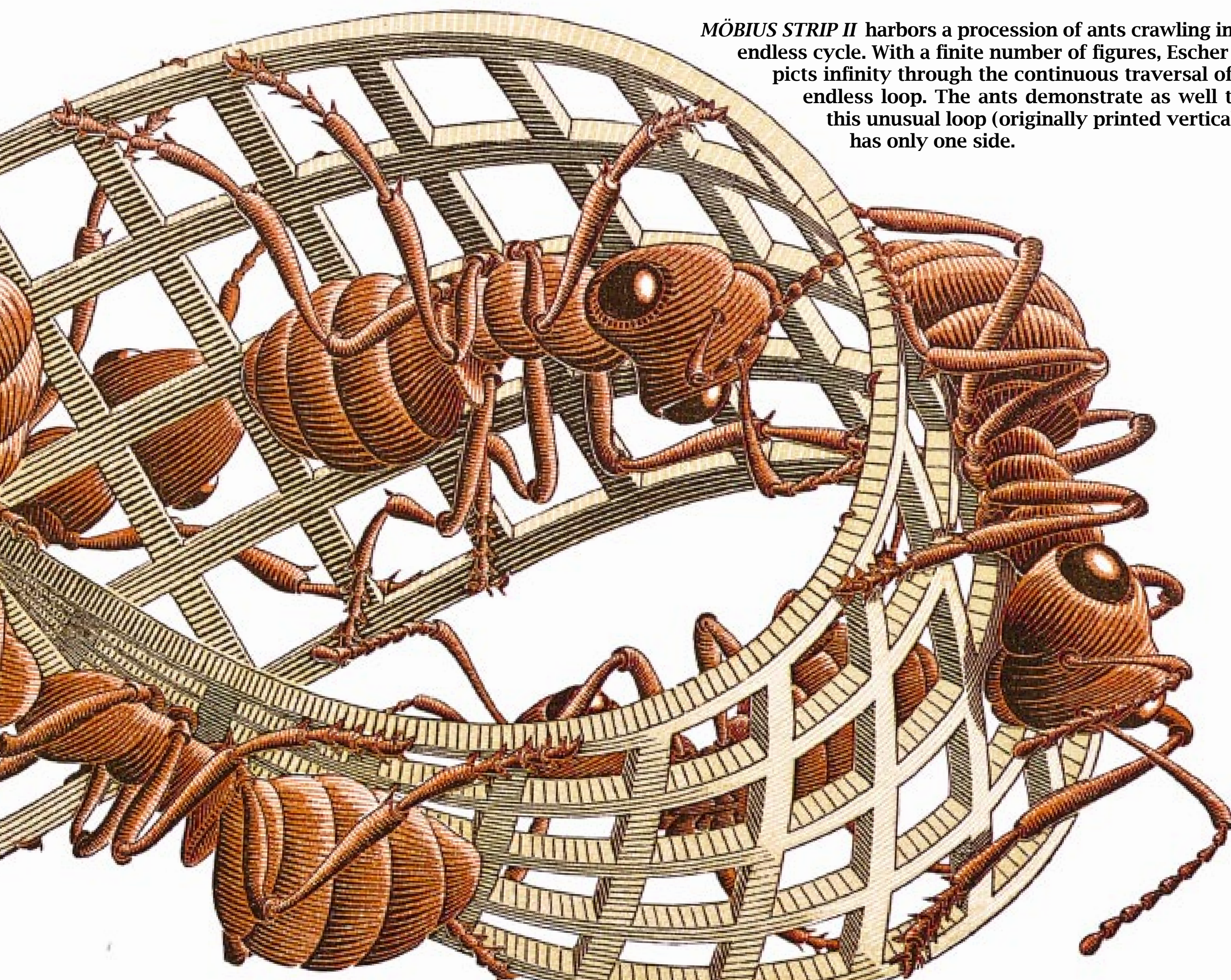
Throughout his life Maurits Cornelis Escher (he used only M. C.) remarked on his inability to understand mathematics, declaring himself “absolutely innocent of training or knowledge in the exact sciences.” Yet even as a child, Escher was intrigued by order and symmetry. The fascination later led him to study patterns of tiles at the Alhambra in Granada, to look at geometric drawings in mathematical papers (with the advice of his geologist brother) and ultimately to pursue his own unique ideas for tiling a plane.

Escher’s attention to the coloring of his drawings of interlocked tiles anticipated the later work of mathematicians and crystallographers in the field of color symmetry. His works are now commonly used to illustrate these concepts. His exhibit in conjunction with the 1954 International Congress of Mathematicians in Amsterdam and the publication of his first book (*The Graphic Work of M. C. Escher*) in 1959 struck a chord with mathematicians and scientists that still resonates strongly. He wrote that a main impetus for his work was “a keen interest in the geometric laws contained by nature around us.” In expressing his ideas in graphic works, he provided arresting visual metaphors for fundamental ideas in science.

Escher was born in 1898 in the town of Leeuwarden, Holland. The youngest son of a civil engineer, he grew up with four brothers in Arnhem. Although three of his brothers pursued science or engineering, Escher was a poor mathematics student. With the encour-

***Self-Portrait,* 1943**

**M. C. ESCHER views himself in a mirror in this “scratch” drawing with lithographic ink.**



***MÖBIUS STRIP II* harbors a procession of ants crawling in an endless cycle. With a finite number of figures, Escher depicts infinity through the continuous traversal of an endless loop. The ants demonstrate as well that this unusual loop (originally printed vertically) has only one side.**

agement of his high school art teacher, he became interested in graphic arts, first making linoleum cuts.

In 1919 he entered the School for Architecture and Decorative Arts in Haarlem, intending to study architecture. But when he showed his work to Samuel Jessurun de Mesquita, who taught graphic arts there, he was invited to concentrate in that field. De Mesquita had a profound influence on Escher, both as a teacher (particularly of woodcut techniques) and later as a friend and fellow artist.

After finishing his studies in Haarlem, Escher settled in Rome and made many extensive sketching tours, mostly in southern Italy. His eyes discerned striking visual effects in the ordinary—architectural details of monumental buildings from unusual vantage points, light and shadow cast by warrens of staircases in tiny villages, clusters of houses clinging to mountain slopes that plunged to distant valleys and, at the opposite scale, tiny details of nature as if viewed through a magnifying glass. In his studio, he would transform the sketches into woodcuts and lithographs.

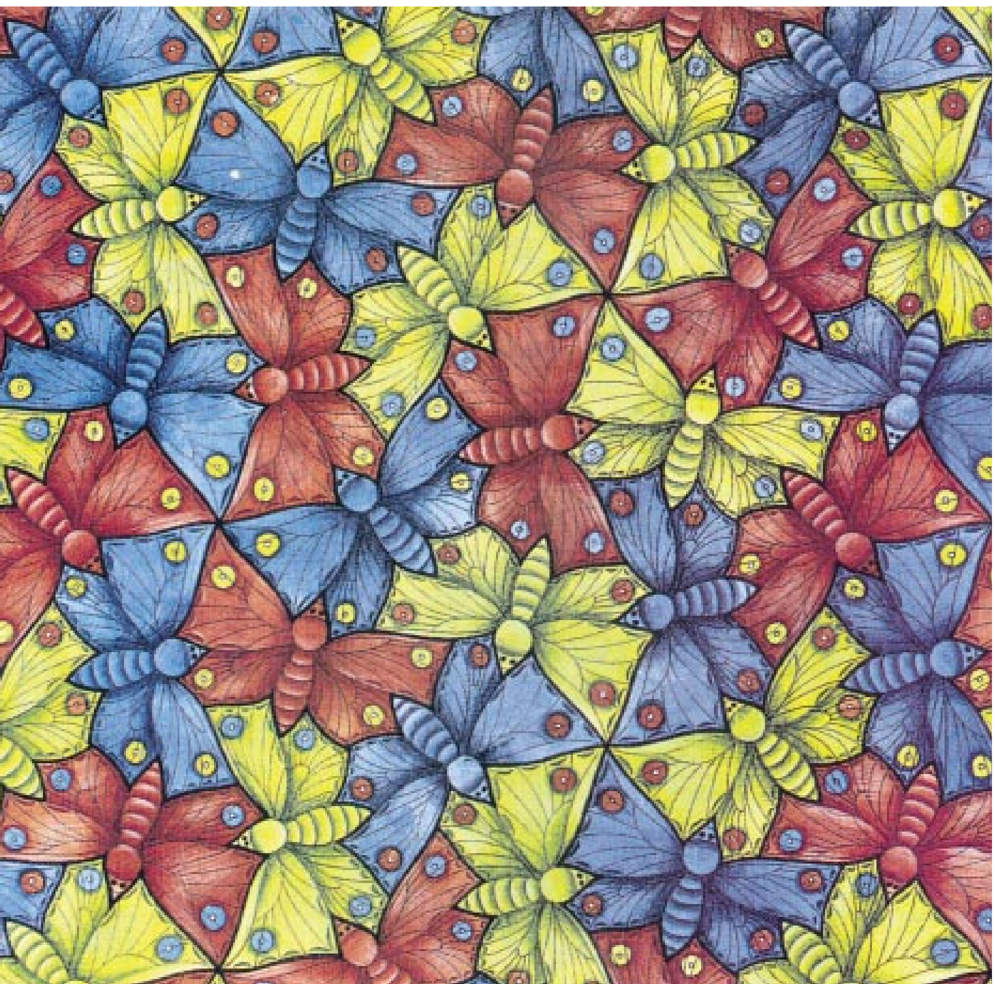
In 1935 the political situation became unendurable, and with his wife and young sons, Escher left Italy forever. After two years in Switzerland and then three years in Uccle, near Brussels, they settled permanently in Baarn, Holland. These years also brought an abrupt turn in Escher's work. Almost all of it from this time on would draw its inspiration not from what his eyes observed but rather from his mind's eye. He sought to give visual expression to concepts and to portray the ambiguities of human observation and understanding. In doing so, he often found himself in a world governed by mathematics.

Escher was fascinated, almost obsessed, with the concept of the "regular division of the plane." In his lifetime, he produced more than 150 color drawings that testified to his ingenuity in creating figures that crawled, swam and soared, yet filled the plane with their clones. These drawings illustrate symmetries of many different kinds. But for Escher, division of the plane was also a means of capturing infinity. Although a tiling such as the one using butterflies [*see illustration below*] can in principle be continued indefinitely, thus giving a suggestion of infinity, Escher was challenged to contain infinity within the confines of a single page.

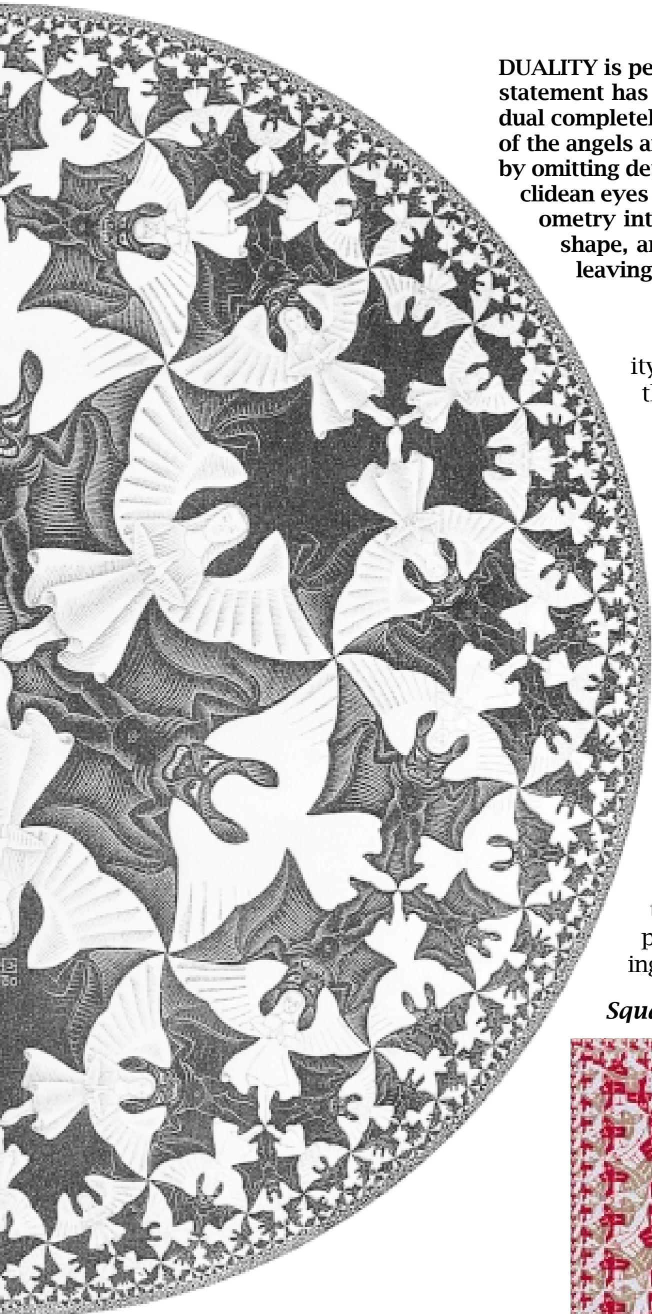
"Anyone who plunges into infin-

***Circle Limit IV,*** **1960**

***Triangle System I $B_3$ Type 2,*** **1948**



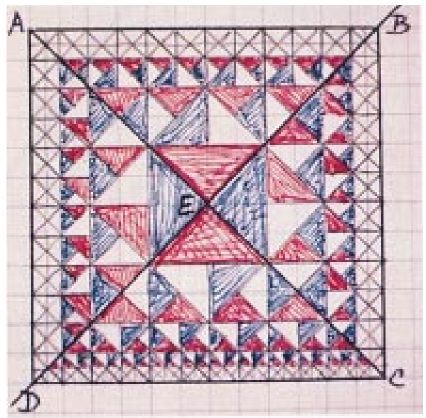
**SYMMETRY is a structural concept that shapes many mathematical and physical models. In Escher's drawing the butterflies seem to fill the page randomly, yet each one is precisely placed and surrounded in exactly the same way. Always six (in alternating colors) swirl about a point where left front wingtips meet; always three (in different colors) spin about a point where right back wings touch; and always pairs (of different colors) line up the edges of their right front wings. Along with rotational symmetry, the drawing has translational symmetry based on a triangular grid. The pattern can continue forever in all directions and so provides an implicit metaphor of infinity. Escher's attention to coloring anticipated discoveries by mathematicians in the field of color symmetry.**



**DUALITY is perhaps the most prevalent theme in Escher's later prints. In mathematics, a statement has a negation, and a set has a complement; in every case, the object and its dual completely define each other. In *Circle Limit IV,* there are no outlines. The contours of the angels and devils define one another. Either is figure or ground (Escher reminds us by omitting detail in half the figures). In this hyperbolic tiling the figures appear to our Euclidean eyes to become more distorted as they diminish in size. Yet measured by the geometry intrinsic to the world of the print, every angel is exactly the same size and shape, and so is every devil. An infinite number of copies repeat forever, never leaving the confines of the circle.**

ity, in both time and space, farther and farther without stopping, needs fixed points, mileposts as he flashes by, for otherwise his movement is indistinguishable from standing still," Escher wrote. "He must mark off his universe into units of a certain length, into compartments which repeat one another in endless succession."

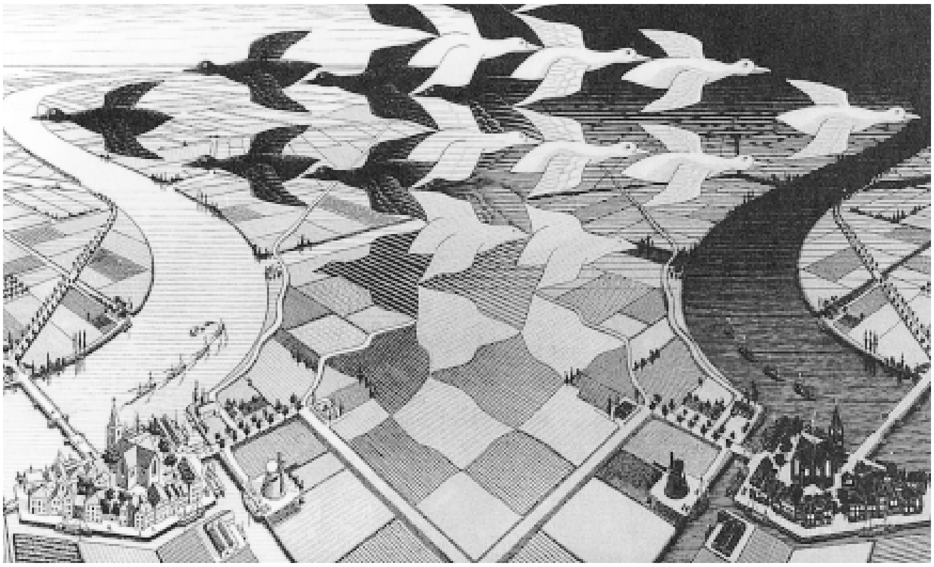
After completing several prints in which figures endlessly diminish in size as they approach a central vanishing point [*see* Whirlpools *on page 71*], Escher sought a device to portray progressive reduction in the opposite direction. He wanted figures that repeated forever, always approaching—yet never reaching—an encircling boundary. In 1957 the mathematician H.S.M. Coxeter sent Escher a reprint of a journal article in which he illustrated planar symmetry with some of Escher's drawings. There Escher found a figure that gave him "quite a shock"—a hyperbolic tessellation of tri-




***Square Limit,* 1964**

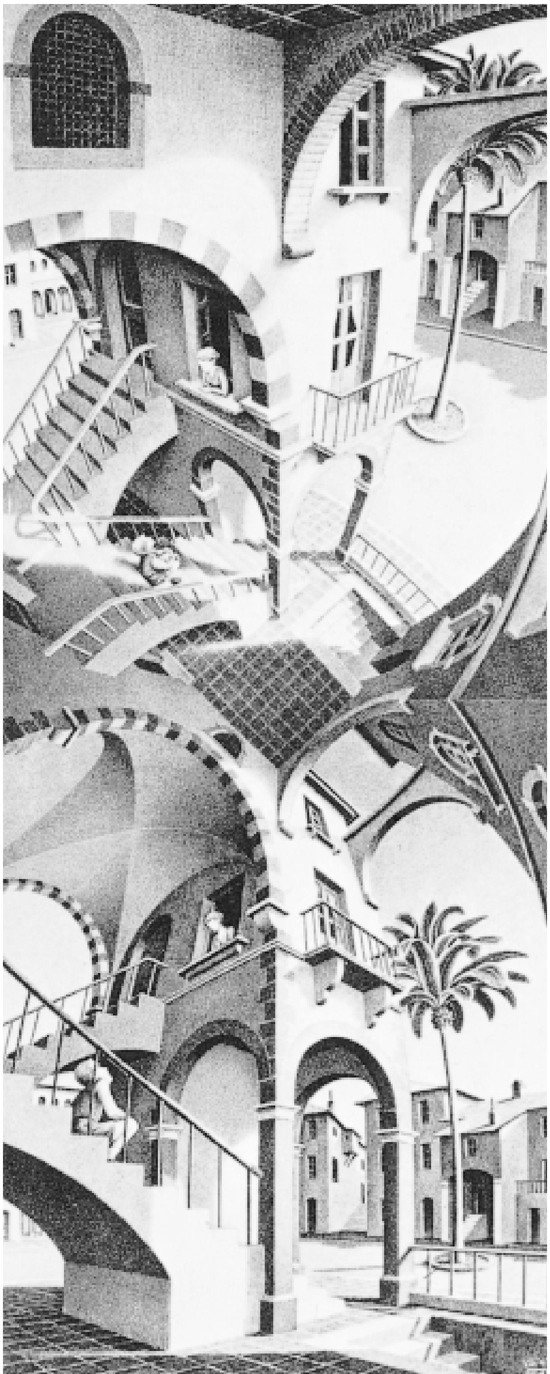
**SELF-SIMILARITY is illustrated in the print *Square Limit,* constructed using a recursive scheme of Escher's own invention. A set of directions that is applied to an object to produce new objects, then applied to the new objects and so on, ad infinitum, is called a recursive algorithm. The end product is self-similar when all the final objects are the same as the original, except for changes of scale, orientation or position. A sketch (*top right*) sent by Escher to the mathematician H.S.M. Coxeter to explain the print shows that the underlying grid involves a recursive splitting of isosceles triangles. In executing the print, Escher carved the woodblock only for a triangle having its apex at the center of the square and its base as one side of the square—and printed the block four times.**

*Day and Night,* 1938



**DIMENSION is that concept which clearly separates point, line, plane and space. To illustrate the ambiguities in the perception of dimension, Escher exploited the printed page—which always must fool the viewer when it depicts a three-dimensional scene. In *Day and Night,* the flat checkerboard of farmland at the bottom of the print metamorphoses into two flocks of geese. The print also illustrates the concept of topological change, in which a figure is deformed without being cut or pierced. Reflection and duality are present as well: black geese fly over a sunlit village, whereas white ones wing over a night view of a mirror image of the same scene.**

*High and Low,* 1947



**RELATIVITY states that what an observer sees is influenced by context and vantage point. In the lithograph *High and Low,* Escher presents two different views of the same scene. In the lower half the viewer is on the patio; in the upper half the viewer is looking down. Now draw back from the print: Is that tiled diamond at the center of the print a floor or a ceiling? Escher uses it for both in order to marry the two views. It is impossible to see the entire print in a logical way. The scene also illustrates how pasting local views together to form a global whole can lead to contradictions.**

angles that showed exactly the effect he sought. From a careful study of the diagram, Escher discerned the rules of tiling in which circular arcs meet the edge of an encompassing circle at right angles. During the next three years, he produced four different prints based on this type of grid, of which *Circle Limit IV* [*see top illustration on preceding two pages*] was the last.

Four years later Escher devised his own solution to the problem of infinity within a rectangle [*see bottom illustration on preceding page*]. His recursive algorithm—a set of directions repeatedly applied to an object—results in a self-similar pattern in which each element is related to another by a change of scale. Escher sent Coxeter a sketch of the underlying grid, apologizing: "I fear that the subject won't be very interesting, seen from your mathematical point of view, because it's really simple as a flat filling. None the less it was a headaching job to find an adequate method to realise the subject in the simplest possible way." In a lecture a few summers ago mathematician William P. Thurston, director of the Mathematical Sciences Research Institute at the University of California at Berkeley, illustrated the concept of self-similar tiling with just such a grid, unaware of Escher's earlier discovery.

Curiously, self-similar patterns provide examples of figures that

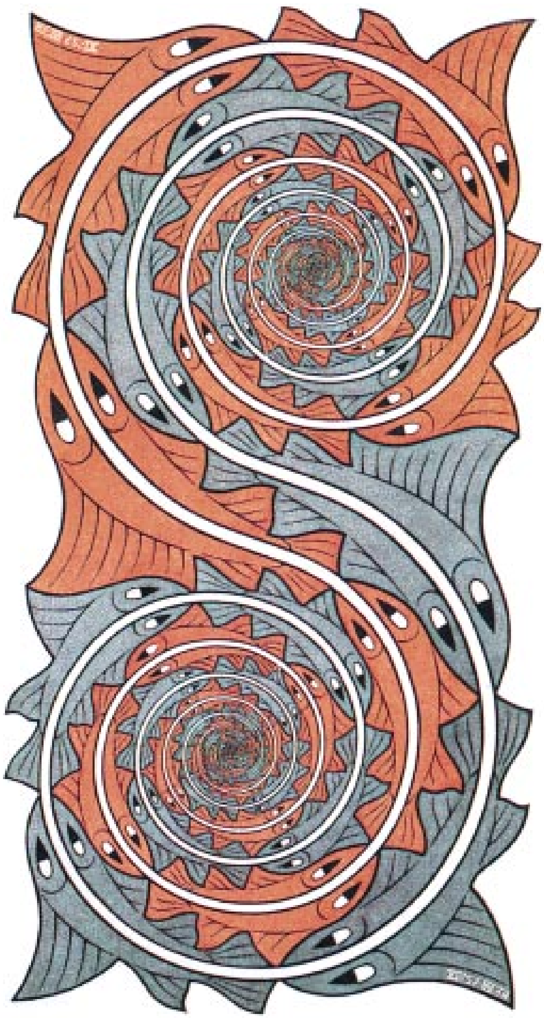
DORIS SCHATTSCHNEIDER, professor of mathematics at Moravian College in Bethlehem, Penn., received her Ph.D. from Yale University in 1966. Active as a teacher, writer and lecturer, the former editor of *Mathematics Magazine* has strong interests in both geometry and art. She is co-author (with Wallace Walker) of a popular work with geometric models, *M. C. Escher Kaleidocycles.* Her 1990 book, *Visions of Symmetry,* culminated a research project on the symmetry studies of M. C. Escher.

**REFLECTION allows phenomena to be observed that are too small, too far away or too obscure to be seen directly. *Puddle* directs our eyes to a woodland trail imprinted by boots and tires—yet in the puddle are also revealed silhouetted trees arching overhead against a moonlit sky. Escher reminds us of the unseen worlds below, behind and above our limited gaze.**

***Puddle*, 1952**

**INFINITY is confined within the finite space of a print in *Whirlpools.* The artist draws a flat projection of the curve (a loxodrome) that is traced out on the globe by a path that cuts across all meridians at a constant angle. As any navigator knows, sailing such a "rhumb line" results in a never-ending, ever tightening spiral about the earth's pole. Escher used one woodblock for both colors. Printing once for the red, he turned it halfway around and printed for the gray.**

***Whirlpools,* 1957**



have fractional, or fractal, dimension, an ambiguity that Escher would doubtless have enjoyed. In 1965 he confessed: "I cannot help mocking all our unwavering certainties. It is, for example, great fun deliberately to confuse two and three dimensions, the plane and space, and to poke fun at gravity." Escher was masterful at confusing dimensions, as in *Day and Night* [*see upper illustration on opposite page*], in which two-dimensional farm fields mysteriously metamorphose into three-dimensional geese. He also delighted in pointing out the ambiguities and contradictions inherent in a common practice of science: pasting together several local views of an object to form a global whole [*see lower illustration on opposite page*].

Near the end of his life (he died in 1972), Escher wrote, "Above all, I am happy about the contact and friendship of mathematicians that resulted from it all. They have often given me new ideas, and sometimes there is even an interaction between us. How playful they can be, these learned ladies and gentlemen!"

FURTHER READING


THE GRAPHIC WORK OF M. C. ESCHER. M. C. Escher. Ballantine Books, 1971.

THE MAGIC MIRROR OF M. C. ESCHER. B. Ernst. Random House, 1976.

ANGELS AND DEVILS. H.S.M. Coxeter in *The Mathematical Gardner.* Edited by David A. Klarner. Prindle, Weber and Schmidt, 1981.

