VIRUS-KILLING NETWORKS • PROVING FERMAT'S THEOREM • THE LOST CITY

SCIENTIFIC AMERICAN

NOVEMBER 1997 $4.95

THE DANGERS OF "LAUNCH ON WARNING"

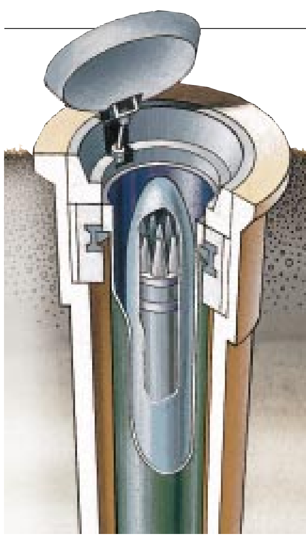
A NUCLEAR MISTAKE MIGHT BE ONLY 15 MINUTES AWAY

Mercury's long, hot afternoon

SCIENTIFIC AMERICAN

November 1997 Volume 277 Number 5

Mercury is the neglected child of the planetary system. Only one spacecraft has ever ventured near it, whereas scores have probed the moon, Venus and Mars. The scant facts available show this strange, blazingly hot planet is full of surprises: its anomalous density and magnetic field suggest that Mercury may be where to seek clues to the origin of the solar system. The poles may even hold pockets of ice.



Because of outdated "launch on warning" policies, an unexplained blip on a radar screen could trigger a nuclear strike by the U.S. or Russia in as little as 15 minutes. Given the frayed state of Russia's military, the risk of accidental or unauthorized attack is alarming. These authors present a plan, based on detailed weapons surveys and discussions with military overseers, for taking weapons systems out of perpetual readiness without compromising either nation's security.

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Periodicals postage paid at New York, N.Y., and at additional mailing offices. Canada Post International Publications Mail (Canadian Distribution) Sales Agreement No. 242764. Canadian BN No. 127387652RT; QST No. Q1015332537. Subscription rates: one year $34.97 (outside U.S. $47). Institutional price: one year $39.95 (outside U.S. $50.95). 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 info@sciam.com **Subscription inquiries: U.S. and Canada (800) 333-1199; other (515) 247-7631.**

About the Cover
The ferocious sun scorches the planet Mercury, which because of its slow rotation and rapid orbit has a dawn-to-dusk day longer than its 88-Earth-day year. Painting by Don Dixon.

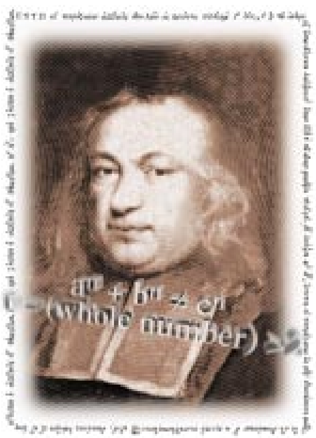
Visit the Scientific American Web site (http://www.sciam.com) for more information on articles and other on-line features.

FROM THE EDITORS

Prove It

Concerning Fermat's last theorem: I, too, have found a simple proof of the conjecture that for $a^n + b^n = c^n$, there are no integral solutions if n is greater than 2. Unfortunately, the 400-some words of this column are insufficient, so I shall return to it another time. By the way, I also found my own elegant proof of the famous theorem that no more than four colors are needed to differentiate contiguous regions on a flat map. But I wrote it on the back of a laundry receipt, and now it's gone. A dog ate my squaring-the-circle proof. So much for greatness. I'm very good at the math; it's the paperwork that gives me headaches.

Curse Pierre de Fermat and his maddening marginalia. Personally, I'm of the camp that when he scribbled his famous note, he was either joking or mistaken. Even granting his mathematical genius, I find it hard to believe 300 years of mental toil by countless professionals and amateurs could fail to reconstruct his reasoning, were he correct.



PIERRE DE FERMAT and his little joke.

But of course, we'll never really know, will we? And so it is the nagging hunch that Fermat's tidy statement must spring from an equally tidy principle that drives people back to their desks and their well-chewed pencils.

The theorem has been proved, by Andrew J. Wiles of Princeton University, but as Simon Singh and Kenneth A. Ribet explain in "Fermat's Last Stand" (see page 68), that proof involves excursions into brands of geometry undreamed of in Fermat's time. Nevertheless, Singh and Ribet at last make Wiles's proof understandable even to the computationally dysfunctional, including (ahem) yours truly.

Next month I will explain where the missing side of a Möbius strip goes. Assuming I have the space. . . .