M. C. ESCHER: HIS LIFE AND COMPLETE GRAPHIC WORK. Edited by J. L. Locher. Harry N. Abrams, 1982.

M. C. ESCHER: ART AND SCIENCE. Edited by H.S.M. Coxeter, M. Emmer, R. Penrose and M. L. Teuber. North-Holland, 1986.

ESCHER ON ESCHER: EXPLORING THE INFINITE. M. C. Escher. Translated by Karin Ford. Harry N. Abrams, 1989.

VISIONS OF SYMMETRY: NOTEBOOKS, PERIODIC DRAWINGS, AND RELATED WORK OF M. C. ESCHER. D. Schattschneider. W. H. Freeman and Company, 1990.

# Secure Distributed Computing

*Networks and computer security often do not go well together, but the developers of the Athena system have yet to see their protocols fail*

by Jeffrey I. Schiller

The advent of the global Internet has brought about a significant change in the way people communicate and do business. Yet as this global village expands, so does the number of less than honest citizens. Furthermore, much of the hardware and software that constitutes the Internet was designed with only the trustworthy user in mind. As a result, intruders can invade the privacy of network communications as well as read or alter stored information.

Recent public discussions of computer security have focused for the most part on active attackers, who introduce false data into a network to corrupt its normal functions. Many hackers have broken into computers on the Internet by exploiting well-known (yet often unrepaired) system vulnerabilities. Now passive attacks—simply listening to network traffic without disrupting its flow—are becoming an ever more important form of violation. Indeed, perhaps one of the most insidious threats to security in modern networks is the ease with which they can be employed for eavesdropping. For example, as long as access to computers depends on passwords typed over a network, a passive attacker can mine digital gold simply by listening.

At the Massachusetts Institute of Technology, computer-system administrators have been coping with problems of security for more than 10 years. In 1983, in collaboration with IBM and the Digital Equipment Corporation, M.I.T. undertook Project Athena to provide advanced computing power to students and faculty. Athena is based on the client-server model: users' workstations (the clients) do most of the computing, but specialized machines on a campus-wide network, MITnet, handle file storage, printing and other needs. (This model has since become a widely used paradigm for much of the Internet.)

When we were designing the Athena system, we knew the network represented a major security threat. Quite frankly, we assumed that our students would quickly figure out how to program campus workstations so that they would act as efficient password-grabbing machines. To reduce this danger, we developed the Kerberos authentication system, which avoids ever sending readable passwords across vulnerable network links. This namesake of the great three-headed dog that once stood guard at the entrance to the underworld is the heart of Athena's security. Unlike its predecessor, however, our cybernetic watchdog has yet to fail in its vigil.

Security challenges posed by computer networks are as much conceptual as they are practical. It turns out that a few lengths of copper wire, coaxial cable or optical fiber can threaten to confuse the very categories

JEFFREY I. SCHILLER has managed the Massachusetts Institute of Technology's campus computer network, which includes the Athena distributed computing system, since its inception in 1984. He is one of the authors of the Kerberos authentication system and leads the security discussion of the Internet Engineering Steering Group and the Internet Engineering Task Force. Schiller has been involved in the development of Internet standards for privacy-enhanced E-mail, and he recently arranged the release of a version of the popular PGP encryption program that is legal to use in the U.S. Schiller received his S.B. in electrical engineering from M.I.T. in 1979.

that people use to decide what is secure and what is not. Consider the notion of computer-system boundaries. For years, specialists have divided their concerns into two classes: perimeter security, which prevents those on the "outside" from getting "inside" a system, and internal security, which keeps users inside from interfering with one another or otherwise violating the security policy of the system.

Defining the perimeter of a traditional mainframe system is fairly simple. The central-processing unit, memory and disk drives are inside, and everything else is outside. Input-output devices such as card readers, tape readers and terminals are the perimeter; any information entering the system must pass through them. When users sit down at a terminal, for example, they authenticate themselves by typing an account name and password. Only the owner of a particular account should know its password, and so the two in conjunction suffice to identify the user to the system.

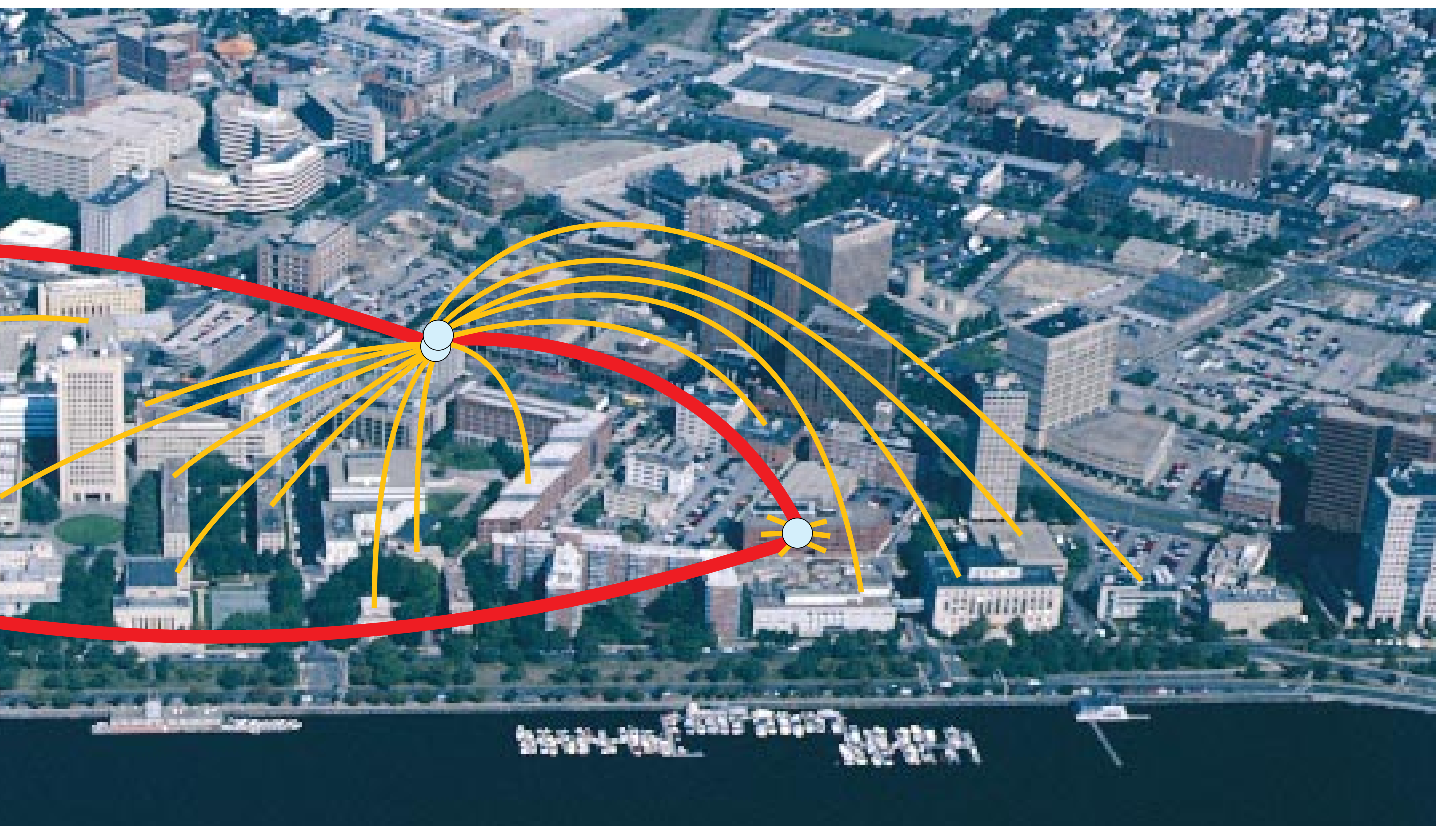
It is important to note that the authentication of the system to the user is implicit—the user assumes that he or she is communicating with the intended computer. If the terminal is connected directly to a specific computer, the user knows empirically that the terminal "speaks" for that system. If users dial a mainframe via modem, they trust the telephone company to connect the telephone call to the computer system that corresponds to the dialed number. (Users and systems managers also trust the mechanism by which the appropriate telephone number is distributed.)

As long as a computer's internal security is not subverted, the mutual trust between user and mainframe remains in force until the user logs out. The system assumes that all the keystrokes it receives have been typed on behalf of the user. Similarly, the user assumes that all information appearing on the screen comes from the appropriate computer system.

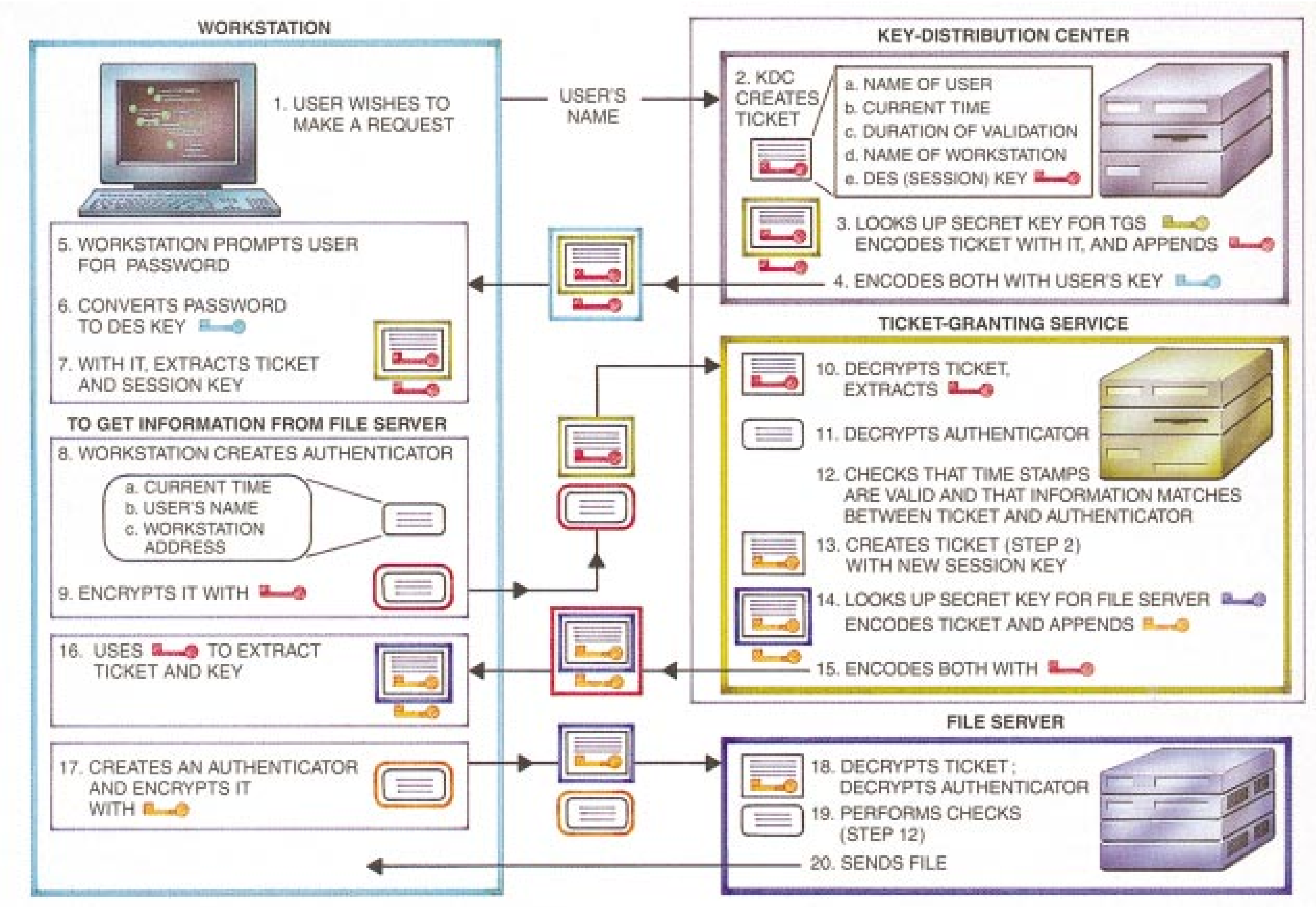
These basic axioms of mainframe security fail once workstations and networks come into the picture. First, the perimeter of a networked system is not easily defined. Rather than a single computer system around which a line can be drawn, a distributed computing environment consists of many independent systems connected by links of uncertain trustworthiness or reliability.

Most networks tend to be physically large and difficult to secure. At M.I.T., wires crisscross the entire campus, extending even into the residence halls. It is simply too expensive to lock all the wires inside telephone closets and unbreachable conduits. In addition, most networks are also broadcast based: every one of the computers connected to a network cable has access to all the information that flows through that cable. Hardware and software within the computers sort out the data going by; each one generally processes only that information addressed to it. An intruder who has control of a computer, however, can readily program it to receive either all the data on the network link it is connected to or all the data intended for some other computer. An intruder may also be able to send information while making it appear to have come from somewhere else.

The ease with which an intruder can perform such illicit acts means that the network is not, properly speaking, inside the security perimeter of the distributed computing environment. If the



**CAMPUS NETWORK links more than 10,000 computers at the Massachusetts Institute of Technology. Its backbone consists of a 100 megabit-per-second fiber-optic ring (*depicted in part by red lines*). Routers (*blue circles*) strategically located across the campus transfer data between more than 100 Ethernet local-area networks (*some shown here in yellow*) and the fiber-optic backbone as needed. About 1,200 of the computers attached to MITnet are part of the Athena computing environment; interactions among them are protected by cryptographic protocols.**




**KERBEROS PROTOCOLS rely on encryption keys, known only to the appropriate parties in a transaction, to protect information sent across an open network. The diagram above shows the sequence that takes place when users first log in to the system and gain access to their files. Only the first exchange requires a user's password; subsequent requests rely on the session key (*red*) shared by the user and the ticket-granting service.**

network is outside the perimeter, then one must somehow protect data packets as they carry information between workstations and servers. Every packet must be authenticated as it crosses the security perimeter represented by the network.

In the Athena system, this authentication is the task of Kerberos, a set of distributed software that employs a series of encrypted exchanges of information to allow a user access to servers. Kerberos also provides for cryptographic checks to make sure that data passing between workstations and servers are not corrupted either by accident or by third-party malice.

The version of Kerberos now in use at M.I.T. employs the data encryption standard (DES) to encode its communications. DES works by breaking messages into discrete blocks of information (usually eight characters—64 bits) and transforming them into blocks of ciphertext according to a 56-bit "key." Decryption requires the same key be used to convert ciphertext blocks back into the original message. Before sending a packet, a workstation can encrypt it using a key known solely to that workstation and to the server for which the data are intended. Intercepting network transmissions will do an attacker no good, because without the key each packet is digital gibberish.

Furthermore, the nature of the DES algorithm makes it easy to detect any hostile alteration of information passing through the network. Virtually any changes to a packet will cause decryption to yield garbage—random characters completely unrelated to the original message. Such corrupted messages are easy to detect, and a workstation or server can simply discard them and request a retransmission.

Cryptographic methods for sending information safely across unsecured networks are only the foundation of Kerberos. On them we have built a superstructure of protocols that can identify the individuals requesting computing services. It can also arrange for workstations and servers to have a secret key known to both machines and no one else (so that cryptographic protections will in fact be effective). The protocols start with a secret 56-bit DES key for each user. Each network service also has a secret key. People cannot remember a random string of 56 bits (which corresponds to a 20-digit number), and so the system permits them to pick a password of six to 128 characters. An additional encryption step converts the password into a DES key. For example, the character string "mydog" translates into 12,322,343,883,628, 311, 502 (10101011000000011100001001111001101111111101111101111001111001110).

All the keys of both users and servers are known only to a special server called the key-distribution center (KDC), which mediates all transactions by means of encrypted digital certificates we call tickets. When users walk up to workstations and type their login names (the name by which the system knows them), they initiate an exchange between the workstation and the KDC that eventually results in a ticket to use a specified network service. First, the workstation tells the KDC that the user wishes to make a request—perhaps to use data stored on a file server. The KDC creates a data packet—the ticket—that contains the name of the user, the current time, the length of time the ticket will be valid, the name of the workstation and a randomly generated DES key, called a session key. It then looks up the secret key for the file server and encrypts the ticket so that only the file server can read it.

The KDC then encodes this already encrypted ticket with the user's secret key and sends the result back to the workstation. When the workstation receives this information, it prompts the user for a password. It converts the password to a DES key and employs it to decrypt the ticket and accompanying session key. Only if the user supplies the correct password to the workstation will it be able to decrypt this information properly.

When the user wants to request information from the file server, the workstation forwards the ticket and an additional packet of data, called an authenticator, to the server. (The authenticator consists of the current time, the user's name and the workstation address, all encrypted with the session key.) The file server decrypts the ticket and extracts the session key, with which it can then decrypt the authenticator. It makes sure that the user and workstation named in the ticket match those in the authenticator; it also checks to make sure that the time stamps are valid. If all these credentials pass inspection, the file server accepts the ticket and processes the user's request.

Except for the user's name, all the information in these transactions passes across the network in encrypted form and is thus safe from prying eyes. Furthermore, in addition to helping to authenticate the user, the session key enclosed in the ticket enables the user to encrypt data passing between his or her workstation and the file server.

For situations in which it may be tricky to determine whether data have been corrupted, Kerberos makes provisions for a message authenticity check (MAC). This additional safeguard is typically a 128-bit number derived from the original unencrypted data and a secret key shared by the workstation and the server; it is appended to a message. When the message arrives, the MAC is computed again. If the two values match, the recipient can be sure it was not tampered with. Indeed, if the data needing to be sent from client to server are not confidential, then a message authenticity check may be sufficient to protect it against unauthorized tampering, without having to bother with actually encrypting it.

Although the basic Kerberos protocol is secure, it is not well suited for ordinary users. In operation, it reveals a potentially troublesome weakness. Each new service requires a different ticket, and so a user could be required to supply a password any number of times during a session. Storing the password on a workstation puts it at risk. Many events could induce a user to leave a workstation without logging out (and thus erasing the stored password). A clever intruder could simply walk up to such an unattended workstation and steal the password from it.

Prompting the user for a password whenever one is needed solves the unattended-workstation problem, but in the long run the repeated prompts are equally risky. One cannot always predict when the user will need a new service. For example, the Athena Post Office server, which holds unread electronic mail, is not the same as the file server where ordinary files reside. Transactions with the Post Office thus require their own tickets, and so a user might have to supply a password every time new E-mail arrived. Assuming that users hold their temper at being asked repeatedly to prove their identity, they will get used to supplying a password to any program that requests it. They then become easy prey for an intruder who provides a program (perhaps a computer game) that prompts for a password but instead of using it for Kerberos authentication stores it away for later pickup.

The solution to this dilemma is the ticket-granting service (TGS). The TGS runs on the same system as the key-distribution center and has access to its database of users, services and keys. Users provide a password once when logging in, to fetch a ticket for the TGS from the KDC. Subsequent requests for tickets to other services go to the TGS, which encrypts them not with the user's password but rather with the session key that accompanied the initial TGS ticket.

Consequently, a user's password need not be stored on the workstation. It resides in memory just long enough to permit the workstation to decrypt the TGS ticket. If a user leaves his or her workstation unattended, an intruder may be able to obtain tickets and session keys from the workstation. These tickets, however, are usable only from that workstation (because they contain the name of the workstation), and each is valid only for a few hours (10 hours at most in M.I.T.'s configuration).

Using Kerberos and the associated encryption does not ensure that a dis-

## The Athena Network

Project Athena employs two basic kinds of computing elements: workstations and servers. Workstations contain the CPU horsepower needed to run users' computations, but they have no facilities for long-term storage of private information. The workstations' local hard disks contain only the minimum set of programs needed to connect them to the campus computer network, MITnet. Once connected, a workstation can obtain a full set of operating-system software from a file server.

File storage, like other features of the computing environment, is provided by "server" computers. These may be advanced workstations, minicomputers, mainframes or any other machine that system managers deem appropriate, just as long as they can communicate using the proper protocols. Athena employs servers to store private data files, to run printers, to operate the electronic-mail system and to run other subsidiary services.

Workstations are expensive, and so the Massachusetts Institute of Technology does not require students to purchase their own systems. Instead machines are available for student use in certain public areas on campus. A typical cluster may contain between 12 and 120 workstations. Recently we have expanded MITnet into on-campus dormitories and off-campus living groups (fraternities and sororities). Students can thus connect their own computers to the network. These systems can function either independently or as workstations integrated into the Athena environment.

**SERVER COMPUTERS for the Athena environment reside, for the most part, in this basement space at M.I.T.'s network operations center. A locked room (dubbed "the dungeon") within the server area holds the computer that runs the Kerberos key-distribution center.**

tributed computing environment is secure. Other measures are necessary to prevent attackers from making an end run around the security that Kerberos can give. First, the servers must be secure in their own right. Kerberos provides a means for workstations (or servers) to confirm individual identities. After that, however, a server must still make the proper decision about whom it will grant access to a particular resource. Fortunately, the task is not that hard. Most servers can be configured so that only system managers can log in to them directly (that is, without requesting services via the ticket mechanism). Security is easier to ensure because only trusted people have access.

Even more important, the Kerberos KDC itself must be physically safe. It has a copy of every user's key, and so an intruder who gains access to the KDC and reads this information in essence gets everyone's password. Indeed, intruders do not need to modify any information on the KDC itself to do this; they merely need to read it (or read the backup tapes!). The M.I.T. implementation of the KDC has some software features that make it difficult for an intruder to read out the DES keys directly, but it is still vital that physical security be maintained by guarding both the KDC and the vaults where its backup tapes are kept.

The KDC must be secure, but it must also always be running so that users can log in to Athena. This requirement poses a conundrum because the traditional approach to improving reliability in a distributed computing environment is to replicate critical services on multiple machines (usually in different locations) as a hedge against power failure or other problems. Yet each additional KDC must also be physically protected. An increase in the number of key-distribution centers enhances reliability, but it also amplifies the risk that the system will be compromised.

Kerberos seems solid in theory, but how does it work in practice? We began using it at M.I.T. in 1986, and the current version (version 4) has been in widespread operation on campus since 1987. We know of no situations in which Athena security has been compromised because of a defect in the Kerberos security mechanisms. Nevertheless, Kerberos does not protect systems from such traditional breaches as the guessing of poorly chosen passwords, deception of system managers by wily intruders or abuse of privileges by trusted individuals. The last two are almost impossible to extirpate. But in 1991 we installed a bad-password filter in our password-changing program; it will not permit people to choose an easily guessed password (such as their name or a common word).

Many other institutions have also adopted Kerberos. Most notably, the Open Software Foundation (an organization that promotes standardized operating systems) has made version 5 a vital component of its distributed computing environment. This version of Kerberos has some new features; in addition, it is less dependent on the details of the Unix operating system and can therefore be adapted to other computing environments. It is now undergoing testing.

Discussions about Kerberos are also under way in the Internet community, where software is desperately needed to protect data as they pass around the globe. We are working on new protocols that might allow us to build distributed systems like Athena that would serve hundreds of thousands of users instead of the 25,000 that we now support. Public-key cryptography (a form of encryption that uses two keys but requires that just one of them be kept secret) will probably play a crucial role in these systems. Furthermore, public-key techniques could be one way to reduce the physical security requirements for key-distribution centers and minimize the risk from multiple copies of the KDC.

The success to date of Kerberos comes from a number of sources. An obvious one is experience: Athena was the first widely used distributed computing system to address security and authentication. Another is simplicity. Users logging in to Athena see the same messages they would if they were communicating by means of a typical multiuser system—indeed, most users are completely ignorant of Kerberos's internal workings. Finally, Kerberos is open: M.I.T. has a free distribution policy (as it does for other software packages such as the X Window system). So the source code for Kerberos is available to potential users and computer scientists anywhere. Many of them have offered their experience and skills in refining it and ferreting out bugs or potential weaknesses, and we expect they will continue to do so.

FURTHER READING


USING ENCRYPTION FOR AUTHENTICATION IN LARGE NETWORKS OF COMPUTERS. R. M. Needham and M. D. Schroeder in *Communications of the ACM,* Vol. 21, No. 12, pages 993–999; 1978.

SECTION E.2.1: KERBEROS AUTHENTICATION AND AUTHORIZATION SYSTEM. S. P. Miller, B. C. Neuman, J. I. Schiller and J. H. Saltzer in *Project Athena Technical Plan.* MIT Project Athena, December 1987.

KERBEROS: AN AUTHENTICATION SERVICE FOR OPEN NETWORK SYSTEMS. J. G. Steiner, B. C. Neuman and J. I. Schiller in *Usenix Conference Proceedings,* Dallas, Tex., pages 191–202; February 1988.

DISTRIBUTED COMPUTING, IMPLEMENTATION AND MANAGEMENT STRATEGIES. Edited by Raman Khanna. Prentice Hall, 1993.

# Why Children Talk to Themselves

*Although children are often rebuked for talking to themselves out loud, doing so helps them control their behavior and master new skills*

by Laura E. Berk

As any parent, teacher, sitter or casual observer will notice, young children talk to themselves—sometimes as much or even more than they talk to other people. Depending on the situation, this private speech (as modern psychologists call the behavior) can account for 20 to 60 percent of the remarks a child younger than 10 years makes. Many parents and educators misinterpret this chatter as a sign of disobedience, inattentiveness or even mental instability. In fact, private speech is an essential part of cognitive development for all children. Recognition of this fact should strongly influence how both normal children and children who have trouble learning are taught.

Although private speech has presumably been around as long as language itself, the political climate in Russia in the 1930s, and the authority of a great Western cognitive theorist, prevented psychologists and educators from understanding its significance until only very recently. In Russia more than six decades ago, Lev S. Vygotsky, a prominent psychologist, first documented the importance of private speech. But at that time, the Stalinist regime systematically persecuted many intellectuals, and purges at universities and research institutes were common.

In fear, Soviet psychologists turned on one another. Some declared Vygotsky a renegade, and several of his colleagues and students split from his circle. According to the recollections of one of Vygotsky's students, the Communist party scheduled a critical "discussion" in which Vygotsky's ideas would be the major target. But in 1934, before Vygotsky could replicate and extend his preliminary studies or defend his position to the party, he died of tuberculosis. Two years later the Communist party banned his published work.

In addition to not knowing about Vygotsky, Western psychologists and educators were convinced by the eminent Swiss theorist Jean Piaget that private speech plays no positive role in normal cognitive development. In the 1920s, even before Vygotsky began his inquiries, Piaget had completed a series of seminal studies in which he carefully recorded the verbalizations of three- to seven-year-olds at the J. J. Rousseau Institute of the University of Geneva. Besides social remarks, Piaget identified three additional types of utterances that were not easily understood or clearly addressed to a listener: the children repeated syllables and sounds playfully, gave soliloquies and delivered what Piaget called collective monologues.

Piaget labeled these three types of speech egocentric, expressing his view that they sprang only from immature minds. Young children, he reasoned, engage in egocentric speech because they have difficulty imagining another's perspective. Much of their talk then is talk for themselves and serves little communicative function. Instead it merely accompanies, supplements or reinforces motor activity or takes the form of non sequiturs: one child's verbalization stimulates speech in another, but the partner is expected neither to listen nor understand. Piaget believed private speech gradually disappears as children become capable of real social interaction.

Although several preschool teachers and administrators openly questioned Piaget's ideas, he had the last word until Vygotsky's work reached the West in the 1960s. Three years after Joseph Stalin's death in 1953, Nikita S. Khrushchev criticized Stalin's "rule by terror" and announced in its place a policy that encouraged greater intellectual freedom. The 20-year ban on Vygotsky's writings came to an end. In 1962 an English translation of Vygotsky's collection of essays, *Thought and Language,* appeared in the U.S. Within less than a decade, a team led by Lawrence Kohlberg of Harvard University had compiled provocative evidence in support of Vygotsky's ideas.

In the late 1970s some American psychologists were becoming disenchanted with Piaget's theory, and at the same time, a broader range of Vygotsky's writings appeared in English. These conditions, coupled with Kohlberg's results, inspired a flurry of new investigations. Indeed, since the mid-1980s the number of studies done on private speech in the West has increased threefold. Most of these studies, including my own, corroborate Vygotsky's views.

In his papers Vygotsky described a strong link between social experience, speech and learning. According to the Russian, the aspects of reality a child is ready to master lie within what he called the zone of proximal (or potential) development. It refers to a range of tasks that the child cannot yet accomplish without guidance from an adult or more skilled peer. When a child discusses a

LAURA E. BERK is currently a professor of psychology and Outstanding University Researcher at Illinois State University. She received her B.A. in psychology from the University of California, Berkeley, and her M.A. and Ph.D. in educational psychology from the University of Chicago. Berk has been a visiting scholar at Cornell University, at the University of California, Los Angeles, and at Stanford University, and her research has been funded by the U.S. Office of Education and the National Institute of Child Health and Human Development. She is co-editor of *Private Speech: From Social Interaction to Self-Regulation* and author of two widely distributed textbooks, *Child Development* and *Infants, Children, and Adolescents.* She has also written numerous journal articles.

PRIVATE SPEECH enables all children to direct their own behavior, acquire new skills and otherwise work through situations that are unfamiliar to them. When any child encounters a new task, he or she will state out loud those features of the problem that seem puzzling. As the youngster's competence grows, this private speech turns into inaudible muttering. Finally, when the cognitive operations necessary to succeed at that task are well practiced, the child thinks words silently.

challenging task with a mentor, that individual offers spoken directions and strategies. The child incorporates the language of those dialogues into his or her private speech and then uses it to guide independent efforts.

"The most significant moment in the course of intellectual development," Vygotsky wrote, "...occurs when speech and practical activity, two previously completely independent lines of development, converge." The direction of development, he argued, is not one in which social communication eventually replaces egocentric utterances, as Piaget had claimed. Instead Vygotsky proposed that early social communication precipitates private speech. He maintained that social communication gives rise to all uniquely human, higher cognitive processes. By communicating with mature members of society, children learn to master activities and think in ways that have meaning in their culture.

As the child gains mastery over his or her behavior, private speech need not occur in a fully expanded form; the self, after all, is an extremely understanding listener. Consequently, children omit words and phrases that refer to things they already know about a given situation. They state only those aspects that still seem puzzling. Once their cognitive operations become well practiced, children start to "think words" rather than saying them. Gradually, private speech becomes internalized as silent, inner speech—those conscious dialogues we hold with ourselves while thinking and acting. Nevertheless, the need to engage in private speech never disappears. Whenever we encounter unfamiliar or demanding activities in our lives, private speech resurfaces. It is a tool that helps us overcome obstacles and acquire new skills.

Currently two American research programs, my own and that of Rafael M. Diaz at Stanford University, have sought to confirm and build on Vygotsky's findings. Our respective efforts began with similar questions: Do all children use private speech? Does it help them guide their actions? And does it originate in social communication? To find out, I chose to observe children in natural settings at school; Diaz selected the laboratory.

Ruth A. Garvin, one of my graduate students, and I followed 36 low-income Appalachian five- to 10-year-olds, who attended a mission school in the mountains of eastern Kentucky. We recorded speech in the classroom, on the playground, in the halls and in the lunchroom throughout the day—paying special attention to those remarks not specifically addressed to a listener.

Our findings revealed that egocentric speech, Piaget's focus, seldom occurred. Most of the comments we heard either described or served to direct a child's actions, consistent with the assumption that self-guidance is the central function of private speech. Moreover, the children talked to themselves more often when working alone on challenging tasks and also when their teachers were not immediately available to help them. In either case, the children needed to take charge of their own behavior.

Furthermore, we found evidence suggesting that private speech develops similarly in all children and that it arises in social experience. The private speech of the Appalachian students changed as they grew older in ways that were much like those patterns Kohlberg had reported a decade and a half earlier.

Middle-class children, such as those Kohlberg observed, speak out loud to themselves with increasing frequency between four and six years of age. Then, during elementary school, their private speech takes the form of inaudible muttering. The Appalachian children moved through this same sequence but did so more slowly. At age 10, more than 40 percent of their private speech remained highly audible, whereas Kohlberg's 10-year-olds spoke out loud to themselves less than 7 percent of the time.

To explain the difference, we studied Appalachian culture and made a striking discovery. Whereas middle-class parents frequently converse with their children, Appalachian parents do so much less often. Moreover, they usually rely more on gestures than on words. If Vygotsky's theory is correct, that private

## Varieties of Private Speech

| CATEGORY | DESCRIPTION | EXAMPLE |
|---|---|---|
| **Egocentric Communication** | Remarks directed to another that make no sense from the listener's perspective. | David says to Mark, who is sitting next to him on the rug, "It broke," without explaining what or when. |
| **Fantasy Play** | A child role-plays and talks to objects or creates sound effects for them. | Jay snaps, "Out of my way!" to a chair after he bumps into it. |
| **Emotional Release** | Comments not directed to a listener that express feelings, or those that seem to be attempts to review feelings about past events or thoughts. | Rachel is sitting at her desk with an anxious look on her face, repeating to herself, "My mom's sick, my mom's sick." |
| **Self-Direction** | A child describes the task at hand and gives himself or herself directions out loud. | Carla, while doing a page in her math book, says out loud, "Six." Then, counting on her fingers, she continues, "Seven, eight, nine, 10. It's 10, it's 10. The answer's 10." |
| **Reading Aloud** | A child reads written material aloud or sounds out words. | "Sher-lock Holm-lock, Sherlock Holme," Tommy reads, leaving off the final "s" in his second, more successful attempt. |
| **Inaudible Muttering** | Utterances so quiet that an observer cannot understand them. | Angela mumbles inaudibly to herself as she works on a math problem. |

speech stems from social communication, then this taciturn home environment might explain the slow development of private speech in Appalachian children.

While our Appalachian study was under way, Diaz and one of his graduate students, Marnie H. Frauenglass, videotaped 32 three- to six-year-olds in the laboratory as the youngsters matched pictures and solved puzzles. Frauenglass and Diaz also found that private speech becomes less audible with age. Yet their results, along with those of other researchers, posed serious challenges to Vygotsky's theory. First, many children emitted only a few utterances, and some none at all—seeming proof that private speech is not universal.

**LEV S. VYGOTSKY, shown here with one of his daughters, first argued that private speech plays a positive role in the development of all children. Because of political conditions in the former Soviet Union, however, his work remained largely unknown until the 1960s.**

Another difficulty arose. If private speech facilitates self-regulation, as Vygotsky believed, then it should relate to how a child behaves while working and how well the child performs. Yet in Frauenglass and Diaz's study, children who used more private speech did worse on the tasks set before them! Other researchers had reported weak and sometimes negative associations between private speech and performance as well.

Diaz crafted some insightful explanations for these outcomes. After a close look at Vygotsky's definition of the zone of proximal development, Diaz concluded that perhaps the tasks typically given in the laboratory were not suitable for evoking private speech in all children. Some children may have been so familiar with solving puzzles and matching pictures that the cognitive operations they needed to succeed were already automatic. Other children may have found these tasks so difficult that they could not master them without help. In either case, self-guiding private speech would not be expected. Furthermore, Diaz reasoned that since private speech increases when children encounter difficulties, it would often coincide with task failure. He suggested that the beneficial impact of private speech might be delayed.

Returning to the classroom—this time, to the laboratory school at Illinois State University—I embarked on a series of studies to test these intriguing possibilities. My team of observers carefully recorded the private speech and task-related actions of 75 first to third graders as they worked alone at their desks on math problems. Their teachers considered this work to be appropriately challenging for each child. Graduate student Jennifer A. Bivens and I then followed the first graders and monitored their behavior as second and third graders.

Every child we observed talked to himself or herself—on average 60 percent of the time. Also, as in previous studies, many children whose remarks described or otherwise commented on their activity received lower scores on homework and achievement tests taken that same year. Yet private speech that was typical for a particular age predicted gains in math achievement over time. Specifically, first graders who made many self-guiding comments out loud or quietly did better at second-grade math. Likewise, second graders who often muttered to themselves grasped third-grade math more easily the following year.

Also, the relationship we noted between a child's use of private speech and his or her task-related behavior bolstered Vygotsky's hypothesis that self-guiding comments help children direct their actions. Children whose speech included a great deal of task-irrelevant wordplay or emotional expression often squirmed in their seats or chewed on or tapped their pencils against their desks.

In contrast, children who frequently made audible comments about their work used more nonverbal techniques to help them overcome difficulties, such as counting on fingers or tracking a line of text using a pencil. Finally, children who most often used quiet private speech rarely fidgeted and were highly attentive. Overall, children who progressed most rapidly from audible remarks to inner speech were more advanced in their ability to control motor activity and focus attention. The development of private speech and task-related behavior thus went hand in hand.

In a later investigation, Sarah T. Spuhl, another of my graduate students, and I attempted to witness in the laboratory the dynamic relationship Vygotsky highlighted between private speech and learning—namely, private speech diminishes as performance improves. We added a new dimension to our research as well: an exploration of how the interaction between a child and an adult can foster self-regulation through private speech.

We asked 30 four- and five-year-olds to assemble Lego pieces into a reproduction of a model. Each subject attempted the exercise in three 15-minute sessions, scheduled no more than two to four days apart. This timing permitted us to track their increasing competence. We pretested each child to ensure that the Lego tasks would be sufficiently challenging—something that had not been done before. Only novice Lego builders participated. Two weeks before the sessions began, we videotaped each mother helping her child with activities that required skills similar to those involved in Lego building, such as fitting blocks together and matching their colors and shapes.

Next we evaluated the communication between the mothers and their children as they solved problems together. According to previous research, parenting that is warm and responsive but exerts sufficient control to guide and encourage children to acquire new skills promotes competence. (Psychologists term such parenting authoritative.) In contrast, both authoritarian parenting (little warmth and high control) and permissive parenting (high warmth and little control) predict learning and adjustment problems. Based on this evidence, we thought that the authoritative style might best capture those features of adult teaching we wished to identify.

Our results revealed that children who have authoritative mothers more often used self-guiding private speech. Among the four-year-olds, those experiencing authoritative teaching showed greater improvement in skill over the course of the three Lego-building sessions. Furthermore, we did a special statistical analysis, the outcome of which suggested that private speech mediates the relationship between authoritative parenting and task success—a finding consistent with Vygotsky's assumptions.

Unlike previous laboratory research, every child in our sample used private speech. As expected, the children's comments became more internalized over the course of the three sessions as their skill with the Lego blocks increased. And once again, private speech predicted future gains better than it did concurrent task success. In particular, children who used private speech that was appropriate for their age—audible, self-guiding utterances at age four and inaudible muttering at age five—achieved the greatest gains.

Next I turned my attention to children having serious learning and behavior problems. Many psychologists had concluded that elementary school pupils who were inattentive, impulsive or had learning disabilities suffered from deficits in using private speech. To treat these children, researchers had designed and widely implemented training programs aimed at inducing children to talk to themselves. In a typical program, children are asked to mimic a therapist acting out self-guiding private speech while performing a task. Next the therapist demonstrates lip movements only and finally asks the children to verbalize covertly.

Despite the intuitive appeal of this training, the approach most often failed. I suspected that the design of these treatments might have been premature. The procedures were not grounded in systematic research on how children having learning and behavior problems use private speech. The spontaneous self-regulatory utterances of such children remained largely uninvestigated.

To fill this gap in our knowledge, my graduate student Michael K. Potts and I studied 19 six- to 12-year-old boys who had been clinically diagnosed with attention-deficit hyperactivity disorder (ADHD), a condition characterized by severe inattentiveness, impulsivity and overactivity. Once again, we observed private speech as the subjects worked on mathematics problems at their desks. We compared these observations to the private speech of 19 normal boys matched in age and verbal ability.

Contrary to the assumptions underlying self-instructional training, ADHD boys did not use less private speech. Instead they made substantially more audible, self-guiding remarks than did normal boys. Furthermore, we examined age-related trends and found that the only difference between the two groups was that ADHD boys made the transition from audible speech to more internalized forms at a later age.

We uncovered a possible explanation for this developmental lag. Our results implied that ADHD children's severe attention deficit prevented their private speech from gaining efficient control over their behavior. First, only in the least distractible ADHD boys did audible self-guiding speech correlate with improved attention to math assignments. Second, we tracked a subsample of ADHD subjects while they were both taking and not taking stimulant drug medication, the most widely used treatment for the disorder. (Although stimulants do not cure ADHD, a large body of evidence indicates that they boost attention and academic performance in most children who take them.) We found that this medication sharply increased the maturity of private speech in ADHD boys. And only when these children were medicated did the most mature form of private speech, inaudible muttering, relate to improved self-control.

The promising nature of these findings encouraged me to include children having learning disabilities in the research. My colleague Steven Landau joined me in observing 112 third to sixth graders working on math and English exercises at their desks. Half of the children met the Illinois state guidelines for being classified as learning disabled: their academic achievement fell substantially below what would be expected based on their intelligence. The other half served as controls. As in the ADHD study, we found that the children who had learning disabilities used more audible, self-guiding utterances and internalized their private speech at a later age than did children who did not have a disability. When we looked at a subgroup of learning disabled children who also displayed symptoms of ADHD, this trend was even more pronounced.

EXPERT ADVICE gives children the framework they need to use private speech effectively. When an adult assists a child with a challenging task, he or she can offer spoken directions or strategies to help that child succeed. The child can then incorporate the language from these conversations into private speech. Later, the child can use this language to guide his or her own efforts.

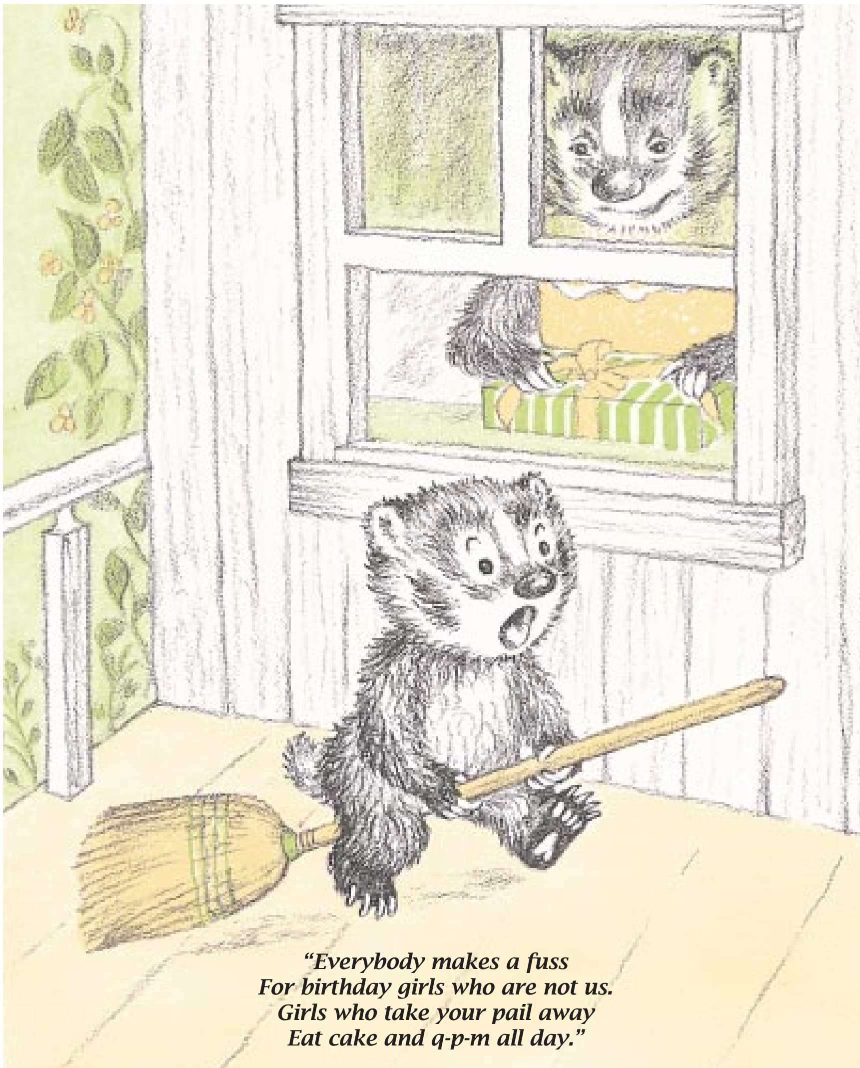
Research on children suffering from persistent learning difficulties vigorously supports Vygotsky's view of private speech. These children follow the same course of development as do their unaffected age mates, but impairments in their cognitive processing and ability to pay attention make academic tasks more difficult for them. This difficulty in turn complicates verbal self-regulation. Our findings suggest that training children who have learning and behavior problems to talk to themselves while performing cognitive tasks amounts to no more than invoking a skill they already possess. Furthermore, interventions that push children to move quickly toward silent self-communication may be counterproductive. While concentrating, ADHD and learning-disabled pupils show heightened dependence on audible private speech in an effort to compensate for their cognitive impairments.

How can our current knowledge of private speech guide us in teaching children who learn normally and those who have learning and behavior problems? The evidence as a whole indicates that private speech is a problem-solving tool universally available to children who grow up in rich, socially interactive environments. Several interdependent factors—the demands of a task, its social context and individual characteristics of a child—govern the extent and ease with which any one child uses self-directed speech to guide behavior. The most profitable intervention lies not in viewing private speech as a skill to be trained but rather in creating conditions that help children use private speech effectively.

When a child tries new tasks, he or she needs communicative support from an adult who is patient and encouraging and who offers the correct amount of assistance given the child's current skills. For example, when a child does not understand what an activity entails, an adult might first give the child explicit directions. Once the child realizes how these actions relate to the task's goal, the adult might offer strategies instead. Gradually, adults can withdraw this support as children begin to guide their own initiatives.

Too often, inattentive and impulsive children are denied this scaffold for learning. Because of the stressful behaviors they bring to the adult-child relationship, they are frequently targets of commands, reprimands and criticism, all of which keep them from learning how to control their own actions.

Finally, parents and teachers need to be aware of the functional value of private speech. We now know that private speech is healthy, adaptive and essential behavior and that some children need to use it more often and for a longer period than others. Still, many adults continue to regard private speech as meaningless, socially unacceptable conduct—even as a sign of mental illness. As a result, they often discourage children from talking to themselves. At home, parents can listen to their child's private speech and thus gain insight into his or her plans, goals and difficulties. Likewise, teachers can be mindful of the fact that when pupils use more private speech than is typical for their age, they may need extra support and guidance. Certainly, we have much more to discover about how children solve problems using spontaneous private speech. Nevertheless, Vygotsky's theory has greatly deepened our understanding of this phenomenon. Today it is helping us design more effective teaching methods for all children and treatments for children suffering from learning and behavior problems. One can only regret that earlier generations of psychologists and educators—and those they might have helped—did not have the advantage of Vygotsky's insights.




**FRANCES, a badger starring in a series of children's books, sings to herself—in this case, to help her cope with the jealousy she feels toward her younger sister, who is soon to celebrate a birthday. Private speech is in fact common in children's literature. The most famous example may well be from *The Little Engine That Could,* in which the protagonist coaches herself, "I think I can."**


FURTHER READING

DEVELOPMENT OF PRIVATE SPEECH AMONG LOW-INCOME APPALACHIAN CHILDREN. Laura E. Berk and Ruth A. Garvin in *Developmental Psychology,* Vol. 20, No. 2, pages 271–286; March 1984.

A LONGITUDINAL STUDY OF THE DEVELOPMENT OF ELEMENTARY SCHOOL CHILDREN'S PRIVATE SPEECH. J. A. Bivens and L. E. Berk in *Merrill-Palmer Quarterly,* Vol. 36, No. 4, pages 443–463; October 1990.

VYGOTSKY: THE MAN AND HIS CAUSE. Guillermo Blanck in *Vygotsky and Education: Instructional Implications and Applications of Sociohistorical Psychology.* Edited by Luis C. Moll. Cambridge University Press, 1990.

DEVELOPMENT AND FUNCTIONAL SIGNIFICANCE OF PRIVATE SPEECH AMONG ATTENTION-DEFICIT HYPERACTIVITY DISORDERED AND NORMAL BOYS. Laura E. Berk and Michael K. Potts in *Journal of Abnormal Child Psychology,* Vol. 19, No. 3, pages 357–377; June 1991.

PRIVATE SPEECH: FROM SOCIAL INTERACTION TO SELF-REGULATION. Edited by Rafael M. Diaz and Laura E. Berk. Lawrence Erlbaum Associates, 1992.

PRIVATE SPEECH OF LEARNING DISABLED AND NORMALLY ACHIEVING CHILDREN IN CLASSROOM ACADEMIC AND LABORATORY CONTEXTS. Laura E. Berk and Steven Landau in *Child Development,* Vol. 64, No. 2, pages 556–571; April 1993.

# Resolving Zeno's Paradoxes

*For millennia, mathematicians and philosophers have tried to refute Zeno's paradoxes, a set of riddles suggesting that motion is inherently impossible. At last, a solution has been found*

by William I. McLaughlin

Once upon a time Achilles met a tortoise in the road. The tortoise, whose mind was quicker than his feet, challenged the swift hero to a race. Amused, Achilles accepted. The tortoise asked if he might have a head start, as he was truly much slower than the demigod. Achilles agreed happily, and so the tortoise started off. After taking quite a bit of time to fasten one of his sandal's ankle straps, Achilles bolted from the starting line. In no time at all, he ran half the distance that separated him from the tortoise. Within another blink, he had covered three quarters of the stretch. In another instant, he made up seven eighths and in another, fifteen sixteenths. But no matter how fast he ran, a fraction of the distance remained. In fact, it appeared that the hero could never overtake the plodding tortoise.

Had Achilles spent less time in the gym and more time studying philosophy, he would have known that he was acting out the classic example used to illustrate one of Zeno's paradoxes, which argue against the possibility of all motion. Zeno designed the paradox of Achilles and the tortoise, and its companion conundra (more about them later), to support the philosophical theories of his teacher, Parmenides.

WILLIAM I. MCLAUGHLIN is a technical manager for advanced space astrophysics at the Jet Propulsion Laboratory in Pasadena, Calif., where he has worked since 1971. He has participated in many projects for the U.S. space program, including the Apollo lunar-landing program, the Viking mission to Mars, the *Infrared Astronomical Satellite* (*IRAS*) and the Voyager project, about which he wrote an article for *Scientific American* in November 1986. He received a B.S. in electrical engineering in 1963 and a Ph.D. in mathematics in 1968, both from the University of California, Berkeley. McLaughlin conducts, in addition, research in epistemology.

Both men were citizens of the Greek colony of Elea in southern Italy. In approximately 445 B.C., Parmenides and Zeno met with Socrates in Athens to exchange ideas on basic philosophical issues. The event, one of the greatest recorded intellectual encounters (if it really took place), is commemorated in Plato's dialogue *Parmenides.* Parmenides, a distinguished thinker nearly 65 years old, presented to the young Socrates a startling thesis: "reality" is an unchanging single entity, seamless in its unity. The physical world, he argued, is monolithic. In particular, motion is not possible. Although the rejection of plurality and change appears idiosyncratic, it has, in general outline, proved attractive to numerous scholars. For example, the "absolute idealism" of the Oxford philosopher F. H. Bradley (1846–1924) has points in common with the Parmenidean outlook.

This portrayal of the world is contrary to our everyday experience and relegates our most fundamental perceptions to the realm of illusion. Parmenides relied on Zeno's powerful arguments, which were later recorded in the writings of Aristotle, to support his case. For two and a half millennia, Zeno's paradoxes have provoked debates and stimulated analyses. At last, using a formulation of calculus that was developed in just the past decade or so, it is possible to resolve Zeno's paradoxes. The resolution depends on the concept of infinitesimals, known since ancient times but until recently viewed by many thinkers with skepticism.

The tale of Achilles and the tortoise depicts one of Zeno's paradoxes, usually denoted "The Dichotomy": any distance, such as that between the two contenders, over which an object must traverse can be halved ($1/2$, $1/4$, $1/8$ and so on) into an infinite number of spatial segments, each representing some distance yet to be traveled. As a result, Zeno asserts that no motion can be completed because some distance, no matter how small, always remains. It is important to note that he does not say that infinitely many stretches cannot add up to a finite distance (glancing at the geometry of an infinitely partitioned line shows immediately, without any sophisticated calculations, that an infinite number of pieces sum to a finite interval). Rather the force of Zeno's objection to the idea of motion comes from the obligation to explain how an infinite number of acts—crossing one interval—can be serially completed.

Zeno made a second attack on the conceptual underpinnings of motion by viewing this first argument from a slightly different perspective. His second paradox is as follows: Before an object, say, an arrow, gets to the halfway mark of its supposed journey (an achievement granted in the preceding case), it must first travel a quarter of the distance. As in Zeno's first objection, this reasoning can be continued indefinitely to yield an infinite regress, thus leading to his insistence that motion could never be initiated.

Zeno's third paradox takes a different tack altogether. It asserts that the very concept of motion is empty of content. Zeno invites us to consider the arrow at any one instant of its flight. At this point in time, the arrow occupies a region of space equal to its length, and no motion whatsoever is evident. Because this observation is true at every instant, the arrow is never in motion. This objection, in a historical sense, proved the most troublesome for would-be explainers of Zeno's paradoxes.