Some problems are unsolved for lack of insight. Others are unsolved for lack of will. Too many grave quandaries in human affairs fall into the latter category, and the logjam in efforts to diminish the nuclear menace is one. If "launch on warning" policies ever truly served the best defense interests of the U.S. and the Eastern bloc, they no longer do. In "Taking Nuclear Weapons off Hair-Trigger Alert," beginning on page 74, Bruce G. Blair, Harold A. Feiveson and Frank N. von Hippel explain why these policies must go. More important, they outline a way for the U.S. and Russia to abolish launch on warning without compromising either nation's strategic interests. The authors are now briefing leaders in the Department of Defense on this plan, in the hope that specific resolutions will eventually implement it. SCIENTIFIC AMERICAN is privileged to share this information with its readers as well.

JOHN RENNIE, *Editor in Chief*
editors@sciam.com

GRANGER COLLECTION; SLIM FILMS

SCIENTIFIC AMERICAN®
Established 1845

John Rennie, EDITOR IN CHIEF

Board of Editors
Michelle Press, MANAGING EDITOR
Philip M. Yam, NEWS EDITOR
Ricki L. Rusting, ASSOCIATE EDITOR
Timothy M. Beardsley, ASSOCIATE EDITOR
Gary Stix, ASSOCIATE EDITOR
Corey S. Powell, ELECTRONIC FEATURES EDITOR
W. Wayt Gibbs; Kristin Leutwyler; Madhusree Mukerjee; Sasha Nemecek; David A. Schneider; Glenn Zorpette
Marguerite Holloway, CONTRIBUTING EDITOR
Paul Wallich, CONTRIBUTING EDITOR

Art
Edward Bell, ART DIRECTOR
Jana Brenning, SENIOR ASSOCIATE ART DIRECTOR
Johnny Johnson, ASSISTANT ART DIRECTOR
Jennifer C. Christiansen, ASSISTANT ART DIRECTOR
Bryan Christie, ASSISTANT ART DIRECTOR
Bridget Gerety, PHOTOGRAPHY EDITOR
Lisa Burnett, PRODUCTION EDITOR

Copy
Maria-Christina Keller, COPY CHIEF
Molly K. Frances; Daniel C. Schlenoff; Terrance Dolan; Katherine A. Wong

Administration
Rob Gaines, EDITORIAL ADMINISTRATOR
Sonja Rosenzweig

Production
Richard Sasso, ASSOCIATE PUBLISHER/VICE PRESIDENT, PRODUCTION
William Sherman, DIRECTOR, PRODUCTION
Janet Cermak, MANUFACTURING MANAGER
Tanya DeSilva, PREPRESS MANAGER
Silvia Di Placido, QUALITY CONTROL MANAGER
Carol Hansen, COMPOSITION MANAGER
Madelyn Keyes, SYSTEMS MANAGER
Carl Cherebin, AD TRAFFIC; Norma Jones

Circulation
Lorraine Leib Terlecki, ASSOCIATE PUBLISHER/CIRCULATION DIRECTOR
Katherine Robold, CIRCULATION MANAGER
Joanne Guralnick, CIRCULATION PROMOTION MANAGER
Rosa Davis, FULFILLMENT MANAGER

Advertising
Kate Dobson, ASSOCIATE PUBLISHER/ADVERTISING DIRECTOR
OFFICES: NEW YORK:
Meryle Lowenthal, NEW YORK ADVERTISING MANAGER; Kevin Gentzel; Thomas Potratz; Timothy Whiting.
DETROIT, CHICAGO: 3000 Town Center, Suite 1435, Southfield, MI 48075;
Edward A. Bartley, DETROIT MANAGER; Randy James.
WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025;
Lisa K. Carden, WEST COAST MANAGER; Debra Silver.
225 Bush St., Suite 1453, San Francisco, CA 94104
CANADA: Fenn Company, Inc. DALLAS: Griffith Group

Marketing Services
Laura Salant, MARKETING DIRECTOR
Diane Schube, PROMOTION MANAGER
Susan Spirakis, RESEARCH MANAGER
Nancy Mongelli, ASSISTANT MARKETING MANAGER

International
EUROPE: Roy Edwards, INTERNATIONAL ADVERTISING DIRECTOR, London. HONG KONG: Stephen Hutton, Hutton Media Ltd., Wanchai. MIDDLE EAST: Peter Smith, Peter Smith Media and Marketing, Devon, England. PARIS: Bill Cameron Ward, Inflight Europe Ltd. PORTUGAL: Mariana Inverno, Publicosmos Ltda., Parede. BRUSSELS: Reginald Hoe, Europa S.A. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