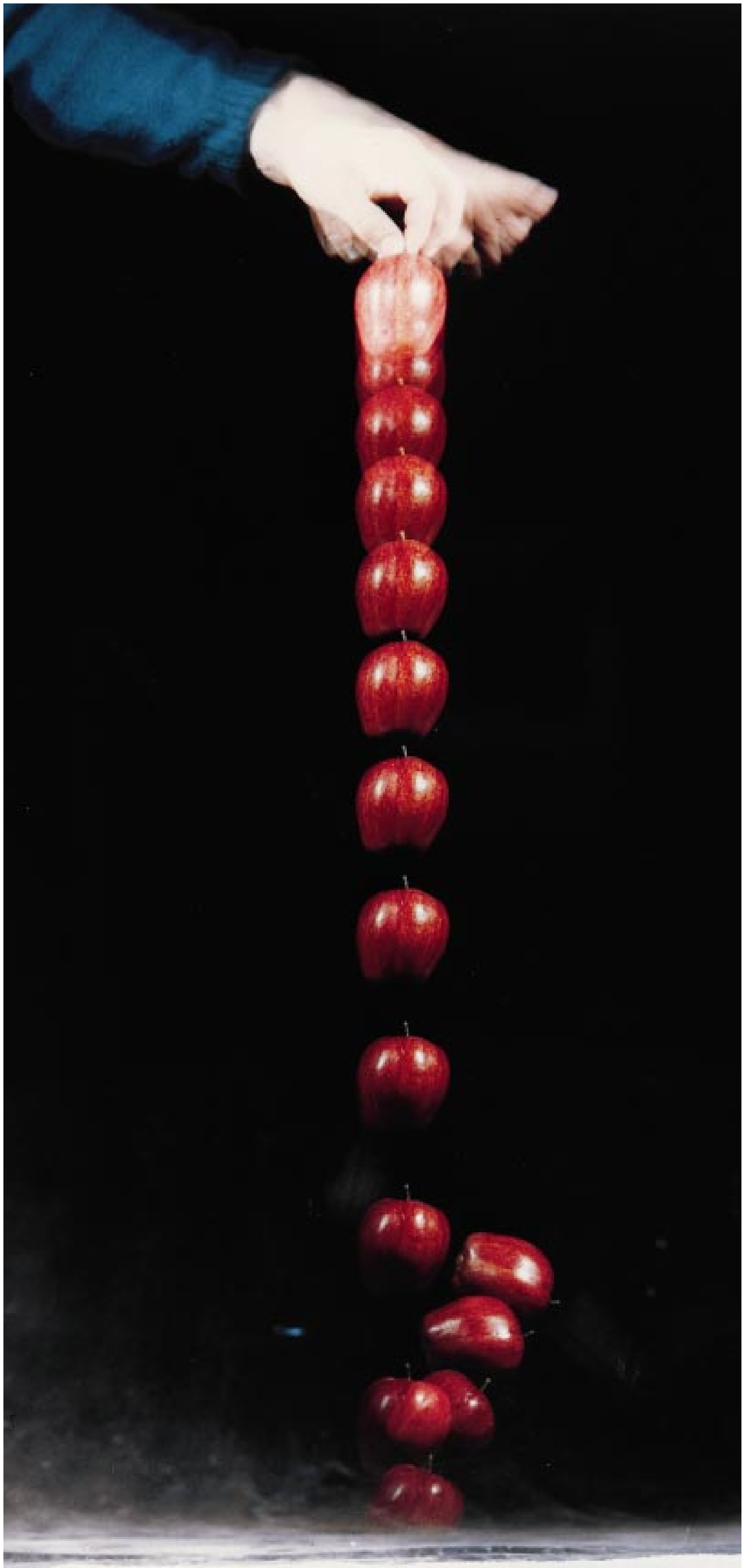
Many philosophers and mathematicians have made various attempts to answer Zeno's objections. The most direct approach has simply been to deny that a problem exists. For example, Johann Gottlieb Waldin, a German professor of philosophy, wrote in 1782 that the Eleatic, in arguing against motion, assumed that motion exists. Evidently the good professor was not acquainted with the form of argument known as reductio ad absurdum: assume a state of affairs and then show that it leads to an illogical conclusion.

Nevertheless, other scholars made progress by wrestling with how an infinite number of actions might occur in the physical world. Their explanations have continually been intertwined with the idea of an infinitesimal, an interval of space or time that embodies the quintessence of smallness. An infinitesimal quantity, some surmised, would be so very near zero as to be numerically impotent; such quantities would elude all measurement, no matter how precise, like sand through a sieve.

Giovanni Benedetti (1530–1590), a predecessor of Galileo, postulated that when an object appeared to be frozen in midair to Zeno, he was in fact seeing only part of the action, as though one were watching a slide show instead of a movie. Between the static images Zeno saw were infinitesimally small instants of time in which the object moved by equally small distances.

Others sidestepped the issue by arguing that intervals in the physical world cannot simply be subdivided an infinite number of times. Friedrich Adolf Trendelenburg (1802–1872) of the University of Berlin built an entire philosophical system that explained human perceptions in terms of motion. In doing so, he freed himself from explaining motion itself.

Similarly, in this century, the English philosopher and mathematician Alfred North Whitehead (1861–1947) constructed a system of metaphysics based on change, in which motion was a special case. Whitehead responded to Zeno's objections by insisting that events in the physical world had to have some extent; namely, they could not be pointlike. Likewise, the Scottish philosopher David Hume (1711–1776) wrote, "All the ideas of quantity upon which mathematicians reason, are nothing but particular, and such as are suggested by the senses and imagination



**FALLING APPLE? Zeno would argue that because the apple appears to be frozen in midair at each instant of its supposed descent, it is never in motion. Moreover, Zeno would assert that there is no proof that the apple will ever reach the ground. Before it arrives there, it must first fall half of the distance between the man's hand and the ground. After that, it must fall half of the remaining distance and half of that again, and so on. How can it be that some fractional distance does not always remain between the apple and the ground? Using similar logic, Zeno would question whether an apple can even begin to fall.**

**RACE between Achilles and the tortoise illustrates one of Zeno's paradoxes. Achilles gives the tortoise a head start. He must then make up half the distance between them, then three fourths, then seven eighths and so on, ad infinitum. In this way, it would seem he could never come abreast of the sluggish animal.**

and consequently, cannot be infinitely divisible."

Either way, the subject of infinitesimals (and whether they exist or not) generated a long and acrimonious literature of its own. Until recently, most mathematicians thought them to be a chimera. The Irish bishop George Berkeley (1685–1753) is noted principally for his idealistic theory, which denied the reality of matter, but he, too, wrestled with infinitesimals. He believed them ill conceived by the mathematicians of the time, including Newton. "They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them ghosts of departed quantities?" He observed further: "Whatever mathematicians may think of fluxions [rates of change], or the differential calculus, and the like, a little reflexion will shew them that, in working by those methods, they do not conceive or imagine lines or surfaces less than what are perceivable to sense."

Indeed, mathematicians found infinitesimals hard to skirt in the course of their discoveries, no matter how distasteful they found them in theory. Some historians believe the great Archimedes (circa 287–212 B.C.) achieved some of his mathematical results using infinitesimals but employed more conventional modes for public presentations. Infinitesimals left their mark during the 17th and 18th centuries as well in the development of differential and integral calculus. Elementary textbooks have long appealed to "practical infinitesimals" to convey certain ideas in calculus to students.

When analysts thought about rigorously justifying the existence of these small quantities, innumerable difficulties arose. Eventually, mathematicians of the 19th century invented a technical substitute for infinitesimals: the so-called theory of limits. So complete was its triumph that some mathematicians spoke of the "banishment" of infinitesimals from their discipline. By the 1960s, though, the ghostly tread of infinitesimals in the corridors of mathematics became quite real once more, thanks to the work of the logician Abraham Robinson of Yale University [see "Nonstandard Analysis," by Martin Davis and Reuben Hersh; SCIENTIFIC AMERICAN, June 1972]. Since then, several methods in addition to Robinson's approach have been devised that make use of infinitesimals.

When my colleague Sylvia Miller and I started our work on Zeno's paradoxes, we had the advantage that infinitesimals had become mathematically respectable. We were intuitively drawn to these objects because they seem to provide a microscopic view of the details of motion. Edward Nelson of Princeton University created the tool we found most valuable in our attack, a brand of nonstandard analysis known by the rather arid name of internal set theory (IST). Nelson's method produces startling interpretations of seemingly familiar mathematical structures. The results are similar, in their strangeness, to the structures of quantum theory and general relativity in physics. Because these two theories have taken the better part of a century to gain widespread acceptance, we can only admire the power of Nelson's imagination.

Nelson adopted a novel means of defining infinitesimals. Mathematicians typically expand existing number systems by tacking on objects that have desirable properties, much in the same way that fractions were sprinkled between the integers. Indeed, the number system employed in modern mathematics, like a coral reef, grew by accretion onto a supporting base: "God made the integers, all the rest is the work of man," declared Leopold Kronecker (1823–1891). Instead the way of IST is to "stare" very hard at the existing number system and note that it already contains numbers that, quite reasonably, can be considered infinitesimals.

Technically, Nelson finds nonstandard numbers on the real line by adding three rules, or axioms, to the set of 10 or so statements supporting most mathematical systems. (Zermelo-Fraenkel set theory is one such foundation.) These additions introduce a new term, standard, and help us to determine which of our old friends in the number system are standard and which are nonstandard. Not surprisingly, the infinitesimals fall in the nonstandard category, along with some other numbers I will discuss later.

Nelson defines an infinitesimal as a number that lies between zero and every positive standard number. At first, this might not seem to convey any particular notion of smallness, but the standard numbers include every concrete number (and a few others) you could write on a piece of paper or generate in a computer: 10, pi, $^1/_{1000}$ and so on. Hence, an infinitesimal is greater than zero but less than any number, however small, you could ever conceive of writing. It is not immediately apparent that such infinitesimals do indeed exist, but the conceptual validity of IST has been demonstrated to a degree commensurate with our justified belief in other mathematical systems.

Still, infinitesimals are truly elusive

entities. Their elusiveness rests on the mathematical fact that two concrete numbers—those having numerical content—cannot differ by an infinitesimal amount. The proof, by reductio ad absurdum, is easy: the arithmetic difference between two concrete numbers must be concrete (and hence, standard). If this difference were infinitesimal, the definition of an infinitesimal as less than all standard numbers would be violated. The consequence of this fact is that both end points of an infinitesimal interval cannot be labeled using concrete numbers. Therefore, an infinitesimal interval can never be captured through measurement; infinitesimals remain forever beyond the range of observation.

So how can these phantom numbers be used to refute Zeno's paradoxes? From the above discussion it is clear that the points of space or time marked with concrete numbers are but isolated points. A trajectory and its associated time interval are in fact densely packed with infinitesimal regions. As a result, we can grant Zeno's third objection: the arrow's tip is caught "stroboscopically" at rest at concretely labeled points of time, but along the vast majority of the stretch, some kind of motion is taking place. This motion is immune from Zenonian criticism because it is postulated to occur inside infinitesimal segments. Their ineffability provides a kind of screen or filter.

Might the process of motion inside one of these intervals be a uniform advance across the interval or an instantaneous jump from one end to the other? Or could motion comprise a series of intermediate steps or else a process outside of time and space altogether? The possibilities are infinite, and none can be verified or ruled out since an infinitesimal interval can never be monitored. Credit for this rebuttal is due to Benedetti, Trendelenburg and Whitehead for their earlier insights, which can now be formalized by means of IST.

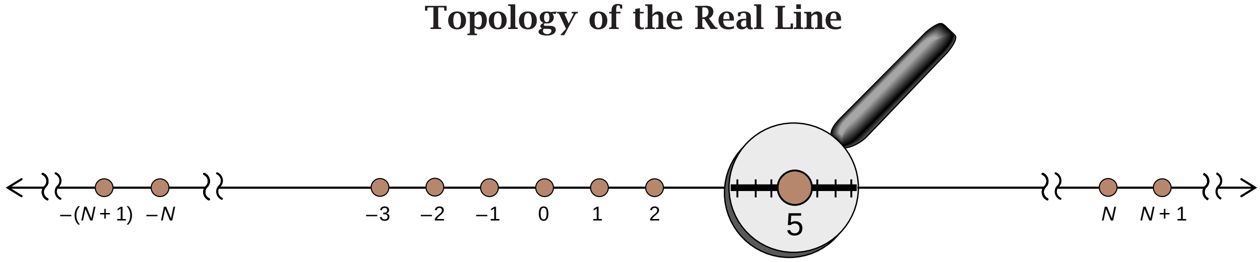
We can answer Zeno's first two objections more easily than we did the third, but we need to use another mathematical fact from IST. Every infinite set of numbers contains a nonstandard number. Before drawing out the Zenonian implications of this statement, it is necessary to talk about the two other types of nonstandard numbers that are readily manufactured from infinitesimal numbers. First, take all the infinitesimals, which by definition are wedged between zero and all the positive, standard numbers, and put a minus sign in front of each one. Now there is a symmetrical clustering of these small objects about zero. To create "mixed" nonstandard numbers, take any standard number, say, one half, and add to it each of the nonstandard infinitesimals in the grouping around zero. This act of addition translates the original cluster of infinitesimals to positions on either side of one half. Similarly, every standard number can be viewed as having its own collection of nearby, nonstandard numbers, each one only an infinitesimal distance from the standard number.

The third type of nonstandard number is simply the inverse of an infinitesimal. Because an infinitesimal is very small, its inverse will be very large (in the standard realm, the inverse of one millionth is one million). This type of nonstandard number is called an unlimited number. The unlimited numbers, though large, are finite and hence smaller than the truly infinite numbers created in mathematics. These unlimited numbers live in a kind of twilight zone between the familiar standard numbers, which are finite, and the infinite ones.

If, as demonstrated in IST, every infinite set contains a nonstandard number, then the infinite series of checkpoints Zeno used to gauge motion in his first argument must contain a mixed, nonstandard number. In fact, as Zeno's infinite series of numbers creeps closer to one, a member of that series will eventually be within an infinitesimal distance from one. At that point, all succeeding members of the series will be nonstandard members of the cluster about one, and neither Zeno nor anyone else will be able to chart the progress of a moving object in this inaccessible region.

There is an element of irony in using infinity, Zeno's putative weapon, to deflate his claims. To refute Zeno's first paradox, we need only state the epistemological principle that we are not responsible for explaining situations we cannot observe. Zeno's infinite series of checkpoints contains nonstandard numbers, which have no numerical meaning, and so we reject his argument based on these entities. Because no one could ever, even in principle, observe the full domain of check-

## Topology of the Real Line




The real numbers consist of the integers (positive and negative whole numbers), rational numbers (those that can be expressed as a fraction) and irrational numbers (those that cannot be expressed as a fraction). The real numbers can be represented as points on a straight line known as the real line (*above*).

The mathematician Edward Nelson of Princeton University labeled three types of numbers as nonstandard within this standard number system. Infinitesimal nonstandard numbers are smaller than any positive standard number yet are greater than zero. Mixed nonstandard numbers, shown grouped around the integer 5, result from adding and subtracting infinitesimal amounts to standard numbers. In fact, every standard number is surrounded by such mixed, nonstandard neighbors. Unlimited nonstandard numbers, represented as $N$ and $N + 1$, are the inverses of infinitesimal nonstandard numbers. Each unlimited number is greater than every standard number and yet less than the infinite real numbers. The nonstandard real numbers prove useful in resolving Zeno's paradoxes.

points that his objection addresses, the objectionable behavior he postulates for the moving object is moot. Many descriptions of motion in the microrealm other than that containing the full series of checkpoints could apply, and just because his particular scenario causes conceptual problems, there is no reason to anathematize the idea of motion. His second argument, attempting to show that an object can never even start to move, suffers from the same malady as the first, and we reject it on like grounds.

We have resolved Zeno's three paradoxes using some technical results from IST and the principle that nonstandard numbers are not suitable for describing matters of fact, observed or purported. Still, more can be said regarding the matter than just the assurance that Zeno's objections do not preclude motion. Indeed, we can construct a theory of motion using a very powerful result from IST. The theory yields the same results as do the tools of the calculus, and yet it is easier to visualize and does not fall prey to Zeno's objections.

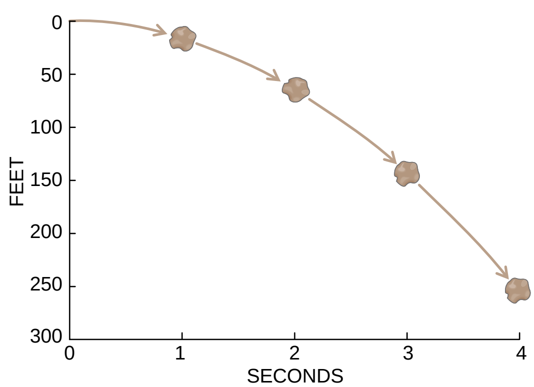
A theorem proved in IST states that there exists a finite set, call it F, that contains all the standard numbers! The corollary that there are only a finite number of standard numbers would seem to be true, but surprisingly, it is not. In developing IST, Nelson needed to finesse the conventional way mathematicians form objects. A statement in IST is called internal if it does not contain the label "standard." Otherwise, the statement is called external. Mathematicians frequently create subsets from larger sets by predicating a quality that characterizes each of the objects in the subset—the balls that are red or the integers that are even. In IST, however, it is forbidden to use external predicates, such as standard, to define subsets; the stricture is introduced to avoid contradictions. For example, imagine the set of all standard numbers in F. This set would be finite because it is a subset of a finite set. It would therefore have a least member, say, $r$. But then $r - 1$ would be a standard number less than $r$, when $r$ was supposed to be the smallest standard number. Thus, we cannot say the standard numbers are finite or infinite in extent, because we cannot form the set of them and count them.

Nevertheless, the finite set F, though constrained as to how it can be visualized, is useful for constructing our theory of motion. This theory can be expressed quite simply as stepping through F, where each member of F represents a distinct moment. For convenience, consider only those members of F that fall between 0 and 1. Let time 0 be the instant when we start tracking a moving object. The second instant when we might try to observe the object is at time $f_1$, where $f_1$ is the smallest member of F that is greater than 0. Ascending through F in this fashion, we eventually reach time $f_n$, where $f_n$ is the largest member of F less than 1. In one more step, we reach 1 itself, the destination in this example. In order to walk through a noninfinitesimal distance, such as the span from 0 to 1 using infinitesimal steps, the subscript $n$ of $f_n$ must be an unlimited integer. The process of motion then is divided into $n+1$ acts, and because $n+1$ is also finite, this number of acts can be completed sequentially.

Of the possible observing times identified earlier, the object's progress could be reported solely at those instants corresponding to certain standard numbers in F. (By the way, $f_1$ and $f_n$ would be nonstandard, as they are infinitesimally close to 0 and 1, respectively.) For example, although we can express a standard number to any finite (but not unlimited) number of decimal places and use this approximation as a measurement label, we cannot access the unlimited tail of the expansion to alter a digit and thus define a nonstandard,

## Calculus by Means of Infinitesimals

To see the relation between infinitesimals and differential calculus, consider the simple case of a falling stone. The distance the stone has traveled in feet can be calculated from the formula $s = 16t^2$, where $t$ equals the time elapsed in seconds. For example, if a stone has fallen for two seconds, it will have traveled 64 feet.




Suppose, however, one wishes to calculate the instantaneous velocity of the stone. The average speed of a moving object equals the total distance it travels divided by the total amount of time it takes. By using this formula over an infinitesimal change in the total distance and time, one can calculate a fair approximation of an object's instantaneous velocity.

Let $dt$ represent an infinitesimal change in time and $ds$ an infinitesimal change in distance. The computation for the velocity of the stone after one second of travel, then, will be as follows: The time frame under consideration ranges from $t = 1$ to $t = 1 + dt$. The position of the stone during that time changes from $s = 16(1)^2$ to $s = 16(1 + dt)^2$. The total change in distance, $32dt + 16dt^2$, divided by $dt$, is the desired average velocity, $32 + 16dt$.

Because $16dt$ is but an infinitesimal amount, undetectable for all intents and purposes, it can be considered equal to 0. Thus, after one second of travel, the formula yields the stone's instantaneous velocity as 32 feet per second.

This manipulation, of course, resembles those used in traditional, differential calculus. There the small residue $16dt$ cannot be dropped at the end of the calculation; it is a noninfinitesimal quantity. Instead, in this calculus, it must be argued away using the theory of limits. In essence, the limit process renders the interval of length $dt$ sufficiently small so that the average velocity is arbitrarily close to 32. As before, the instantaneous velocity of the stone after one second of travel equals 32 feet per second. Similarly, judicious use of infinitesimal regions facilitates the computation of the area of complicated regions, a basic problem of integral calculus. Some think the newer calculus is pedagogically superior to calculus without infinitesimals. Nevertheless, both methods are equally rigorous and yield identical results.

***THE MEASURERS,* a 17th-century Dutch painting attributed to Hendrik van Balen, illustrates the words of the Roman poet Horace: "There is measure in all things." No matter how precise measurements become, however, infinitesimal amounts will forever escape our grasp, since any useful unit of measure must correspond to some standard number.**

infinitesimally close neighbor. Only concrete standard numbers are effective as measurement labels; the utility of their nonstandard neighbors for measurement is illusory.

Much is superfluous in this theory of motion, and much is left unsaid. It suffices, however, in the sense that it can easily be translated into the symbolic notation of the integral or differential calculus, commonly used to describe the details of motion [*see box on opposite page*]. More important in the present context, the finiteness of the set F enables us to jump over the pitfalls in Zeno's first two paradoxes. His third objection is dodged as before: motion in real time is an unknown process that takes place in infinitesimal intervals between the standard points of F; the nonstandard points of F are irrelevant given that they cannot be observed.

For many centuries, Zeno's logic stood mostly intact, proving the refractory nature of his arguments. A resolution was made possible through two basic features of IST: first, the ability to partition an interval of time or space into a finite number of ineffable infinitesimals and, second, the fact that standardly labeled points—the only ones that can be used for measurement—are isolated objects on the real line. Is our work merely the solution to an ancient puzzle? Possibly, but there are several directions in which it might prove extensible.

Aside from its mathematical value, IST is ripe with epistemological import, as this analysis has shown. It might well be modified to constitute a general epistemic logic. Also, infinitesimal intervals, or their generalization, would promise a technical resource to house Whitehead's so-called actual entities, the generative atoms of his philosophical system. Finally, the current theory of motion and the predictions of quantum physics are not dissimilar in that they both restrict the observation of certain events to discrete values. Of course, this theory of motion is not a version of quantum mechanics (nor relativity theory, for that matter). Because the theory resulted from a thought experiment on Zeno's terms, it holds no direct connection to present physical theory. Moreover, the specific rules inherited from IST are probably not those best suited to describe reality. Modern physics might adapt the IST approach by modifying its rule system and introducing "physical constants," perhaps by assigning parameters to the set F.

But maybe not. Still, the simplicity and elegance of such thought experiments have catalyzed research throughout the ages. Notable examples include Heinrich W. M. Olbers, questioning why the sky is dark at night despite stars in every direction, or James Clerk Maxwell, summoning a meddling, microscopic demon to batter the second law of thermodynamics. Likewise, Zeno's arguments have stimulated examinations of our ideas about motion, time and space. The path to their resolution has been eventful.

FURTHER READING


A HISTORY OF GREEK PHILOSOPHY, Vol. 2: THE PRESOCRATIC TRADITION FROM PARMENIDES TO DEMOCRITUS. W. K. Guthrie. Cambridge University Press, 1965.

ZENO OF ELEA. Gregory Vlastos in *The Encyclopedia of Philosophy*. Edited by Paul Edwards. Macmillan Publishing Company, 1967.

NONSTANDARD ANALYSIS. Martin Davis and Reuben Hersh in *Scientific American,* Vol. 226, No. 6, pages 78–86; June 1972.

INTERNAL SET THEORY: A NEW APPROACH TO NONSTANDARD ANALYSIS. Edward Nelson in *Bulletin of the American Mathematical Society,* Vol. 83, No. 6, pages 1165–1198; November 1977.

AN EPISTEMOLOGICAL USE OF NONSTANDARD ANALYSIS TO ANSWER ZENO'S OBJECTIONS AGAINST MOTION. William I. McLaughlin and Sylvia L. Miller in *Synthese,* Vol. 92, No. 3, pages 371–384; September 1992.

# Big-Time Biology

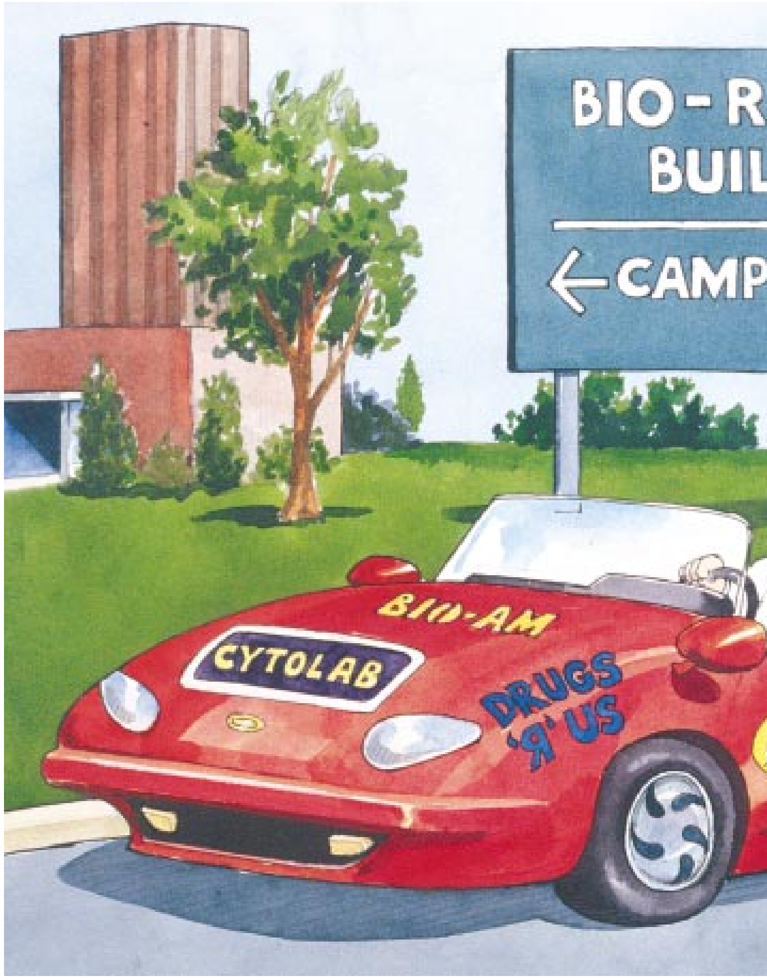
by Tim Beardsley, *staff writer*

Biology just hasn't been the same since 1973, when Stanley N. Cohen of Stanford University and Herbert W. Boyer of the University of California at San Francisco invented the technique that made genetic engineering practical and so launched the biotechnology industry. Cohen and Boyer worked out a straightforward way to transplant functioning genes from different organisms into bacteria, which could then be grown in large quantities. Their achievement opened up vistas that researchers had only dreamed about.

Proteins that previously had to be extracted laboriously from animal or plant tissue could be manufactured at will by using bacteria as miniature factories. At about the same time, researchers learned how to make large amounts of pure antibodies, biological chemicals that can be used to probe cellular processes in fine detail. Life's book, it seemed, was suddenly laid open. And Boyer began tooling around the Bay Area in a Porsche.

The biotechnology boom that followed transformed a good deal more than the techniques used in laboratories. It also changed fundamentally the financial environment and the culture of biological research. Corporations performed half of the nation's health research last year, whereas the National Institutes of Health, the government agency that has been the principal supporter of biomedical research since the 1950s, supported just 32 percent. The number of young investigators applying to the NIH for grants has fallen to half of the 1985 figure. Biotechnology firms, which constitute one of the more volatile sectors of the stock markets, had a combined capitalization of $41 billion in 1994, according to Ernst & Young—$7 billion less than in 1992 but impressive for an industry less than 20 years old.

Has the change been for the good? The privatization of biology has forced researchers and officials to face the question about what role the traditional institutions—the universities and federal laboratories—should have in the future of their science. Ironically, it is these very institutions that produced most of the breakthrough discoveries from which the biotechnology industry has grown. Some senior




*CULTURE CLASH of corporate dollars and an older academic tradition has sometimes caused hard feel-*

scientists even wonder whether research in its traditional form can survive the entrepreneurial deluge. Among them is Paul Berg of Stanford, who is often described as the father of genetic engineering. He fears that the rush to develop lucrative drugs could imperil the unfettered inquiry that is essential if new ideas are to be generated. Others are concerned that the desire to protect proprietary information will short-circuit the open discourse so cherished on university campuses. "Freedom of communication is an area of concern," says Richard F. Celeste, chairman of the Government-University-Industry Research Roundtable, an organization that

*Molecular biology is—not so quietly—*
*evolving from a science into an industry.*
*Can it survive the transformation?*


***ings. Andy Myer saw it this way in 1992 in the June issue of the* Journal of NIH Research.**

studies relationships among those institutions.

Whatever solutions to these issues are devised, turning back the clock will not be among them. The techniques now emerging are so powerful—and the material rewards of success so great—that every other discovery looks like a business opportunity. In the two decades since Cohen and Boyer turned gene shuffling from vaulting accomplishment into an experiment suitable for a high school class, droves of biologists have transmuted themselves into entrepreneurs. Companies aiming to exploit the new technology have proliferated like cells in a petri dish: there are now about 1,300 biotechnology firms in the U.S., and most major agricultural and pharmaceutical corporations have plays in the emerging arena. Five years ago industry overtook government as the primary provider for biomedical research, after decades of federal dominance. The biotechnology industry sponsored $5.7 billion in research in 1993, a 14 percent increase over the previous year, despite a marked tightening of the supply of capital for new investments.

The private sector is filling a vacuum created by the uncertain, fluky currents of public support for biological and other forms of basic inquiry. The NIH underwent a huge expansion in the decades after World War II. Appropriations rose 150-fold between 1945 and 1965, even after inflation, from $26 million to $4 billion in 1988 dollars. That corresponds to a dizzying 28 percent of real growth a year. During that period, the NIH's 312-acre campus in Bethesda, Md., became, arguably, the world's preeminent medical research center.

Just as important were the NIH's extramural grants, which account for about three quarters of its budget. Generations of young investigators at universities around the country came to rely on the agency to fund studies of cells and bodies in health and illness. The in-house and university-based research supported by the NIH was universally understood to be basic in nature, unaffected by commercial imperatives.

The NIH's growth slowed, however, in the 1970s and 1980s. Although the budget increases continued in dollar terms, they became less impressive when allowance was made for the spiraling cost of research. The agency employs a special biomedical research and development index to measure the growth in its costs, and the index climbs significantly faster than do consumer prices. During the past decade, the NIH's budget has been growing at a relatively modest annual rate of 3.9 percent, after allowing for inflation in research costs.

The number of investigators chasing grant dollars has, however, increased dramatically since the early 1970s. In 1972 the NIH reviewed 8,596 grant applications; in 1992 the number was 20,142. Even accomplished researchers now find the odds of winning an NIH grant daunting.

In 1993 only 22 percent of grant applications from scientists 36 years old and younger were funded, down from 33 percent in 1985.

More worrying still, the number of young researchers (those younger than 36 years) applying for grants fell by a startling 54 percent between 1985 and 1993. That trend has caused perplexity. “We don’t understand it yet,” says Harold E. Varmus, the virologist and Nobel laureate who is director of the NIH. Where have all the children gone? Some are probably employed as members of large research teams. Others have migrated to industry or may be working in universities on projects supported by industry. Indeed, the universities, no longer certain of federal support for basic research, have started courting the private sector.

“Universities are looking to supplement federal dollars,” Celeste says. “There has been keen interest in our work on how to negotiate research agreements.” Although high-technology companies typically do a lot of their research in-house, about 7 percent of the health research conducted in universities and medical schools is supported by industry.

The financial ecology started to change significantly in 1980, when Congress passed laws encouraging academic institutions and federal laboratories to make money from technology transfer. The Bayh-Dole Act, which President Jimmy Carter signed that year, was an important milestone: it gave universities the right to patent inventions arising from federally sponsored research and to license them to commercial firms. The Federal Technology Transfer Act of 1986 took things a step further by allowing companies to form partnerships with in-house government scientists at the NIH and elsewhere, through contracts known as collaborative research and development agreements.

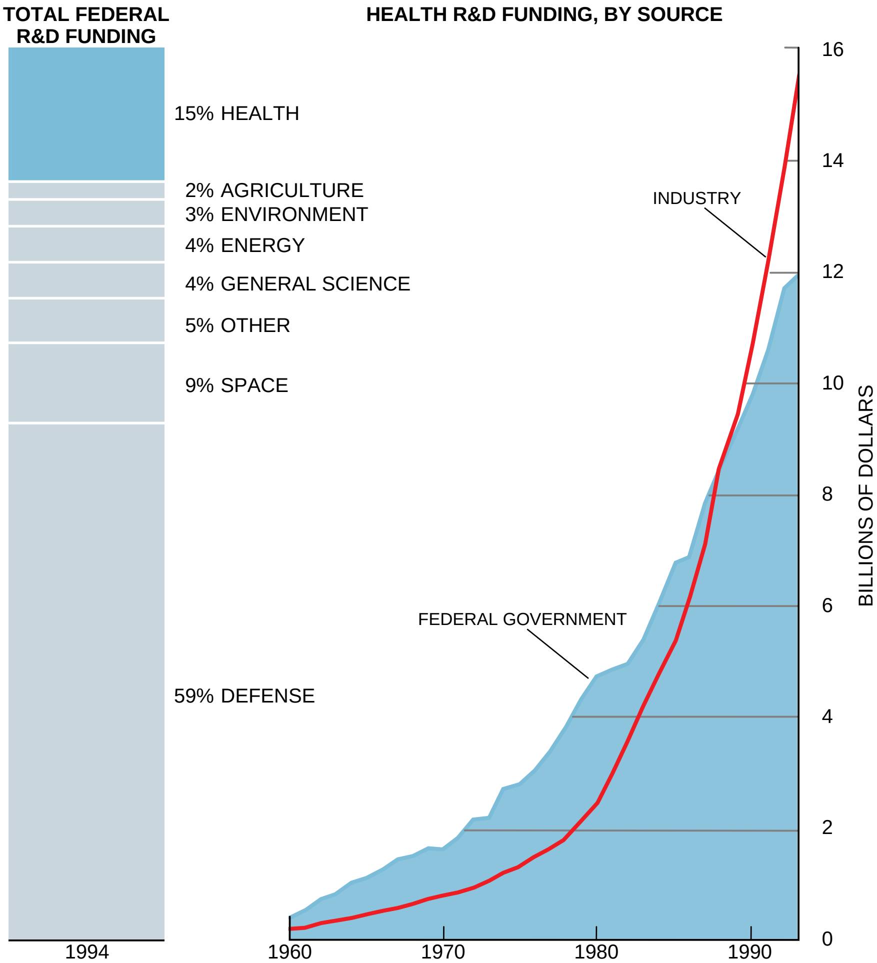
Economic development as well as scholarship suddenly became part of the mission of universities. Government laboratories, too, which had since the 1940s been dedicated to fighting the cold war abroad, were now expected to lend a hand in a different battle, one for national industrial competitiveness.

Universities are embracing their private-sector obligations with gusto. A quarter of all patents awarded to universities since 1969 were granted in 1991–1992 and health and biomedicine were particularly well represented. For some universities, the financial rewards of the new entrepreneurial climate have been enormous. The Cohen-Boyer patent—admittedly an exceptional example—has earned about $100 million in royalties since it was issued in 1980. Profits are split between Stanford University and the University of California. Most university patents are worth far less than that. The Massachusetts Institute of Technology—which is widely considered a pioneer in technology licensing—receives less than 2 percent of its revenue from patent royalties. Yet hope springs eternal.

Industrial research support can mean millions of dollars in overhead payments for an institution. Critics such as Leonard Minsky, executive director of the National Coalition for Universities in the Public Interest, deride the way universities are handling their industrial links. “Universities are corrupt as hell,” he growls. “Most universities have two or three corrupt deals at any one time.”

Many would argue that that is an exaggeration, but anxieties about the propriety of sponsored research keep surfacing. Two years ago Bernadine Healy, then director of the NIH, caused a furor when she criticized a proposed research agreement between the Scripps Research Institute in La Jolla, Calif., and Sandoz, the Swiss pharmaceutical maker. Under the terms of the proposed agreement, Sandoz would have had the right of “first refusal”—the first option to license—for most of Scripps’s research inventions, in exchange for $30 million a year for 10 years. Healy said the proposed agreement could impair the academic freedom of Scripps researchers.

In point of fact, many research administrators did find the proposed Scripps-Sandoz deal troublesome. A survey performed by the NIH confirmed that it was unusual in the degree of exclusivity it allowed Sandoz. The contract was renegotiated so that Sandoz will instead provide Scripps with just $20 million a




*SOURCE: Science and Engineering Indicators 1993, National Science Board*

*SOURCE: NIH Division of Planning and Evaluation, Planning and Policy Research Branch*

**INDUSTRY SUPPORT for health research and development is rising rapidly and overtook the federal government’s contribution five years ago (*graph at right*). Health is the largest category of federally funded R&D after defense, accounting for 15 percent of the total (*bar at left*). (Percentages do not add up to 100 because of rounding.)**

year, in exchange for first-refusal rights to the results of 47 percent of Scripps's research. Scripps receives some $70 million a year from the NIH.

Edward E. Penhoet, president and chief executive of Chiron in Emeryville, Calif., recently co-chaired an NIH panel on sponsored research agreements that was inspired by the Scripps-Sandoz flap. The panel's recommendations were disarmingly published as "considerations" for recipients of NIH research grants who are thinking of accepting corporate sponsorship. The panel's report states that "grantees should ensure that sponsored research agreements preserve the freedom for academic researchers to select projects,...determine the types of sponsored research activities in which they wish to participate, and communicate their research findings at meetings, and by publication and through other means." Penhoet notes that the Scripps-Sandoz deal was unusual in that it gave Sandoz almost blanket ownership of Scripps's research inventions before they had even been devised. The Bayh-Dole Act was, he notes, intended to encourage licensing of existing inventions.

**PRODUCTS of genetic engineering are manufactured on an industrial scale. Fermentation tanks at a Schering-Plough plant in New Jersey produce alpha-2b-interferon, approved in the U.S. to treat hepatitis C, Kaposi's sarcoma and hairy-cell leukemia.**

### Close but Not Touching

The Scripps-Sandoz deal may be exceptional, but universities with close ties to industry are not uncommon. Washington University is a well-known example; over the past 10 years it has received almost $100 million from Monsanto for biological research. A committee consisting of representatives from Monsanto and the university selects projects from proposals put forward by faculty members. Monsanto has the right of first refusal on patentable inventions arising from the work it supports. In fact, the company has chosen to pick up only two ideas from the collaboration for development, making it questionable whether the investment was worthwhile. Originally, Monsanto was willing to pay for "blue sky" research, but it is now focusing on specific areas. It has also decided to reduce its commitment to $5 million a year.

Such links are sensitive. The University of North Carolina, which conducts a lot of research funded by Glaxo, was embarrassed earlier this year when the *News and Observer* in Raleigh reported that the university's president owned stock in Glaxo and that a Glaxo lobbyist was elected to become chairman of the university system's board of governors. Many universities discourage senior administrators from owning stock in companies that fund their research.

Boston University has been criticized for investing some $68 million in Seragen, a biotechnology company founded by a medical school faculty member, John R. Murphy. The university's stake, which amounts to 54 percent of Seragen, is equivalent to some 19 percent of the university's endowment. Seragen returns the vote of confidence by bankrolling research at the university.

Boston University's president John R. Silber says the investment in Seragen was prudent and came out of normal operating expenses. "We spend $70 million a year on medical research for which we get no return," Silber explains. "I couldn't see any reason on earth why we should throw all that away." Others are not convinced. "Institutional conflicts of interest are especially aggravated when a university takes a large position in a start-up company," says Albert A. Barber, special assistant to the chancellor at the University of California at Los Angeles. One danger is that industrial support is hardly a guarantee of stable funding. "The vast majority of academically founded start-ups—like other start-ups—fail," notes Nelson G. Dong, an attorney who specializes in intellectual property at the law firm of Dorsey & Whitney in Minneapolis.

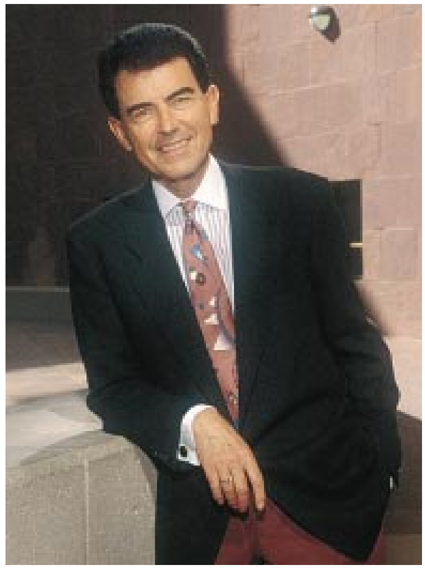
If, however, one moves to the proverbial bottom line, the benefits to society of university-industry collaboration have been immense. Human insulin, interferon and the hepatitis B vaccine all derive in part from university-based research. In all, biotechnology companies have brought about 30 new medical products to market—a boon to humanity as well as to the scientists who developed them and, sometimes, to their investors. Agricultural biotechnology is making slower progress, but its first products—Monsanto's bovine somatotropin and Calgene's engineered tomatoes—are now on sale.

At the same time, the growing corporate presence on university campuses has caused a culture clash. Scholars with industrial support are likely to have more money for equipment and staff than are their colleagues, Dong notes. That can lead to envy, jealousy and misunderstandings. "University people with industry support sometimes feel they have to wear Kevlar buttoned-down shirts," he cracks. (Kevlar is used in, among other things, bullet-resistant vests.)

Little current quantitative information exists about ties between university faculty and the private sector. But research conducted in the 1980s suggests they were extensive even in that early era of biological entrepreneurism. A 1985 Harvard University study found that 47 percent of 800 biotechnology faculty surveyed at 40 research universities consulted for industry. Eight percent owned equity in a privately traded com-

*"We don't want to turn money away, so long as certain guidelines are followed."*
***—HAROLD E. VARMUS,***
*National Institutes of Health*

*"In one sense, universities are obliged to pursue the practical application of federally funded research."*
***—PAUL BERG,***
*Stanford University*



*Biotechnology is "contributing to the economic welfare of the nation, and it has by and large benefited academic progress."*
***—FRANCISCO J. AYALA,***
*University of California at Irvine*

pany that was exploiting their research. Thirty percent of those with industry funding said their choice of research topics had been influenced by the likelihood that the results would have commercial application. Only 7 percent of those without industry support admitted such considerations. So much for following the breezes of pure curiosity.

A survey conducted in 1987 by Linda J. Zimbler of the U.S. Department of Education found that full-time health science faculty members with teaching responsibilities each received more than $88,000 annually from industry for consulting or sponsored research. The sample was small, Zimbler points out.

In light of the continuing growth in the number of biotechnology companies, the number of faculty who have commercial relationships has almost certainly increased since the 1980s. Not only are the links many in number, they take many different forms. Some companies sponsor university-based research through gifts that have few strings attached. Others pay faculty salaries. More common is for a company to sponsor a researcher through contracts that specify the work to be done. And it is not unusual for a company to license a patented technology from a university and maintain a consulting relationship with the faculty inventor.

Companies, particularly small, cash-hungry ones, compensate academics with equity. Moreover, some university departments may do research that is paid for by several different companies. For example, research under way at the University of North Carolina enjoys the support of 40 to 50 different companies, including 10 large corporations, says Thomas J. Meyer, vice chancellor of graduate studies and research at the Chapel Hill campus.

**Serving Two Masters**

Such arrangements inevitably raise questions about how, or whether, a scholar can serve two—or more than two—masters. Berg, who shared a Nobel Prize in Chemistry in 1980 for fundamental research on genetic engineering, fears that universities' new push to transfer technology puts pressure on academics to do work likely to yield practical results. The Bayh-Dole Act, Berg notes, means that "in one sense universities are obliged to pursue the practical application of federally funded research."

The important matter to be decided, as he sees it, is: How aggressively should that be done? "If university technology licensing officers are going around constantly inquiring in research laboratories about what is new and trying to peddle inventions through licensing and so on, I am concerned it is going to distort the path of fundamental research in universities," Berg says. "Young people will be under pressure to do things that bring income to the university."

The research center that Berg directs at Stanford does accept some industrial funding. Berg tries, however, to make sure that investigators choose topics independently of whether the work falls into the area of interest of the commercial sponsor, SmithKline Beecham.

Phillip A. Sharp, who is head of the biology department at M.I.T. and another Nobel laureate, suggests that an increasing emphasis on the practical applications "is in some ways healthy." He would draw the line if universities were to start hiring faculty on the basis of how much money they can bring into their departments. "I hope it won't happen—that would be very unfortunate, and it would be changing the culture of the university," he says. Biomedicine will, Sharp maintains, continue delivering products to improve the human condition as long as the government can see its way to providing strong support for the NIH.

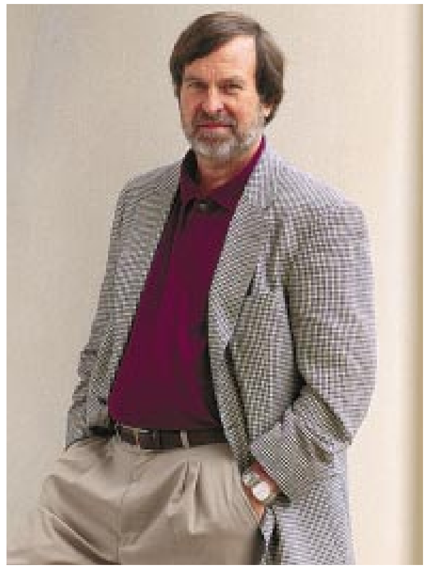
Then there is the problem of secrecy. Universities are normatively places where scholars communicate their ideas freely, unencumbered by the need for confidentiality. Commercial interests, in contrast, inevitably dictate some reticence about an invention until it has been patented. "It will take serious work to preserve the openness of fundamen-

*"I never heard anybody on the industrial side arguing that we should compromise the intellectual process, and nobody on either side wants to do this."*

—EDWARD E. PENHOET, *Chiron Corporation*

*"Unfortunately, the current system often fails to record taxpayer interests in these technologies and to safeguard important public rights."*

—REPRESENTATIVE RON WYDEN, *Oregon*



*"Is it fair that university researchers should take findings produced with the public purse and benefit from them financially?"*

—THOMAS J. MEYER, *University of North Carolina at Chapel Hill*

tal research," Celeste acknowledges. "There will be some challenges for universities and industry."

Not everybody shares Berg's and Celeste's apprehensions. Francisco J. Ayala, a biologist at the University of California at Irvine who is a member of the President's Committee of Advisors on Science and Technology, says he used to be concerned about industry's influence but has changed his mind. "My view of the current situation in biotechnology developments is that it is wonderful," he declares. Biotechnology is "contributing to the economic welfare of the nation, and it has by and large benefited academic progress. The benefits have far outstripped harms that may have occurred."

Bruce M. Alberts, president of the National Academy of Sciences, admits that he would worry if industrial money on a campus meant that some students could not talk about what they were doing. "Communication is part of education," he points out. But Alberts does not believe communication in academia is breaking down yet. "Generally, the system works," he maintains. Sharp is also confident that science is thriving in its new financial environment. "Good science is roaring along, and people are talking about it. First-rate people don't let things get in the way," he insists.

That sentiment is echoed by Frank W. Fitch, a past chairman of the Federation of Associated Societies in Experimental Biology and a pathologist at the University of Chicago. To the extent that secrecy does occur, Fitch speculates, it is probably because some scientists keep their best ideas to themselves so that they can get ahead in the competition for grants and achieve distinction by publishing first. "There have been bastards in science from the beginning," he asserts.

Captains of industry recoil from the notion that they encourage such behavior. They argue that it would not be in their interest to suppress the open communication that is the essence of the academic environment. Penhoet says, "We don't prevent publication even if the results are harmful to Chiron. It's not thinkable. This is a straw man that people set up, and I've never heard of anyone having to do it."

But freedom to publish is, he notes, predicated on the ability to patent valuable ideas. Chiron, along with other companies, does ask researchers it supports at universities to let the firm review draft reports before they are published so that the work can be vetted for patentable ideas. The delay is usually no more than a month or two, Penhoet says. That is not a significant worry, according to Alberts: "Public information is different than talking to colleagues."

A more troubling issue for some academic entrepreneurs is conflict of commitment. Many universities have a rule that faculty should spend no more than one day a week on nonuniversity business. But simply managing time in such a way that everyone gets their due can be a headache, says Leroy E. Hood, a leading molecular biologist now at the University of Washington, who both receives substantial industrial support and has helped found several companies [*see box on next page*].

Management of graduate students is another tricky area for faculty with industrial sponsorship. Alberts is still head of a biochemistry and biophysics laboratory at the University of California at San Francisco, an institution that has been at the forefront of the biotechnology revolution from the beginning. Alberts says a faculty committee was created at the school—as it has been at some other universities—to make certain that students and university funds were not exploited by professors with industry links.

Ultimate dread is reserved for the possibility that unconscious bias by investigators who have financial stakes might influence clinical studies and so affect patients. That fear was raised in 1987 during a congressional inquiry into a clinical trial of tissue plasminogen activator, a drug made by Genentech that dissolves blood clots in heart-attack victims. The late Ted Weiss, a

## Enjoying the Best of Both Worlds

Leroy E. Hood is not just one of the most influential molecular immunologists of the age. He has also made a name for himself as a charismatic and astute businessman who has helped launch several companies commercializing advances in biotechnology.

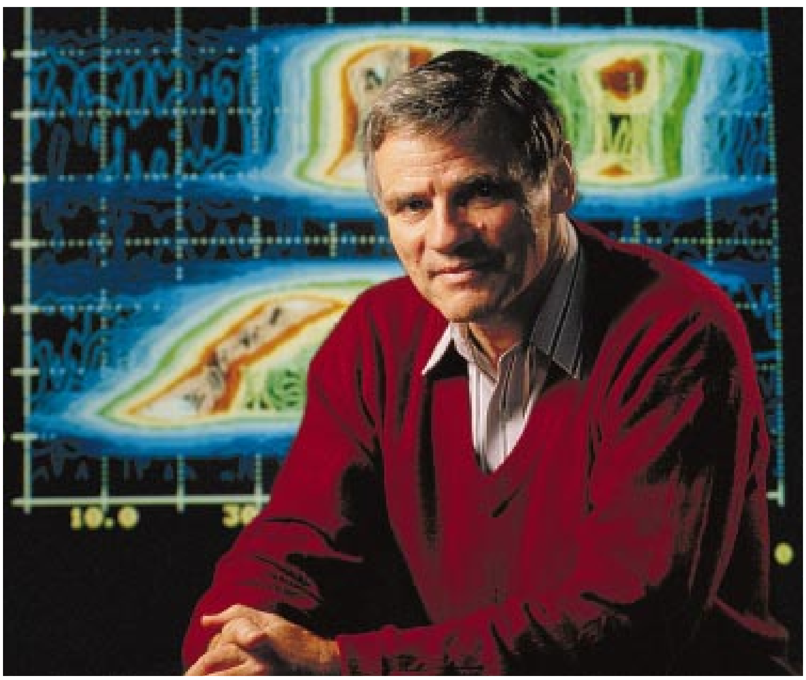
At the California Institute of Technology from 1970 to 1992, Hood devoted much of his time to the challenges of automating the analysis of proteins and, later, DNA. He became a consultant to Applied Biosystems in Foster City, Calif. (now a division of Perkin Elmer), which developed and sold protein and DNA sequencers and other instruments to laboratories all over the world. He also emerged as a prominent champion of the Human Genome Project, the $3-billion international effort to map and sequence all human genes.

In 1992 William H. Gates III, the co-founder and chief executive of Microsoft, gave the University of Washington $12 million to lure Hood there as the head of a new interdisciplinary department of molecular biotechnology. Hood has since persuaded 13 senior scientists to join him. In the new department, computer scientists work alongside geneticists to develop devices for analyzing systems such as the genes that control immune responses and those involved in the early stages of development. A faster DNA-sequencing machine is also on the agenda.

Hood's creative drives seem undiminished. Also in 1992 he, along with Ronald E. Cape, the former head of Cetus Corporation, and several other biotechnology gurus, founded Darwin Molecular in Seattle. Darwin has ambitious plans to discover new drugs using techniques that mimic evolution through natural selection.

Darwin says it is developing proprietary methods that could hugely accelerate the process of identifying small molecules that bind to biological targets. Such molecules may have valuable properties. An important part of the formula is Darwin's expertise in DNA sequencing, which has provided it with numerous targets in the human genome to investigate.

Hood supports the principle of full disclosure as an aid to resolving conflicts of interest. "What you have to do is think carefully about what the nature of the conflict is," he says. "Absolutely the best way to confront these things is to lay it all out up front." He adds: "American industry's strength next century is going to be intellectual. I think we have got to do what we can to foster these developments."



*LEROY E. HOOD is automating the analysis of DNA.*

representative from New York State, embarrassed the NIH by pointing out that more than half of the investigators conducting the trial owned stock in Genentech. Weiss suggested that project managers' financial interests had swayed them to publish a rosier account of the drug's effects than the data warranted.

A more clear-cut abuse occurred at Harvard's Massachusetts Eye and Ear Infirmary in the mid-1980s. A research fellow there, Scheffer C. G. Tseng, allegedly made a large personal profit from selling his shares in Spectra Pharmaceutical Services after he realized that an experimental ointment made by Spectra that he was testing was not proving effective at treating dry eye. Investigations by Harvard found several violations of protocol in the research.

### If At First…

It was Weiss's interest in academic-industry relationships that prompted the NIH in 1989 to propose regulations designed to guarantee that its grantees had no financial conflicts of interest. But the proposals created an uproar. The NIH received 751 letters of complaint and was forced to withdraw its suggestions. The abandoned rules would have required grant recipients to disclose all their financial interests to the government, as well as those of family and even business partners.

The agency published revised proposals this past June. The new approach would require investigators to disclose to their institutions what is termed a "significant financial interest" that would appear to be "directly and significantly" affected by their NIH-supported research. A significant interest could be stock or any kind of remuneration worth more than $5,000 or equity that amounted to more than 5 percent of the ownership of a company.

The regulations give institutions broad discretion on how to manage any conflicts that surface. Administrators could monitor the research to ensure that the financial interest does not bias the project, for example, or require that an investigator relinquish troublesome holdings. Similar regulations for researchers receiving grants from the National Science Foundation were published alongside the NIH proposals. The Food and Drug Administration, too, is expected to issue such regulations in the near future.

The new NIH initiative has certainly been better received than the abortive 1989 effort. "This is an important rule, and we in general endorse it," Fitch declares. The rule was required by a 1993


**POLICING of conflicts of interest by the NIH inspired this barbed new vision from Andy Myer in the *Journal of NIH Research* this past June.**

law governing the NIH's bureaucratic parent, the Public Health Service. Congressional interest in research conflicts remains strong. "We are trying to protect the public interest," says Steven C. Jenning, an aide to Representative Ron Wyden of Oregon, who has taken up the torch of preserving research integrity. Wyden recently held a hearing in which he charged that the NIH has failed to ensure that products arising from federal research are made available to the public at "reasonable" prices.

Not everyone is happy with the NIH's solution to financial conflicts. One potential difficulty concerns how exactly a university should determine whether a particular stock holding is likely to be directly affected by an investigator's project. Martin L. McGregor, a partner in the law firm of Gardere & Wynne in Dallas, Tex., and a co-chair of the Biotechnology Industry Organization's law committee, says he would like to see the rules clarified so that universities can be sure their certifications will not be second-guessed later. "The definition of significant financial interest is overly broad," McGregor complains. "Once anything becomes a problem, potential investors are going to move elsewhere. If you make regulations that make it more difficult to commercialize, public funding in this area fails."

Most players in the biomedical research business have, however, decided, along with the NIH, that disclosure is the best way to take care of conflict of interest. In 1984 the *New England Journal of Medicine* started to ask researchers publishing in its pages to declare any financial interests in companies likely to be affected by their results. Scientific journals such as *Science* and *Nature* followed suit. (*Scientific American* also asks its contributors to mention their relevant private-sector interests.) Many universities already have in place conflict-of-interest rules somewhat like those that the NIH will be insisting on.

What is quite unclear is whether it will ever be possible to satisfy all the conflicting demands on biological entrepreneurship. Institutional conflicts are in some sense unavoidable once academics start going into business, because universities are then trying to do two different things at once: to provide havens of scholarship and also to make a buck. The NIH regulations recognize the inherent ambiguity in the new role of universities by referring to the need for "balance" between competing interests. "All the things you do to control conflict of interest will ultimately result in inhibiting technology transfer," says Barber, who is studying technology transfer and conflicts of interest for the Government-University-Industry Research Roundtable.

Barber believes financial involvements receive more attention than they deserve. "The problem is that conflict of interest is almost a pejorative term," he points out. "People equate it with scientific integrity and misconduct, but the troublesome cases are very, very few." Nevertheless, he thinks the NIH has done the right thing in putting the onus on institutions to monitor their employees' interests.

Meyer of the University of North Carolina agonizes over how his institution will meet the government's charge to boost the economy while keeping the faith with traditional academic values. "We have to face up to developing arrangements and methods to maintain the university's integrity in the public eye and to keep the highest ethical standards we can," he says. But at the same time, "these charges to commercialize research are real ones." The basic dilemma, as Meyer sees it, is: "Is it fair that university researchers should take findings produced with the public purse and benefit from them financially? The best method we have for taking something to the market as quickly as possible is the profit motive."

Hood, the academic entrepreneur at the University of Washington, turns the dilemma on its head. "We scientists have a major obligation not only to discover things but, where possible, to put them to use in society," he argues. "You can transfer knowledge in the form of an instrument or technology or patents, or alternatively you can transfer knowledge in the manner of creating companies. We owe society an obligation because society is paying for science, and this is an obligation that scientists are just starting to take seriously."

Everyone, of course, wants to see new medicines and other wonders come out of the laboratory and into the market. The possibility troubling some observers is whether the rush to commercialize is making it harder to do fundamental research that cannot immediately be turned into profit. Hood is not complacent. He charges that scientists are "guilty of not spending enough time justifying fundamental science." He frets that government, like industry, is taking an increasingly narrow view of what kind of science it will pay for. And he doubts whether industry's underwriting of biology will continue to grow as it has in recent years.

"Industry support is increasingly going to be short-term and very, very pragmatically focused," Hood warns. The big leaps in science, in contrast, usually happen when good scientists are allowed to follow their imaginations. Varmus notes that much of the research that industry funds in universities is basic in nature. But, he admits, "You have to be concerned."

FURTHER READING


BIOMEDICAL RESEARCH: COLLABORATION AND CONFLICT OF INTEREST. Edited by Roger J. Porter and Thomas E. Malone. Johns Hopkins University Press, 1992.

DEVELOPING SPONSORED RESEARCH AGREEMENTS: CONSIDERATIONS FOR RECIPIENTS OF NIH RESEARCH GRANTS AND CONTRACTS. National Institutes of Health, Office of Science Policy and Technology Transfer, in *Federal Register,* Vol. 59, No. 122, pages 32997–33001; June 27, 1994.

OBJECTIVITY IN RESEARCH: NOTICE OF PROPOSED RULEMAKING. Department of Health and Human Services, Public Health Service, Office of the Secretary, in *Federal Register,* Vol. 59, No. 123, pages 33242–33251; June 28, 1994.

THE FUNDING OF YOUNG INVESTIGATORS IN THE BIOLOGICAL AND BIOMEDICAL SCIENCES. Commission on Life Sciences, National Research Council. National Academy Press, 1994.

## Fighting for Survival

*Peace has fizzled for defense contractors*

Early in 1993 President Bill Clinton proposed softening the impact of defense cutbacks with programs to help contractors develop products for civilian markets. But after nearly two years and several billion dollars of federal funds, plowshares are not proving as profitable as swords. In fact, there are rather few plowshares to be seen.

Although many defense contractors have attempted to find nondefense projects, the results have been underwhelming. "Success to date has been very limited from the perspective of new jobs and economic development within the defense community," says Don Nakamoto, research director of the International Association of Machinists and Aerospace Workers, District 725. Environmental technologies—which seemed to offer a promising market for defense contractors because they involve complex challenges, often on a large scale—have created some employment. Yet the numbers are counted in the dozens or hundreds, not in the thousands.

The National Commission for Economic Conversion and Disarmament, a nonprofit organization in Washington, D.C., reports that defense-related jobs are disappearing at an ever quickening pace. In the first half of 1994 alone, 115,000 defense-related manufacturing jobs were cut—close to the total of 164,000 such jobs lost during the whole of 1993. The commission has recorded the loss of 728,000 positions between 1990 and mid-1994, most of them in California. And there is probably worse to come. Defense Secretary William J. Perry has warned that the defense industry base will contract to a half or even a third of its size in the mid-1980s. If so, the defense budget would fall from $249 billion this year to less than $190 billion (in 1994 dollars).

Instead of trying to survive the economic squeeze by converting to civilian-related manufacturing, most companies have opted to merge with or acquire part of a competitor. The new hybrid is then slimmed down. The emergence of Lockheed Martin out of the traumatized remains of Lockheed and Martin Marietta is the most notable recent example. Each partner in the union laid off about 8,000 workers during 1993, according to Gregory A. Bischak, executive director of the National Commission for Economic Conversion and Disarmament. Bischak predicts that some 15,000 additional Lockheed Martin workers will lose their jobs in the inevitable housecleaning.

Several U.S. defense contractors have tried another strategy: cope with the cuts at home by moving into the export market. The share of American arms being sold in developing countries has soared, Bischak's commission reported earlier this year, from 10 percent a decade ago to 60 percent today.

MARTIN MARIETTA CORPORATION

*MARTIN MARIETTA, now part of Lockheed Martin, has found civilian markets for its rockets. This Atlas I launched a weather satellite this past April.*

But what defense contractors have not done, for the most part, is "to develop commercial products, invest in retooling and marketing to become competitive in commercial areas, or retrain and redeploy their workforces," the commission concluded. That assessment is echoed by Daniel Flaming, president of the Economic Roundtable in Los Angeles. Flaming notes that a study conducted by the organization shows that aerospace manufacturers in the Los Angeles region have become more dependent on defense funds since 1991, not less, because of a slowdown in the civilian aviation industry. "My belief is that it is indeed a very bleak picture," Flaming says. "We have virtually nothing you can point to showing that the industry is being reconfigured in an organized way."

Companies that traditionally made their way in the world by supplying the Defense Department with low-volume, specialty products find themselves ill suited to civilian markets, where beating down costs is the name of the game. "They are quasi-nationalized organizations," Flaming argues. Only small and medium-size companies with fewer than 500 employees seem to have the flexibility to make the changes needed to move into a new line of business—and they make little difference to the overall job market.

To be sure, some defense contractors have made serious efforts to change their spots. Hughes Aircraft stands out: the company, which is owned by General Motors, has thrown itself into the television satellite business. Hughes is involved in $75-million worth of projects initiated under the federal Technology Reinvestment Program—out of a total of $190 million spent to date. But few other companies have been as energetic. Bischak's commission blames the administration for not providing "appropriate incentives and resources.” Nakamoto cites a different factor: “Each believes it is the best and thinks it can be one of the handful of survivors.”

Environmental challenges have provided a new business venue for some corporations that have traditionally relied on the Pentagon. Westinghouse, for example, has moved into the business of cleaning up other contractors’ polluted sites. Northrop is designing a bus that operates on fuel cells for a buyers’ consortium. Calstart, a group of California contractors, is developing a range of technologies for electric vehicles. And Raytheon has won a contract from the Northern Illinois Regional Transportation System to investigate “personal rapid transit,” a low-pollution rail-based system in which the carriages take passengers directly to their destinations.

Still, environmental projects are the exceptions, not the rule. Curtis Moore, the co-author with Alan Miller of a recently published book on environmental technologies, *Green Gold* (Beacon Press), blames the government for that missed opportunity. Moore charges that U.S. environmental standards have become increasingly lax in comparison with those of its competitors. That trend, he argues, translates into a dearth of opportunities for American companies. Moore points out that the Clean Air Act amendments of 1990, for instance, gave the U.S. 12 years to bring about pollution reductions less dramatic than those that Germany achieved in six years. As a result, “there is not a market for this stuff in the U.S.,” he says.

Bischak acknowledges the current shortcomings of defense-to-environment conversion: “The bottom line is we are only spending $4 billion a year in cultivating this market.” The recession, strongest in California, has merely increased resistance to stricter environmental regulation. There is growing pressure on that state to eliminate its mandate to manufacturers to sell automobiles with zero emissions of pollutants by 1998. The Big Three automakers—Chrysler, General Motors and Ford—are bitterly opposed to the emission-reduction plan, and if it should be abolished, electric vehicles will look less attractive as development projects.

Some companies have even openly declared that conversion is a dead end. McDonnell Douglas, one of the biggest defense contractors, has been selling off peripheral businesses in information systems and financial services. “We have no intention of diversifying into new and unfamiliar businesses,” the company told shareholders earlier this year.

Despite the generally poor showing to date, many other firms are still hopeful about prospects in civilian markets. Although Norman Augustine—who was head of Martin Marietta and may eventually lead Lockheed Martin—is a well-known cynic about ill-conceived attempts at conversion, Lockheed Martin reckons that 40 percent of its sales will be in civilian markets. The company should be able to capture 20 percent of the global satellite-launch market, according to one estimate.

Bischak, too, sees opportunities ahead. “We’re perched on the edge of a boom in shipbuilding that should occur in the next 10 years,” he predicts. A new government program, Ameritech, could help U.S. companies ride that wave, he believes. But so far conversion has fallen short of providing desperately needed jobs. —*Tim Beardsley*

## A New View for Surgeons

Most people have no use for infrared night-vision goggles or a laser-targeting system, which is why finding a fit between defense know-how and the civilian world is often difficult. But one new project might bring some high-technology military imaging capabilities to a wider community: that of physicians and their patients.

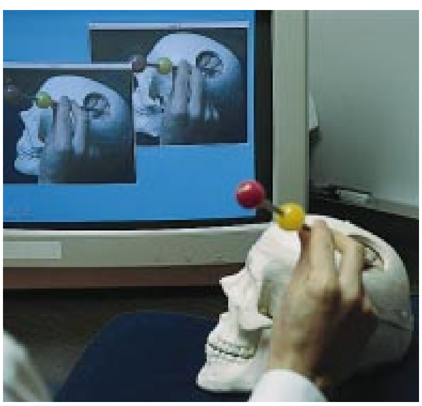
A navigation system that could provide surgeons with a three-dimensional, real-time, see-through view of a surgical site is being developed by the Cleveland Clinic Foundation, with the help of Picker International, a medical equipment manufacturer, and the U.S. Air Force’s Materiel Command. The device could be a lifesaver for patients undergoing difficult operations.

Gene Barnett, a neurosurgeon at the clinic, and his associates have already developed a fixed-image version of the apparatus to help in brain surgery. Before the operation, an image of the patient’s brain is produced in a supercomputer by combining the results of three standard techniques—magnetic resonance imaging, computed x-ray tomography and digital subtraction angiography. During surgery, the computer monitors the positions of the surgical instruments. It does so by timing how long it takes ultrasonic pulses to travel from tiny “beacons” mounted on the instruments to nearby microphones. The computer “knows” the shapes of the instruments, so it can superimpose moving images of the tools onto its stored representation of the brain. The resulting composite is displayed to the surgeon.

At present, the apparatus cannot be used on most other parts of the body. Because breathing displaces tissues, a fixed image obtained before surgery does not remain accurate. Enter James D. Leonard of Wright-Patterson Air Force Base, an expert in “sensor fusion”—the art of combining images generated by different instruments. Leonard hopes to add ultrasound scanning to Barnett’s existing device. The technique would provide a real-time image of the area that reflects tissue movements. Then comes the hard part: data from the ultrasonic scan would be used to “bend” the high-resolution image stored in the computer into the actual shape of the tissue.

The fixed-image equipment currently available uses a television screen mounted next to the surgical site. But Leonard has a better idea for this limitation as well. He wants to experiment with helmet-mounted displays similar to those used by fighter pilots. The computer-generated image would be projected inside the headgear, so that a surgeon would not have to look away appreciably from the work site in order to see the image.

After that technical tour de force, equipping the computer with a voice-recognition system would be easy. A surgeon would then be able to use the equipment without having a computer operator present. “Sensor fusion is not a solved problem,” Leonard cautions. Still, several potential partners for a development effort have expressed interest, including Loral, a defense contractor in Arizona. Leonard thinks the project should be doable in five years. —*Tim Beardsley*



GENE BARNETT *Cleveland Clinic Foundation*

***MILITARY COMPUTER know-how could aid surgeons in the operating room.***

## Productivity Lost

*Have more computers meant less efficiency?*

It has been billed as one of the greatest macroeconomic mysteries in recent times. American corporations have invested enormous sums of money in information technology, supposedly in exchange for streamlined procedures and speedier transactions. Yet such spending seems to have bought very little by way of lower business costs or greater efficiency.

This so-called computer paradox, which has perplexed executives and economists for years, has suddenly taken on new life. Improvements in the economy suggest that old explanations for the schism between computers and productivity need to be revised. The former favorite culprit—shortcomings inherent in the measurement of economic output—is being replaced by a more ominous one: mismanagement.

The computer conundrum first emerged in the 1980s. Expenditures on information technology in that decade rocketed into the billions of dollars. At the same time, white-collar services—such as banking, finance, health care, insurance, telecommunications, advertising and retail—increased their productivity by a meager average of 0.8 percent, well below their historical norm of 2.5 percent a year. Was it coincidence that during that same decade these companies spent more than $860 billion on computer hardware?

The most optimistic and oft-heard answer to this question has run more or less as follows. Because services are primarily intangible goods, they are harder to tally than, say, cars or washing machines. Hence, computers undoubtedly improved productivity in the service sector, but the impact has been underestimated.

Some economists and managers now seem less willing to put blind faith in measurements unseen. The economy has changed for the better in the past two years. Nonfarm output per worker rose a remarkable 4.9 percent—its biggest leap since 1976. “To me, the current improvement is a clear confirmation that there is something troubling in the computer paradox,” says Stephen S. Roach of Morgan Stanley, who has long criticized the measurement explanation. “Had information technology been so productive all along—as the measurement apologists felt—we would not be seeing a continuing upheaval in the services.” He credits recent gains to painful reforms in the service sector, including the six airline companies and scores of Fortune 500 managers downed on the path to recovery.

Daniel E. Sichel of the Brookings Institution in Washington, D.C., has dismissed the measurement gap using a different tack. “Qualitatively, I think the point is right that computers do produce a lot of unmeasurable output,” he says. “But it has been too easy to overstate the importance of the measurement story.” He notes that computer hardware made up only 2 percent of the plant equipment that businesses used in 1992. So even if computers did pump up productivity in certain unmeasurable industries, their contribution to the economy would remain too small to resolve the paradox.

In addition, Sichel points out that many of the idea-based products service companies yield, such as consultation or software, are purchased by other businesses. Because the GDP counts the value of goods and services only in their final stage of production (so as to avoid recording anything twice), these intermediate products are not included. Therefore, even if output in the service sector had been grossly underestimated during the 1980s, the impact on productivity would have been minimal.

Although measurements are hard to blame, managers are not. "Some people are now saying of the services, 'Well, maybe it just took these guys 20 years to learn to use their equipment,' " reports Lester C. Thurow of the Massachusetts Institute of Technology. "But the operative word at the moment is maybe." Thurow, like Roach, has long dismissed the measurement gap as myth and was one of the first to suggest that American managers had failed to make good use of information technology.

Service executives would not be the first to falter in the face of inventions. Thurow observes that a similar delay occurred when electricity first arrived on the factory floor. That breakthrough did not boost productivity until some two decades later, when manufacturers designed and built factories keeping electricity in mind. (The older factories, electrified or not, had placed their machines closest to steam engines, waterwheels, shafts and rods.)

"Unwittingly, as managers leaped into the information age, they spent more than they could afford," Roach comments. "Companies today probably have 10 times as much processing power as they need." Roach feels that in the 1980s most service managers "overmipsed" their back offices—that is, they endowed them with equipment and capacity they did not need. Now they are beginning to rid themselves of the inefficiency such excess created. Roach points to mergers such as the one between Manufacturers Hanover and Chemical banks as attempts to pare down the bulk.

Information technology may have delivered more lasting woes as well. "Courtesy of the computer, the nature of the cost structure in service companies has changed," Roach says. Formerly, the main assets in service companies were people. As a result, services could adjust their expenses quite easily by refiguring the salaries of their workers. Today these companies are increasingly encumbered by less negotiable costs, such as those required to maintain support staffs and upgrade equipment. This shift, Roach argues, has left them less flexible and less fit to compete.

"The lack of payback from investments in information technology was putting more and more pressure on the user community to get their act together," Roach concludes. "The productivity gains we are seeing right now come from that." What the resolution to the paradox will be, no one can say, but surely fewer managers will continue to believe that what you can't measure can't hurt you. *—Kristin Leutwyler*

## Pricing Internet

*Tolls may prevent traffic jams on the data superhighway*

One recent Wednesday at the Massachusetts Institute of Technology, half a dozen network analysts milled around an office stuffed with computer terminals, network cables—and a three-foot iguana named Iggy. Iggy happily devoured a kiwifruit under the hot spotlights of a video camera. He then crawled up a stick, under a sign reading "The Iguana Channel," and relieved himself. The camera captured the moment and fed its frames to one of the workstations. The video was digitized, compressed and then broadcast to some 20 viewers in seven countries.

The Iguana Channel was not a pay-per-view event. It was aired free on the Internet, the information highway on which tens of millions of users swap more than 15 trillion bytes of data monthly. The footage of Iggy ate up twice as much Internet bandwidth as many colleges supply to their entire campus.

If you love iguanas and their antics, this broadcast may seem an entirely proper use of the Internet, most of whose users pay fixed monthly fees for their connections, regardless of their activity. But a number of economists warn that unless people are charged for the services they request, pranks and joyriding will tie up legitimate traffic. A few have built simulations of the Internet to test various pricing schemes.

Scott Shenker and his colleagues at the Xerox Palo Alto Research Center noticed that there are at least two distinct types of Internet services: real-time services, such as videoconferencing, which require a steady connection with minimal delays, and best-effort services, such as E-mail, which can tolerate slower delivery. For both to work well, real-time data must be given priority over ordinary best-effort traffic. But what is to prevent users from cutting in line by mislabeling their E-mail as a video?

Shenker's group hypothesized that a priority pricing scheme that charged more for real-time service would motivate users to be honest. In simulations of various network shapes, speeds and traffic conditions, the team found that there is a range of normal and premium prices that "will induce each user, out of her own self-interest, to ask for the service class that maximizes the network's overall efficiency."

Of course, prices must rise and fall with demand. Shenker suggests that some administrative body use historical evidence to predict peak periods and

CHRIS MURPHY *Massachusetts Institute of Technology*

***IGGY starred in a 24-hour-long video broadcast live over the Internet.***

set prices accordingly, much as telephone companies do. But experiments conducted by Hal R. Varian of the University of Michigan indicate that "delay fluctuates widely across times of day… but follows no obvious pattern."

Varian proposes letting users bid for priority by attaching the highest prices they are willing to pay to their requests for service. In this "smart" market the network switches would continually adjust the price of admission based on the level of congestion. The smart market idea has its drawbacks, however. It cannot guarantee delivery time, so a surgeon directing a remote operation might find her video feed slowing to a crawl if congestion peaked suddenly.

Smart agents—that is, mobile computer programs dispatched by a user to collect or control information—would act as brokers in a pricing scheme developed by Alok Gupta, Dale O. Stahl and Andrew B. Whinston of the University of Texas at Austin. The agents would help users shop around for the best price by soliciting estimates of the current cost and delay from various network servers.

In a simulation of 50 servers and 100 services, Whinston and his colleagues calculated how long users would have to wait for a request to be fulfilled under the current free-access policy and under their pricing mechanism. "You get tremendous benefits by introducing pricing," Whinston claims. Under heavy traffic, for example, he estimates that data taking about a minute to traverse a pay-as-you-go network would take more than an hour to emerge from a fixed-fee network like the Internet.

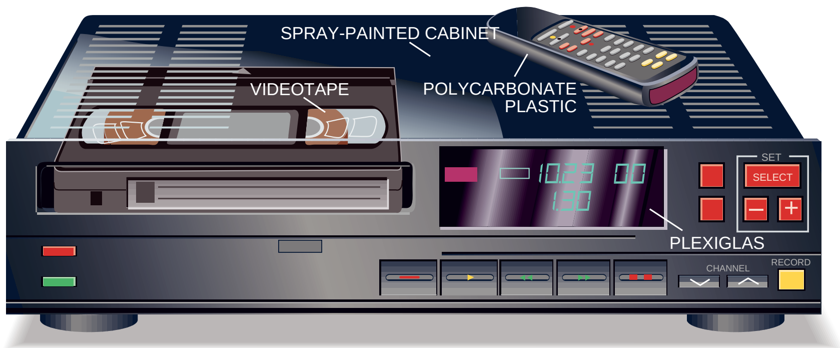
Although most economists seem to agree that some kind of priority pricing could improve the efficiency of the Internet, many details remain unresolved. The NSFNet backbone alone transports more than 40 billion packets of data a month—and that number is doubling every year. An accounting system that can track and bill a trillion entries annually could prove to be too costly.

By next spring the National Science Foundation plans to hand over control of the NSFNet, which handles 75 percent of U.S. Internet traffic, to Pacific Bell, Sprint, Ameritech and Metropolitan Fiber Systems. They will probably charge for use as well as access. After that, Whinston says, "the Internet will become an enormous economic system of a kind unparalleled in history. Tariff issues will become very important—and very unclear." *—W. Wayt Gibbs*

## Ounce of Prevention

### *Cleaner chemicals pay off, but industry is slow to invest*

When chemists gathered last year at the first symposium sponsored by the American Chemical Society and the Environmental Protection Agency to discuss ways of redesigning chemicals to produce less pollution, they heard many good ideas but few good examples. In August the




IAN WORPOLE

*VIDEOCASSETTE recorder incorporates many materials for which cleaner, cheaper alternatives now exist. Supercritical carbon dioxide, rather than harmful solvents, can be used to spray-paint the case and to make the plastic cover for the display. A water-based magnetic tape formulation could cut air pollution. And careful temperature control can produce more recyclable polycarbonate plastics without toxic solvents.*

| MATERIAL | POLLUTANT | REDUCTION WITH CLEANER METHOD |
|---|---|---|
| SPRAYED COATINGS | VOLATILE ORGANIC COMPOUNDS | 80% |
| MAGNETIC TAPE | VOLATILE ORGANIC COMPOUNDS | 100% |
| PLEXIGLAS-LIKE PLASTICS | HAZARDOUS WASTEWATER | 100% |
| POLYCARBONATE PLASTICS | PHOSGENE, METHYLENE CHLORIDE | 100% |

sequel proved more impressive. Cleaner, cheaper techniques for making drugs, spraying paint, manufacturing plastics and coating magnetic tape were announced. There was a bonus: these processes promise not only to reduce hazardous waste but also to boost profits.

The idea of developing chemicals to minimize pollution rather than simply to get the biggest bang per buck is still new to many industrial chemists. The EPA and the National Science Foundation have funded such research only since 1991—and the $6 million or so they have doled out pales next to the $4.4 billion that the chemical industry spends every year to control, treat and dispose of the 638 billion pounds of hazardous waste it produces.

If the recent crop of innovations is any indication, however, a little research can go a long way. "This can have a dramatic bottom-line payoff," asserts Daniel J. Watts, executive director of the Emission Reduction Research Center at the New Jersey Institute of Technology. Last year, for example, Sandoz Pharmaceutical gambled $2.1 million on a joint research project with Watts's center to develop a recyclable alternative to a solvent used in manufacturing its number-two drug. The investment cut the 20 pounds of hazardous waste that Sandoz has been generating for each pound of its drug down to just 1.7 pounds, saving $772,000 a year in disposal and operating costs.

Basic chemical manufacturers, because they have slimmer margins and larger volumes than drugmakers, may stand to profit more from using cleaner compounds. Last year the EPA helped a Du Pont factory cut the pollution created by 15 of its most wasteful processes. After investing $6.7 million in new equipment, Du Pont now saves $1.6 million in disposal fees, $2.2 million in reduced materials costs and $11.1 million in increased capacity—every year.

Computerized expert-design systems may make such success stories more common. A research team led by George Stephanopoulos, a chemical engineer at the Massachusetts Institute of Technology, is completing a system that can warn engineers when their plans would generate lots of waste. The program can also suggest cleaner alternatives. In the meantime, other recent innovations promise to crank emission flows down a turn by improving routine manufacturing materials or techniques.

American industries spray 1.5 billion liters of paints and coatings every year to finish products, according to Marc D. Donohue of Johns Hopkins University. Each liter of paint pumped out of a spray gun releases an average of 550 grams of volatile organic compounds (VOCs)—air pollutants that can increase the production of toxic, ground-level ozone. "More VOC emissions result from painting a car than from the engine exhaust over its lifetime," Donohue points out. Much of this pollution could be avoided. Donohue has worked with Union Carbide to commercialize a technique that eliminates up to 80 percent of the VOCs needed for such sprays.

Union Carbide's alternative dissolves supercritical carbon dioxide—gas that is pressurized until it takes on some properties of a liquid—into the paint solids, rather than dissolving the solids in a VOC solvent. When pumped through a sprayer nozzle, the $CO_2$ depressurizes like a shaken soda, atomizing the paint into an umbrella-shaped mist of even-size particles just perfect for painting.

Donohue claims that "coating quality is so much better that you can use just two coats where before you would have used three." A strong market incentive lies in that 33 percent reduction of labor. "This technology is being adopted not because it prevents pollution but because it saves money," Donohue observes. His assertion is supported by the fact that companies switching to or testing supercritical $CO_2$-spraying equipment—Ecco, GM and Ford, among them—have not leaped to advertise their environmentally friendly gestures.

Supercritical $CO_2$ may find an equally important role in plastics production. Joseph M. DeSimone and his colleagues at the University of North Carolina at Chapel Hill designed a surfactant that allows high-quality Plexiglas to grow in a bath of supercritical $CO_2$ mixed with a stabilizer. Harmful wastewater is eliminated. Although he has yet to demonstrate it, DeSimone thinks the same or similar surfactants should enable cleaner production of common plastics such as polystyrene, PVC and various acrylics.

Supercritical $CO_2$ does not, however, provide a perfect solution. It requires a modest capital investment in high-pressure equipment, and the gas is implicated in global warming. In some cases, researchers are finding, clever chemistry can yield an even cleaner and cheaper alternative solvent: water. David E. Nikles of the Center for Materials for Information Technology at the University of Alabama has developed a water-based formulation for magnetic tape that could help this $14-billion industry meet regulations mandated by recent amendments to the Clean Air Act. The rules will force U.S. tape manufacturers—who currently supply one third of the world market for video, audio and computer tape as well as floppy disks—to reduce emissions of air pollutants or to move elsewhere.

In such a competitive market, Nikles says, it is not enough for an innovation to be clean, cheap and simple; novel technology must also improve the performance of the product. Nikles believes he has met three of those goals and will soon achieve the last. In a pilot trial conducted with Graham Magnetics in Graham, Tex., Nikles's VOC-free process produced videotape at 15 percent lower cost than did a solvent-based process, without slowing the production line. In a typical plant, Nikles estimates, the new process would generate about 10 kilograms of VOC emissions as opposed to the usual 390,400 kilograms. The project has piqued interest among both tape makers and the companies that supply them with raw materials. “Air Products and Miles are already trying to position themselves as suppliers for waterborne processes,” Nikles observes. “They see this as a long-term strategy.”

U.S. chemical companies that ignore the trend risk falling behind their competitors overseas. Kiosuk Komiya of Asahi Chemical reported at the August symposium that his firm has developed a cleaner industrial process for producing polycarbonate. About one million tons of this heat- and impact-resistant plastic are manufactured worldwide every year. The material is incorporated into everything from nursing bottles to VCRs. Unfortunately, almost all polycarbonate is produced using large quantities of methylene chloride (a suspected carcinogen) and phosgene (a lethal neurotoxin). By carefully controlling the reaction temperature, Asahi has eliminated the need for solvents and produced polymers that are easier to recycle.

Michael J. Wallace, who leads the pollution prevention committee at Sandoz, predicts that the savings reported by projects such as his will motivate chemical companies to invest in environmentally benign research. So far only a handful have shown real interest. A 1992 study of 181 hazardous-waste reduction activities at 27 plants by Mark H. Dorfman found that nearly all focused only on improving the efficiency of existing production lines. “They didn’t ask what can we change about this process,” says Dorfman, who is senior associate at INFORM, a nonprofit environmental research organization.

Some research may deliberately be hidden to avoid tipping off the competition. And Dorfman admits that in recent updates to his survey, “more seems to be happening.” But this year in New Jersey, fewer than half of the chemical companies required to submit pollution prevention plans to the state complied with the order, Wallace reports.

“It’s the old story of overcoming paradigms,” he says. “People have always thought that environmental control costs money. We’re just becoming aware now that it’s not only environmentally sound, but it also makes a heck of a lot of business sense.” But Dorfman, citing obvious efficiency improvements that chemical companies have overlooked for decades, is less convinced that self-interest will suffice. “It’s laudable that the EPA has given that a try,” he says, “but I think it is time to require a little more accountability from industry.” —*W. Wayt Gibbs*

# Bettering Batteries

*Lithium cells are becoming safer and less weighty*

No battery's perfect. But a smaller, lighter, cheaper one would be ideal for the next generation of electronics. Which is why chemical engineers have pinned their hopes on lithium, the third lightest atom after hydrogen and helium. The metal provides greater energy per volume than do the traditional battery standards, nickel and cadmium. Although dogged by a few safety problems, the buoyant element has already found its way into some familiar batteries, including those powering portable computers. And now it looks as though lithium-battery technology is undergoing a significant advance, one that may begin to address safety concerns.

Much lithium-battery research has centered on what is termed the lithium-ion system. These cells contain an anode (the negative electrode of a battery) in which lithium ions are enmeshed within the lattice structure of graphite during a process called intercalation. The ions typically shuttle between the anode and cathode (or positive electrode) through a liquid solution known as an electrolyte.

Although the carbon that houses the lithium and the container that prevents the toxic liquid from escaping add unwanted weight to the batteries, both are needed to ensure safety. Electrolyte leaks can leave lithium exposed to the air, where it reacts rapidly with water vapor. Proper casing prevents such reactions while protecting users and the environment from exposure to solvents. Sequestering lithium by intercalating it within carbon further decreases its volatility.

Advances made in an alternative system, the lithium-polymer battery, could obviate these issues. Such cells, which have largely been the focus of North American and European research efforts, incorporate an anode of pure lithium foil and a solid-state plastic electrolyte. Without the carbon required for intercalation or the bulky liquid containers, total battery weight can be cut significantly. Although preventing pure lithium from reacting with the air remains crucial to guaranteeing safety, the solid-state electrolyte reduces such risk. As May-Ying Chu of PolyPlus Battery Company explains, the electrolyte "won't drain out if you spring a leak," making exposure of the metal and of a dangerous electrolyte in the case of cell rupture unlikely.

These advantages suggest that the future of lithium batteries most likely lies in a combination of lithium-polymer and lithium-ion technology. In such an integrated battery, intercalated lithium-carbon at the anode would be surrounded by a solid electrolyte. Intercalated lithium-carbon electrodes can be fairly inexpensive when mass-produced because carbon is affordable and readily available. Such electrodes combined with a solid electrolyte would almost completely prevent lithium from contact with air and would avoid leakage.

Of course, intercalated lithium-carbon anodes would increase the weight of the cells. But Chu notes that the required carbon-to-lithium ratio continues to fall, making these anodes more reasonable to include in a solid-state battery.

The weight of a combination lithium-polymer/lithium-ion system is not the only hurdle researchers have to overcome. Lithium moves through solids very slowly—a problem for any battery with a solid-state electrolyte. Investigators at Bell Communications Research, under the direction of Jean-Marie Tarascon, have demonstrated a potential solution. The group devised a lithium-ion cell with a semisolid plastic electrolyte. Tarascon says that although the plastic is half-liquid, it feels similar to a telephone cord: dry yet flexible. Adding a solvent that remains trapped within the electrolyte allows the rapid flow of lithium ions while preserving the safety features of an all-solid electrolyte, he explains. Tarascon describes his team's findings as more of an engineering and processing breakthrough, and one that happened quickly. "After several years of studies on various components of the cell, everything fit together within three or four months," he comments.

Whether lithium batteries become widely used remains to be seen. Despite the recent laboratory successes of advanced lithium batteries, the technology retains only a small portion of the battery market: 0.5 percent. Lithium batteries may last longer than do traditional nickel-cadmium cells, but they can be nearly three times as expensive—so consumers not factoring in lifetime savings may steer away from the new batteries. Although people may be willing to spend slightly more for a lighter lithium battery in a portable computer, the extra cost seems unreasonable for a Dustbuster or an electronic toy. Indeed, Brian M. Barnett of Arthur D. Little believes "lithium batteries will never completely replace nickel-cadmium."

Barnett further cautions against setting expectations too high. "The technology we haven't built always looks better than what we already have," he advises. Nevertheless, because of their high energy density and safety, solid-state lithium cells remain the front-runner in the long-term race for the Holy Grail: a cell that could safely power a car for 300 miles on one charge. The U.S. Advanced Battery Consortium—a partnership between the Big Three automakers, the U.S. Department of Energy and the Electric Power Research Institute—has made more than $90 million available to researchers investigating lithium technology for electric vehicles.

Clearly, such applications will remain a challenge well into the next century. By all accounts, though, nothing outlasts the energized lithium battery, and the hunt for the perfect one just keeps going and going. —*Sasha Nemecek*

DOUGLAS L. PECK

***SOLID-STATE LITHIUM batteries may provide consumers with the ultimate cell: one that is light, cheap and energetic. This tiny model is approximately 150 microns thick.***

# An Economic Uncertainty Principle

"If you're so smart, why ain't you rich?" It's a good question to keep in mind when watching infomercials about easy money in real estate and when listening to economists engage in learned predictions of GNP, interest rates or unemployment.

The most important discovery in economic science over the past 30 years is that this query, which I call the American Question, has no good answer. The early Latin poet Ennius sneered at forecasters "who don't know the path for themselves yet show the way for others." But only in recent decades have economists put principles and hard numbers behind this ancient observation.

It is rare that a scientific principle can make you money, but if I had learned the meaning of the American Question earlier and better, I would be $10,000 richer. It turns out that you, too, can get rich by learning it (or at least avoid losing money, which in economic terms amounts to the same thing).

The American Question is a slayer of chutzpah. It's funny, but it's serious, too, and it cuts deep. See how it skewers an economist who claims to predict next month's interest rate. Clearly, there is a fortune to be made with such knowledge: the price of interest-bearing securities such as bonds rises when rates fall, and vice versa. Knowing next month's interest rates is equivalent to knowing next month's bond prices. So why isn't the economist savant rich? Seriously, now. Keep quiet, get a second mortgage, buy bonds today, then sell them at a fat profit in a month. It's a sure thing.

Of course, a claim that the interest rate will fall next month is also a claim that the rest of the chumps in the market do not know this fact. If they did, the interest rate would have fallen already, and there would be no profit to be made buying bonds today. That pattern reflects how markets work when they get information about their future. Furthermore, there's no way around this conundrum: orders to buy or sell bonds convey information about each trader's beliefs regarding next month's interest rate. An economist with perfect foreknowledge would end up revealing it to everyone else in the market in short order.

The notion that exclusive information is dissipated by the very act of using it goes beyond financial markets. Anyone who has watched the electronics industry during the past 15 years can recall a dozen or more brilliant new ideas that seemed like a license to print money when only one company had thought of them: memory chips, video games, IBM PC clones, word-processing software. Then dozens of entrants jumped into the same niche, boldly assuming that 40 could enjoy the same profit margin as one. Often even the original innovators lost their shirts.

Economists are still discovering the full depth of the American Question—that is why even today some of them go around offering predictions about the price of bonds or the turning point of the business cycle. You can tell the history of the American Question by looking at the Nobel Prizes in Economics that were awarded for asking it. Friedrich A. von Hayek (1974) led the "Austrian" school of economics, which has been saying for a long time that if it were possible to outwit a society's judgment of what is profitable, social scientists would be rich. Merton M. Miller (1990) observed that if, in fact, the hot tips your stockbroker sells you had any value, then running the broker's little formulas for picking "incorrectly valued" stocks should make you rich. Statistics show that it ain't so.

> *"If I had learned the meaning of the American Question earlier and better, I would be $10,000 richer."*

Robert Lucas (put your money on 1998) has pointed out that the American Question could also be asked of governments trying to fool some of the people some of the time. Lucas's work with Thomas Sargent and others is called rational expectations, a phrase with enough arrogance and mystery to be controversial. But Lucas would agree that at bottom he is asking the same goofy old question: If the deep thinkers at the Federal Reserve Board can outguess the public, Lucas asks, why aren't they rich? Or, at the very least, why isn't the government solvent?

Paul A. Samuelson (1970)—who is rich, thank you—did not get the Nobel Prize for asking the Question. But he has held all along, as he put it in 1982, that "it's a mug's game for a dentist—or an associate professor of econometrics—to think that he can have an edge over those who count the cocoa pods in Africa and follow the minute-by-minute arrival of new information."

The Question is beginning to constrain economic argument the way Heisenberg constrains quantum mechanics or Gödel constrains mathematical logic. In all probability, you cannot understand a social fact and make money from it at the same time. And if you could, the market would adapt to render your understanding invalid. (The economic statistics that were once used to predict recessions, for example, now predict instead when the government will make policy changes intended to ward off recessions.)

So here is the practical use of this deep scientific principle. Fire your stockbroker the instant he says that Fly-by-Nite Canadian Gold Mining Ltd. is "undervalued" or that "we're recommending" Whitewater Real Estate, Inc. Or, to take my own sad case, do not believe your brilliant former student at the University of Chicago when he comes up with a scheme, in which other economists have invested, to make money out of a glitch in the foreign-exchange market. On that one I lost half my $10,000 in a weekend.

So be smart and rich. Do what I say, not what I do. Constrain economic science with the American Question, and if someone offers you a tip, ask yourself why they are not using it themselves. As a Damon Runyon character said, "Now, Herbie, I do not doubt your information, because I know you will not give out information unless it is well founded. But I seldom stand for a tip.... So I thank you, Herbie, just the same, but I must do without your tip."

*DONALD N. McCLOSKEY is professor of economics and history at the University of Iowa. His most recent book is* Knowledge and Persuasion in Economics *(Cambridge University Press, 1994).*

# Playing Chess on a Go Board

It had been a long, wet day, and the O'Nair triplets—Millicent, William and Deborah—were quickly running out of ideas.

"I'm bored," Millicent sighed.

"Why not play another game of Go?" William suggested.

"We've played Go so many times that everything I look at turns into black and white circles," Millicent replied.

"Let's play chess then," Deborah offered. "The Puzzles and Games Ring of the Archimedean Society—that's a student society at the University of Cambridge in England—has invented a way to play chess on a Go board. Well, something a bit like chess," she amended. "Not exactly the same, you understand. Well, I suppose if I were pressed I'd say it's rather different, really, but—"

"Oh, do stop waffling. Is it fun?"

"Definitely. It's a very unusual game, as you'll see if you let me explain it to you. The pieces sort of break up and reassemble as the game goes on. The game is called gess, pronounced 'guess,' because it's a cross between Go and chess. You play gess on a board with a grid of 18 × 18 squares—which, coincidentally, is how the lines on a Go board are arranged. Of course, in conventional Go you play on the intersections of the grid, but you play gess in the squares. You use the Go stones to make pieces—though in a very curious way.

"There are two players, black and white, each having 43 stones of his or her own color. They take turns to move, with black starting. The starting position is just as it is in chess [*see left illustration below*].

"Each player's turn consists of moving one piece, according to rules I'll explain. But a piece isn't a single stone; it's a group of them. Specifically, a black piece is any 3 × 3 region of the grid that contains at least one black stone and no white stones. In fact, the piece can go partly off the edge of the board, as long as it is not totally off the board [*see right illustration below*]. Similarly, a white piece is any 3 × 3 region that contains at least one white stone and no black stones. With me so far?"

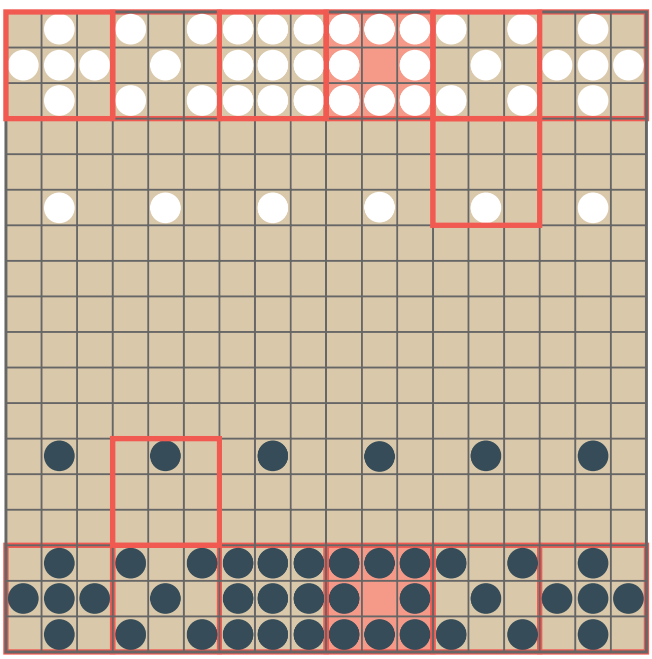
"Certainly," Millicent and William answered in unison.

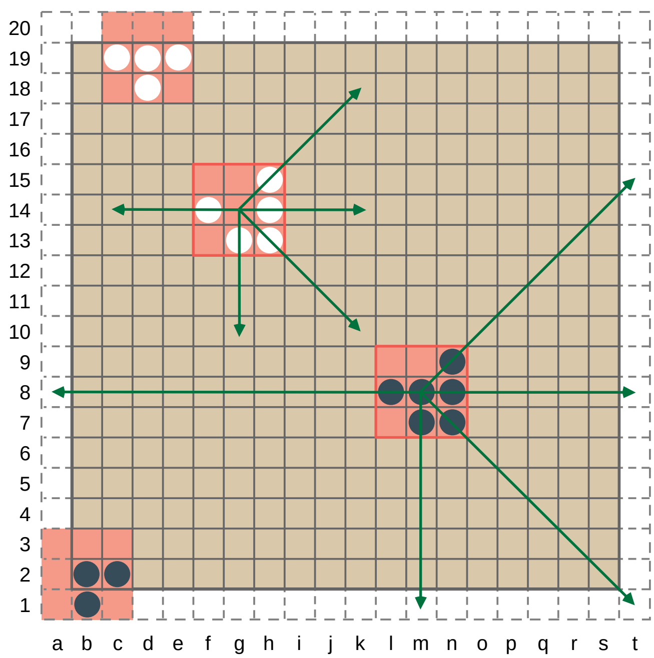
"A piece moves as a single unit. How it moves is determined by the arrangement of stones within it. The central square of the 3 × 3 region tells you how far it can move, and the outer squares tell you in which directions. If the central square is unoccupied, the piece can move at most three grid squares. If the central square has a stone in it, the piece can move as far as it likes in any legal direction—unless it is obstructed, which I'll come back to later. The stones in the outer squares determine which directions are legal. For instance, if a piece has a stone at the northwest corner, it can move diagonally in the northwest direction—but it can't return along the southeast direction unless it has a stone in its southeast corner. If it has a stone in the north position, it can move north, and so on."

"So the shape of a piece is a kind of map of where it can go and how far?" William asked.

"That's right. That's one advantage gess has over chess: the pieces remind you how they can move."

"I'm beginning to see why you said the starting position looks like chess," Millicent said. "If you group the stones in 3 × 3 squares, you get pieces whose






***STARTING POSITION for gess* (left): *the red outline marks groups of stones whose moves resemble those of a rook, bishop, queen, king and pawn. The players' rings are shaded. A piece is any 3 × 3 region that contains at least one stone and no stones of different colors. As demonstrated on the right, the stones in a piece determine the directions in which it can move and also the maximum distance—three grid squares if the central square is unoccupied; unlimited if it is occupied.***

legal moves are those of the rook, bishop and queen in chess. And one piece, with eight stones and a hole in the middle, moves pretty much like a king."

"Right," Deb confirmed. "It can go up to three grid squares in any direction—up to three because the central square is empty and any direction because all the outer squares contain stones. That piece—or any other with the same shape—is called a ring, and it's important that you protect it."

"Why?"

"You lose if your ring is damaged. But it's not quite that simple, because during the game you may be able to form extra rings, and if you do, your opponent has to damage them all to win."

"I suppose those single stones up in front are like pawns," William wondered.

"Precisely. The only chess piece that is not represented, even approximately, is the knight. That's because all moves in gess are slides—no jumps. Now, as I said, let's suppose it's black's turn to move. She chooses any one of her pieces—some suitable 3×3 region. Note that pieces may overlap one another. That doesn't matter; it just gives more choices of move. Having chosen a piece, black works out what directions she can move in and how far. Then she moves the piece in that direction, as far as she wants to, subject to the limits on that piece. But if there is an obstacle in the way, things are more complicated."

"What do you mean by 'obstacle'?"

"First, define the footprint of a piece to be the entire 3×3 region that it occupies. As it moves, imagine sliding the footprint along with it. It can continue moving in its chosen direction only if its footprint does not hit any other stones, black or white. The stones in the moving piece are not obstacles—they just move together—but any other stone is.

"When the footprint of a moving piece first covers one or more obstacles, the piece stops moving, and those obstacles are removed—whatever color they may be [*see illustration on this page*]. Of course, you generally try to remove your opponent's stones, but there's no rule against removing some of your own. After black has moved one piece and removed the appropriate obstacles, then it's white's turn, and so on.

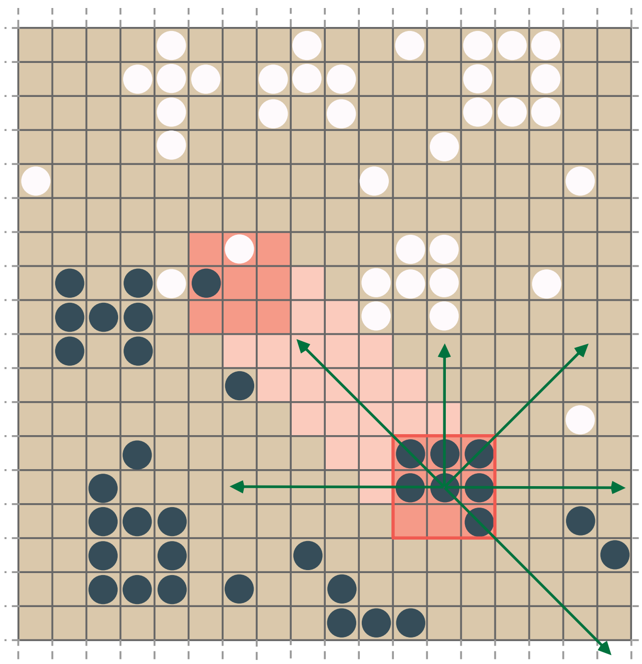
"You can stop moving a piece before it hits an obstacle—just as in chess, you're not obliged to take a piece on every move. A single stone sitting at the center of a 3×3 square can move as far as it likes but in no direction."

"Meaning it can't really move at all?"

"You got it. Oh, and any move that leaves the board looking unchanged isn't allowed. The game ends, as I said, if one player has no complete rings, in which case that player loses. So you must protect at least one of your rings—just as chess players have to protect their kings. But if during play, you create extra rings, you need protect only one of them. Of course, you must be careful not to break up your own ring by moving a piece that only partly overlaps it. Unless, again, you have a spare intact ring somewhere else."

"I see," William interjected. "Although the opening position is inspired by chess, you don't have to keep choosing the 'chesslike' 3×3 regions that determine the initial position. You can break up the original groups of stones by moving overlapping 3×3 squares, and of course pieces can gain stones by moving next to them, or they can lose them when they are taken by an opponent—or yourself. That must make the game rather different from chess."

"Of course, that's what makes it fun. But it follows many of the same general principles—protect your ring, keep a strong 'pawn' formation, occupy open lines, avoid having 'pinned' pieces that can't move and so on. Chess players soon adapt to it."



***PIECE must stop moving when its footprint* (shown shaded above) *first encounters other stones, but it may stop sooner. Any stones covered by its final footprint are removed, whatever their color may be.***

"Let's play a game," Millicent proposed.

"Right. I'm going to record the moves on paper, so I must also explain the notation for grid squares. It's quite easy. The squares on the grid are numbered from 2 to 19 vertically and lettered from b to s horizontally. So the square three across and two up from the lower left-hand corner has coordinates d3, and so on."

"Those are funny choices."

"Remember, pieces can go off the board, so there are 'invisible' rows 1 and 20 and columns a and t as well. A piece is referred to by the coordinates of its central square. The piece g5 covers the squares f4, f5, f6, g4, g5, g6, h4, h5, h6."

"Okay. I'll be black," said Millicent, "so I go first, right? Good. I'll move f6–f7 [*see illustration on next page for the moves in this game*]."

"Ah," said William, kibitzing. "You're obviously preparing to form a powerful diagonally moving piece later on."

"Am I? I thought I was just tentatively advancing a pawn."

"I'll reply with p15–m12," Deborah declared.

"Hmm. Trying to control the center—and you're preparing to create a second ring, I see. Two can play at that game. I move e3–e6."

"Two can play at that game, too. I'll move p18–p15."

"Now what?" said Millicent, temporarily stumped for a good move.

"Try b3–e3," William urged.

"Can I do that? Oh, yeah, pieces can go off the edge—partially. And I form a second ring, securing my position."

"For the moment," Deborah countered. "My move is e15–h12."

"Still strengthening your center, I see."

"And opening a line for an attack on your new ring," William observed.

"True. How about m6–l7? Opening up my own line?"

"Not bad."

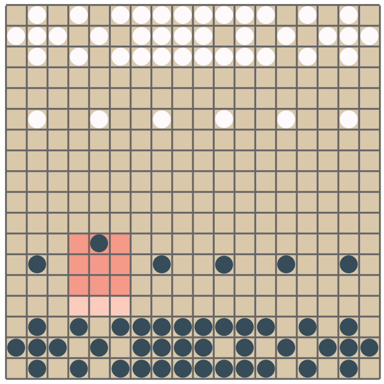
"Humph. I'll fix you: m12–j9. Pawn takes pawn. Sort of."

"That's not such a good move," William pointed out. "Your stone on l14 is pinned. You've got a ring behind it and an open file leading to a very powerful black piece in front of it."

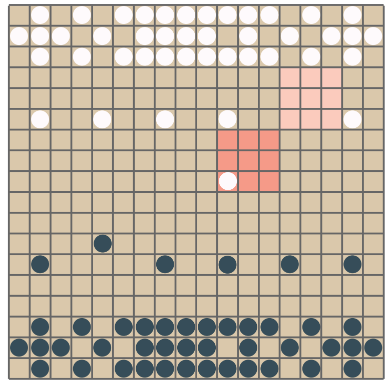
"Too late now," Millicent chuckled. I'll take one of your stones now: i6–i7."

"Very unimaginative," William criti-

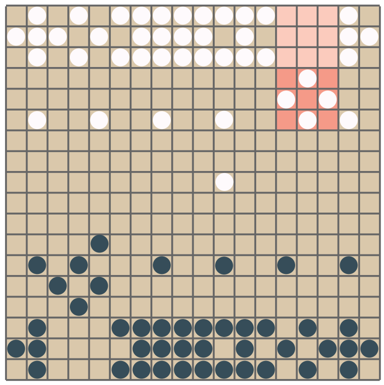
*1* BLACK BEGINS



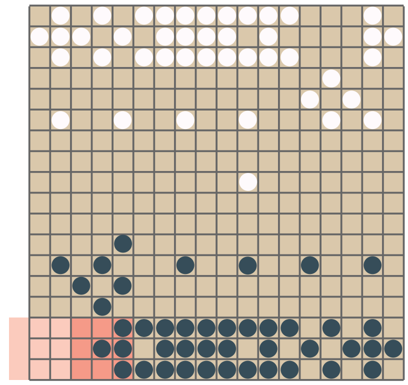
*2* WHITE ADVANCES A PAWN



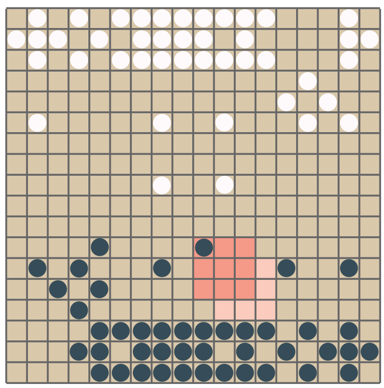
*4* WHITE MATCHES THE MOVE



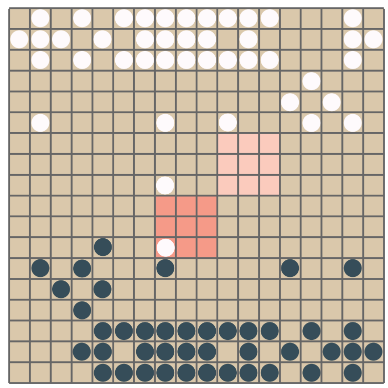
*5* BLACK FORMS SECOND RING



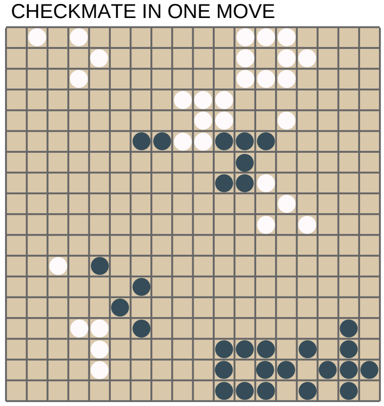
*7* BLACK ALSO MOVES INTO CENTER



*8* WHITE PAWN TAKES BLACK PAWN



. . . CHECKMATE IN ONE MOVE



***OPENING MOVES** in a typical game of gess appear in the first nine squares. Light shading shows the original position of a piece, and dark shading, its new position. The game ends in the last square: white to move and mate in one.*

cized. "I would advise h6–i7. Or j6–k7?"

"Oh, do shut up."

"I'll reinforce my ring and open up my 'queen' with h15–k12," Deb rejoined.

"I can open up the game, too: i7–i10."

"You're exposing one of your rings," William objected.

"Yeah, but I've got another one," Millicent said. "Your move, Deb."

"Uh, uh—m15–j12."

"Your pawns are overadvanced."

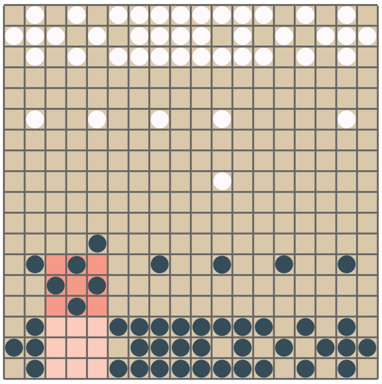
"As Milly said, shut up, William!"

The game—and argument—continued for several more moves until it reached the position shown in the final illustration on this page, with white to move. Deb thought for several minutes, then announced, "Checkmate."

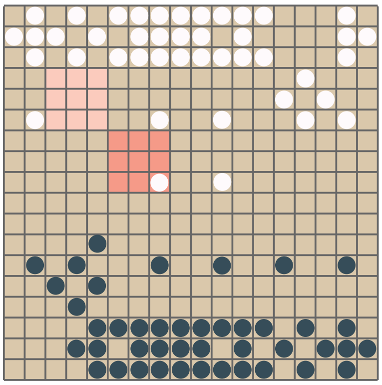
"What?" Millicent yelled. "Where?"

Can you spot the winning move? For

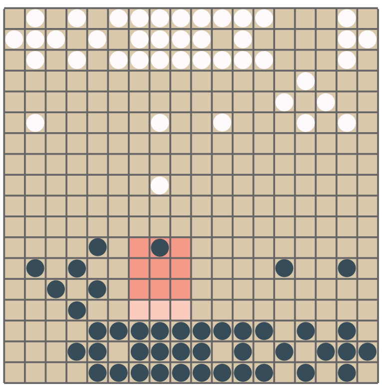
3 BLACK PREPARES TO FORM SECOND RING



6 WHITE ADVANCES INTO CENTER



9 BLACK PAWN TAKES WHITE PAWN. . .



the answer, see the end of this article.

The O'Nair triplets played gess for weeks, until one day Millicent suddenly said, "I'm bored."

"The Archimedean Society's Puzzles and Games Ring has invented a really neat game that combines croquet and alligator wrestl—" Deborah began.

### *ANSWER*

The winning move is o11–j6. Now, whatever black does, her ring won't survive a diagonal thrust.

FURTHER READING

GESS THE GAME. Paul Bolchover in *Eureka,* Vol. 53, pages 24–27; 1994.

## Of Ants and Men

**JOURNEY TO THE ANTS: A STORY OF SCIENTIFIC EXPLORATION,** by Bert Hölldobler and Edward O. Wilson. Illustrations in black-and-white and color. Harvard University Press, 1994 ($24.95).

The authors are two senior professors, close friends for three decades, collaborators on more than 200 publications, who meet still for an annual working holiday in the field. The burly Hölldobler and the wiry Wilson are shown peering together, last year, into a nest of carpenter ants. The same men are shown apart in much older images: two young teenage entomologists, the one in Bavaria, the other in Alabama, each armed with his collecting net. "To put the matter... simply... we, having entered our bug period as children, were blessed by never being required to abandon it."

The two entered science an ocean apart along different academic paths, although they share an intellectual grandfather, more or less their own professor's professor's professor. That worthy man—William Morton Wheeler of Harvard before World War I—was a gifted savant. Wilson still works at Wheeler's big old desk in homage, but early on he was persuasively exhorted to outdo his famous forebear. "We can and will do better; we must," insisted Wilson's not so old mentor, the celebrated ant finder and classifier William L. Brown, long at Cornell. The task they all joyfully undertook was "to bring ants into the mainstream of biology."

They have fulfilled their duty; the old "eccentric tinge" is washed away by the tide of deepening understanding. Today the old tools—forceps, microscope and steady hand, even the newer cunning interventionary nest and swarm manipulation—are complemented by the full instrumental arsenal of biochemist and molecular biologist. The language current among all those ant nanominds is mainly gestural and chemical, and those who would decipher its toughest texts must learn first to read molecules.

But they remain naturalists, opportunists zealous to learn as much as possible about "all aspects" of the "totem organisms" they venerate. Wilson, a gifted writer and a versatile scholar, manifestly enjoys the literature and the culture of science as well. He recalls a *National Geographic* article he read at age 10, whose happy author recalled one Christmas day in the Cuban mountains when he had found under a rock "brilliant green metallic ants... an unknown species." "Imagine!" thought the boy, "prospecting in a faraway place for new species of ants that resemble living emeralds." Ed Wilson, 15 years later an ambitious graduate student, repeated the experience closely, finding the same species aglitter again under a rock in the same locality, "a reassurance of the continuity of the natural world, and of the human mind." The human community of the myrmecologists, a happy few, is as far-flung as the myriad ants themselves.

About 10 years ago Hölldobler, after 20 years at Harvard, began to receive

***LEAFCUTTER ANTS** cultivate fungus on bits of leaves that they cut* **(top left)** *and transport back to the nest* **(bottom left).** *Sometimes members of a smaller caste ride on the leaf fragment* **(right)** *to protect the carrier ant from parasitic flies.*

"irresistibly attractive offers" from Europe. He and Wilson decided to write as thorough a reference treatise as possible before their daily partnership should end, a book to supplant at last Wheeler's own great vade mecum of Edwardian years. In 1990 *The Ants* appeared. The quality of their writing is easy to assess: it won the Pulitzer Prize for General Nonfiction, the first "unabashed" work of science ever to be so recognized.

Uncle Bill Brown himself reviewed *The Ants* in these pages; he found it fully worthy, the "masterpiece" he had insisted on. So it is, yet all the same it is a data-filled tome weighing seven pounds. Now the partners—Hölldobler did return to his own university, Würzburg, where his father had professed entomology—have made this book for the public, beautifully written and illustrated, to expound at modest length on the ants. The loss of the rich quantitative evidence that lies behind their gripping narrative of a bizarre territory is in part assuaged for general readers by the inclusion of the personal story of those who explore it, the very material we have just sampled. These 15 chapters are a bustling but well-organized ant heap, full of wonders natural and intellectual, a linked modern narrative of many ants and some humans.

That history is startling, its intricate diversity evolving under stringent law: fossil Ur-ants waspy in amber, self-exploding kamikaze ants, rampant parasitism (how easily the nearly clueless laborers are fooled by poorly forged passwords), ants that mimic the termites, air-conditioned nests, dirt used as camouflage (turned into white when those ants are reared amid plaster), ant squeaks and taps that augment their basically chemical language.

One extraordinary finding is all we can describe. The Harvard offices of the two men became festooned for a decade by plastic tubes and clear chambers, within which prospered bustling settlements of tropical leafcutters, a species of *Atta*. Like most human societies today, these ants depend on agriculture. They collect fresh leaf cuttings to nourish a domesticated fungus, a kind of bread mold not to be found wild, on which the entire nest feeds, from larvae to queen. The garden is minutely tended along an intricate assembly line of ants, a bobbing, weaving procession busiest at midnight. These daughters of their queen are all physically varied sisters. Ants of different castes differ greatly in size. The most abundant caste, dwarfed gardeners an eighth of an inch long, mine, patrol, weed and reap within the gray, tangled masses of the fungus strands. Larger ones crush and mold leaf fragments into moist pellets, adding them to nourish the spongy crop. Still bigger ants clip the supply of new leaves into millimeter-size fragments for the crushers, and the largest big-jawed workers, the size of a big housefly, steadily forage abroad, to bring back big, fresh leaf sections for the clippers.

This account is not itself new. Such specialized body forms and behavior had led Wheeler to speak of a "superorganism," the ant colony in metaphor: its individual ants as types of cells, its castes tissues and organs, its generative system the queen and her drones. What the Harvard pair made clear for the first time was how real the powerful metaphor is, how predictive, how quantitative: the colony is integrated as though it were in fact one organism ruled by a genome that constrains behavior as it also enables it. The experimenters could act at will to change these living "tissues" by depleting or increasing the ant population of each caste, as though one should remove for a few days the thumb from a farmer's hand. Vivisection was their tool, but vivisection that is bloodless, painless and quickly reversible without harm. Then they measured the metabolic balance of each altered superorganism, the nutrient energy gathered against energy costs, measured by total oxygen uptake.

The result was a clear statement, valid to a few percent. Small-jawed foragers and dwarf gardeners are enough while the colony remains small and its leaf lands narrow. Add more ants with a millimeter too much jaw, and the economy suffers, for ants that are too large eat too much for what they produce. Permit only too few of the big jaws, and the leaf take suffers. The physical superorganism acts to adjust the demographic mix so as to optimize its energy economy. A big enough *Atta* colony, its foragers widespread, a royal treasure in its fungus gardens, even develops a standing army for the long run, a new caste of much larger soldiers, their great jaws fit to cut and clip not mere leaves but tough enemy insect bodies.

Ant societies have a foreign policy, too, one as highly utilitarian as their domestic polity. It is summed up here: "Restless aggression, territorial conquest, and genocidal annihilation... whenever possible. If ants had nuclear weapons, they would probably end the world in a week."

We have all encountered some principles of ant superorganisms propounded by philosophers of our own species. The austere rules allow of no play, no art, no empathy, as demonstrably they tend to maximize one simple quantity. Here she is, elegant yet real; in her purity we recognize her, Economic Ant.

## Coming Attractions

**THE HOT ZONE**, by Richard Preston. Random House, 1994 ($23).

This intimate, suspenseful chronicle is not fiction. And "hot zones," in the jargon of biohazards research, are not places in need of cool air. Their heat is an infestation by lethally infectious microorganisms. One such zone is a set of stainless-steel research rooms, labs of the Biosafety Level 4 suite at the U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, Md., 25 miles north of the District of Columbia line. The labs are held under reduced air pressure behind airlocks and special showers, all elaborately defended against any leak to life outside. In the locker room that raises the first real barrier, at minimal Biosafety Level 0, a sign offers the opening command: "Remove everything touching the skin: clothing, rings, contact lenses, etc. Change into sterile surgical scrubs." For Level 4, you need to have scaled all the steps into your heavy blue, pressurized and decontaminated plastic space suit. Then you must climb out again carefully, sign by sign, level after level.

The second hot zone, uncertain and obscure, is an African mountain cave once frequented—perhaps once even excavated—by elephants hungry for the salt in its walls! The Kitum Cave is now a tourist destination on Mount Elgon, an old Rift volcano that rises high out of the plains. Right on the border between Uganda and Kenya, 50 miles north of the Lake Victoria shore, Elgon's wide slopes hold many villages within the fast-diminishing rain forest.

The two zones have somehow been linked by the transfer of certain virus particles. One particular class of virus forms the link. Called filoviruses, thin, protein-walled molecular tubes a few microns long can be seen in the electron microscope as telltale bent, crooked or looped threads. They enclose fragile tapes of RNA. Let several of these tapes enter intact into just the right living cell of human, monkey or guinea pig, and they multiply inordinately; host matter turns to virus; cells grow entire crystals of packed threads.

Typical responses are appalling for all of us potential hosts. First the virus incubates, for a few days or a couple of weeks. Once enough threads have been produced and set free, they clog the blood flow in many capillaries by congealing the platelets, so many cells soon die all over the body. Larger and larger blood vessels break as they in turn clog and swell, until the host mammal bleeds outwardly through all orifices, and inwardly too, wherever fluid blood and new clots find available space. Blood spills into lung and abdominal cavities and may even puff out a whole new space between the weakened skin and the flesh. Organs, save for skeletal muscle and bone, soften to jelly as abundant virus transforms the supporting collagen. The animal enters into a state of total shock, circulation utterly awry, and may "crash and bleed out," as the medical jargon describes this style of death. Whereas one in 20 of yellow fever–infected patients treated in a modern hospital may die, the virulent thread viruses may in a day or two kill one in four, or up to nine in 10, depending on the filovirus subtype.

So far about half a dozen limited human outbreaks of thread virus have been endured. The first of them remains only an anecdote: an English veterinarian, who was then working to inspect cattle around Mount Elgon, recalls credible accounts he heard during the early 1960s, monkeys and humans nearby dying of a bleeding disease. In 1967 the virus was first identified at a plant in north Germany where they routinely grew monkey cell cultures to yield vaccines. Those virologists of Marburg first saw filovirus in dying monkeys just imported from somewhere in Uganda and soon within 31 infected people around their plant, seven of whom quickly died.

In 1976 the fiery threads flared twice, killing hundreds of villagers at two rainforest locales 600 miles apart and more or less as far from Mount Elgon. This time the world paid attention. Emerging viruses of such virulence are riveting; the lesson has been taught again by the scourge of HIV. Within a few days the U.S. Centers for Disease Control team in Atlanta had a new blood sample; they found it to be a new thread subtype and named it Ebola Zaire, for the small river valley north of the Congo where it had first appeared. Its antibodies are quite distinct from the Marburg form; this virus was a true "slate-wiper"; nine out of 10 persons infected had died.

Worse, it was demonstrably loose in Kinshasa city, where two million people crowd together. A light airplane had just flown in two nuns, one of them already ill, from the upriver town where their hospital had been decimated by the newfound virus. One young local nurse who had cared for the dying visitor had wandered the city for two days and was now herself ill. Soon all three women had crashed out bloodily, and "full-scale panic" gripped the World Health Organization in Geneva.

Zaire set tight military quarantines in place. The American centers flew strong teams to the country. Upriver they found the disease had burned out so swiftly within the regional hospital that there had been little outside contact. At Kinshasa the activists of the CDC quickly organized a hospital ship to moor in the wide river, as a refuge for the doctors to come, held safely away from the hot zone of the big city. The 82nd Airborne was alerted to help bring out the 1,000 Americans from the city, once the virus kindled the flame. But Ebola never ignited; not one more case was seen in Kinshasa. These threads were very hot—indeed, lethal—but not infectious through casual contact, sparing even the person who had shared a bottle of soda pop with Nurse Mayinga as she fell ill that threatening day in 1976.

Ebola subtypes of the filovirus soon became a research program at the Army Institute. They were too risky to ignore; neither origin nor virulence was understood. In 1983 a controlled Level 4 experiment showed that Ebola Zaire could indeed pass airborne from monkey to monkey, probably via droplets: this certainly was bad news, Kinshasa's reprieve notwithstanding.

During the 1980s, two new samples of fatal threads reached the U.S. from Nairobi Hospital. First a man and then a boy had crashed out from thread virus there, although the two cases were years apart. Only one earlier crossing of their paths in life could be teased out by careful tracing; each had visited the Elgon elephant cave not long before his final illness.

Off to test the heat of the suspect cave went a sizable Frederick team to join expert Kenyan partners, laden with battery-driven field pressure suits (not blue but orange), radios, refrigerators, bat traps, tick traps, insect traps, cages, monkeys, guinea pigs. They audited the cave in all its complexities, living and inert. But not a sign of thread virus was found in any of their samples. No one expects a mammal to be the steady host for Ebola; the virus is too intemperate therein. ("Did you ever try to draw blood from an elephant? We didn't.") But in spider, frog or newt, in dust or on rock…nothing.

In the fall of 1989 Ebola surprisingly took the offensive. Threads arrived only 25 miles from Frederick, their vector 100 big macaque monkeys sent from the Philippines, to remain a while in a commercial quarantine colony in Reston, Va., near Dulles Airport. The monkeys, intended for eventual distribution to U.S. research users, died steadily after fierce internal bleeding; the responsible veterinarian suspected a virus and called the Institute for specific advice.

So far only grim prologue: half the book tells the story of this crisis in vivid and personal detail. A filovirus, plainly fulminating, hemorrhagic and airborne, its antibodies matching those of hottest Ebola Zaire, was now abundant and growing fast within three tons of caged monkeys, half an hour from our nation's capital. They had to act.

Not one person among the five or six surely exposed at that time to subtype Ebola Reston was ever infected. The new subtype simply never jumped out of its macaque hosts. This joyous anticlimax to a grim threat is all true. No one has yet found any differences that can explain why, even after close study of the seven proteins all Ebolas share. How did a Central African virus enter from the Philippines anyhow? No one knows that either; maybe wealthy hunters had imported African game to stock some island rain forest.

The monkeys were deftly sacrificed, sampled and disposed of. Their cinderblock house was first sealed, then utterly vacated and finally rendered certifiably devoid of all life within the week by a dozen or two young soldiers, men and women from Fort Detrick. Trained animal-care technicians, 18 years age and up, they volunteered that Monday morning to outface Ebola daily in orange space suits under Level 4 conditions. (Their typical army job classification, designated 91-Tango, rates a steady seven dollars an hour; frugal army paymasters reckon you don't deserve hazard pay for hot-zone work as long as you are within a regulation pressurized suit.) The technicians had a genuine working leadership of six or eight expert officers and civilians whom they knew and could trust.

The Reston monkey house is abandoned now. Ebola burned there once, consumed only monkeys and subsided. But Richard Preston ends with a stoic's reasoned forecast: "It will be back."

In this tale, real person after real person is presented, from a touching family funeral to a cut glove in Level 4. A high-budget motion picture made from the book is now in postproduction and will come to a theater near you early in 1995. Celebrated stars will play the real virologist, Karl M. Johnson, who first found and then challenged Ebola in Kinshasa, and the real pathologist, Colonel Nancy Jaax, whom we first meet at home, getting ready for her initiation into Level 4. Meanwhile read the book; the public health flows from unceasing devotion.

# Words at Play on the Internet

When Michael Faraday needed new and fitting terms for his discoveries in electrochemistry, he consulted William Whewell, professor of moral philosophy at the University of Cambridge. The professor produced "electrode" (from the Greek, meaning "a way"), "ion" (from the Greek, meaning "I go") and "electrolysis" (from the Greek, meaning "I dissolve"). Nowadays Marvel Comics is as likely a source for technical neologisms as *Skeat's Etymological Dictionary,* particularly on the Internet, that freewheeling collection of roughly 30,000 interconnected networks and 2.5 million attached computers. The buoyant, fast-growing Internet lexicon combines metaphor, pop allusions and whimsy in playful etymologies that are as good-humored as they are eclectic.

Consider, for instance, the term "gopher." On the Internet, gopher is an information retrieval system. Its name is a play on gofer (the person who goes for something) and on gophers, any of various short-tailed, burrowing animals in the Geomyidae family (a play on the gopher program's ability to help users tunnel through a maze of databases to get the information they need). The term also slyly acknowledges gopher's developer, the University of Minnesota. (Minnesota is the Gopher State; the University of Minnesota is home of the Golden Gophers.) To access gopher, you need a gopher "client"—special software installed on a computer attached to the Internet. Once you are in gopherspace, your client will look for servers that have what you seek—say, the E-mail address of a colleague.

Gopher—which neatly combines metaphor, allusion and general jeu d'esprit—is just one of the evocatively named search services on the Internet. There is also Veronica (Very easy rodent-oriented netwide index to computerized archives, accessed by a gopher client), which is descended from Archie, the first search engine in this linguistic family. Jughead (Jonzy's universal gopher hierarchy excavation and display) is a refinement of Veronica that lets you search yet more easily for specific information. (This is probably the first time anyone has considered Jughead a refinement of Veronica.)

Archie also produced Anarchie, a program written for Macintosh computers by Peter Lewis that both locates and downloads files automatically. There may be no further developments in the Archie family (although Betty suggests itself as the next linguistic candidate) because gopher has a supplanter in the wings. The alliterative World Wide Web (known as W3), run out of CERN, the particle physics laboratory near Geneva, is the preferred Internet database of databases for a growing number of people who use it via the metaphorically named Mosaic. Mosaic, developed at the National Center for Supercomputing Applications at the University of Illinois, is a hypertext shell—to use it, people point and click with a mouse instead of typing a command string, browsing easily by hyperlinks (the underlined or shaded/colored words that, when clicked, lead to the next documents). Like most software that has come through the public domain, Mosaic comes in flavors (customized versions rather than, say, the plain vanilla of the original source code)—including ones for Apple Macs, Microsoft Windows and X Window (X Window is a public domain window management system that can be used on Unix).

Programs are not the only items on the net with zippy names; there are also all those machines linking up to help researchers share data, send E-mail, poke through distant library catalogues and check on weather in the Southwest before calling their parents in New Mexico. An early Vax machine at the University of California at Berkeley was dubbed Covax, a punning salute to the comedian Ernie Kovacs. At my university, where Mosaic (the program) runs on Newton (the machine), we also have machines named Photon, Prism and—latest addition in the light theme, chosen by contest—Quasar; Darwin, Edison and Kepler; Tasha, Ishara and Worf (Worf is the Klingon); and George and Gracie.

Verbs, too, are graphic, metaphorical and informal on the net. For instance, users can zip and unzip (compress and decompress) files (type "prog stuffit" to find decompression programs on Archie). For pure vividness, stuffit joins "finger," a command to eavesdrop on who is using which machine, and "lurking," the expression for the voyeuristic pleasure of reading letters distributed to mailing lists. (People subscribe to mailing lists by sending their names to the computer, called listserve, that manages the list. After that, any letter posted to the list bounces out to all subscribers.)

Within the vast readerships of mailing lists, bulletin boards and conference connections such as Internet Relay Chat—a kind of mad, addictive party line where hundreds of people, typically students, exchange messages in real time—many people use the net as a convenient way to blow off steam, a practice referred to by the most vivid and least playful of metaphors used on the net, "flaming." As the term "hacker" before it, which gradually became associated with criminal behavior, flaming has come to have a darker meaning than its original sense of the ill-considered, ad hominem attack on authors of Usenet postings. (Usenet is a popular segment of the Internet composed of discussion groups, typically in the form of electronic bulletin boards. The readership is estimated at about 10 million people.) Today flames range from the scathing messages sent to two Arizona lawyers who placed unsolicited advertisements on Usenet bulletin boards (this kind of scattershot posting is known as spamming, from the tinned meat, which splatters if thrown) to hate E-mail—a recent flame applauded Rwanda "in that their human weeds are being removed with the selfsame tool used for unwanted plants, the machete."

While flaming is becoming a word of deep disapprobation, most terms for obnoxious acts on the Internet remain true to the playful spirit of the terminology—even when the acts verge on the criminal. For instance, the programs hackers devise to observe and record chunks of data (packets) when people log on—the better to steal user identifications and passwords—have picked up a name characteristic of the net's cheerful, intensely figurative vocabulary. The programs are called packet sniffers.

---

*ANNE EISENBERG (aeisenb@photon.poly.edu) answers her E-mail from Polytechnic University in Brooklyn, N.Y., where she is a professor and writer.*