Business Administration
Joachim P. Rosler, PUBLISHER
Marie M. Beaumonte, GENERAL MANAGER
Alyson M. Lane, BUSINESS MANAGER
Constance Holmes, MANAGER, ADVERTISING ACCOUNTING AND COORDINATION

Chairman and Chief Executive Officer
John J. Hanley

Corporate Officers
Robert L. Biewen, Frances Newburg, Joachim P. Rosler, VICE PRESIDENTS
Anthony C. Degutis, CHIEF FINANCIAL OFFICER

Program Development
Linnéa C. Elliott, DIRECTOR

Electronic Publishing
Martin O. K. Paul, DIRECTOR

Ancillary Products
Diane McGarvey, DIRECTOR

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue • New York, NY 10017-1111
(212) 754-0550

PRINTED IN U.S.A.

LETTERS TO THE EDITORS

HOT TOPIC

As an occupational physician and toxicologist who has treated thousands of patients whose health and lives have been stolen from them by the mineral asbestos, I cannot sit by without comment on the July article "Asbestos Revisited," by James E. Alleman and Brooke T. Mossman. Asbestos is a chronic poison and proved human carcinogen in all its forms. Does the need for better mailbags provide a rationale for the continued use of this killer or the loss of even one life? And how could there have been no mention of the late Irving Selikoff's definitive research on the asbestos plague? Why is no reference made to Cesare Maltoni's work on the basic science and epidemiology of asbestos? Alleman and Mossman's article may be couched in a charming literary style, but it is filled with smoke and mirrors.

DANIEL THAU TEITELBAUM
Medical Toxicology
Denver, Colo.

STEVE FRISCHLING *AP Photo*

ASBESTOS REMOVAL in New York City.

Alleman and Mossman dismiss the health concerns related to low doses of asbestos as emanating solely from the class of asbestos known as amphiboles. This is an entirely inadequate and inaccurate assessment of the issue. As I explain to each resident in our occupational and environmental medicine training program, the increased risk of developing lung cancer is associated with all types of asbestos, including Alleman and Mossman's "safer chrysotile form."

PETER ORRIS
Cook County Hospital
Chicago, Ill.

Alleman and Mossman reply:

We wrote "Asbestos Revisited" as a history of asbestos use rather than as an article about the many contributions of medical researchers who have studied the health effects of asbestos. The true "smoke and mirrors" can be found in Teitelbaum's references to an "asbestos plague" caused by a "chronic poison." Whereas this misleading information may fuel asbestos litigation, expensive and unnecessary removal of intact asbestos, and general hysteria, it is incorrect: the rates of mesothelioma in the U.S. appear to be declining. And unlike a contagious disease transmitted by brief contact, asbestos fibers must be airborne and inhaled for extended periods at high concentrations to cause an increased risk of cancer.

We thank Orris for his comments emphasizing that lung cancer is associated with asbestos workers exposed to all types of asbestos. It is worth noting, however, that tumors are rarely found in nonsmokers, and several studies of workers (predominantly smokers) exposed to chrysotile in cement plants indicate that their risk for lung cancer is not any higher than the risk among smokers in the general population. These results, along with several other studies, suggest that chrysotile may be a less potent form of asbestos in the development of lung cancer.

ALL THE WORLD'S A STAGE

In his review of Tom Stoppard's play *Arcadia*, Tim Beardsley observes that the play "is poised to reach a much larger audience now that general production rights are available in the U.S." ["Sex and Complexity," Reviews and Commentaries, July]. Quite so. A few weeks after I saw the play in Houston, I was privileged to be at a dinner with Tom Stoppard at the Ransom Center at the University of Texas at Austin.

I asked Stoppard if anyone has ever created a "Coverly set," a mathematical creation that, in the play, was generated by young Thomasina Coverly's set of equations on a laptop screen not seen by the audience—a trick reminiscent of Fermat's famous notation in the margin. Stoppard told me that there was no Coverly set when he wrote the play but that there is now. The set has been created by Andrew J. Wiles, who proved Fermat's last theorem.

BILL HOBBY
Houston, Tex.

HISTORY LESSON

In the article "Rights of Passage" [News and Analysis, July], Philip Yam writes, "In spite of persecution, scientists have invariably advocated free thinking, political openness and other human rights." Invariably? I can think of several counterexamples to this sweeping judgment. I know of no systematic study classifying oppressors and murderers according to their academic training, but if one were done, I doubt any discipline would emerge unscathed. After all, Joseph Stalin was once a theology student.

RACHEL E. FAY
Mary Esther, Fla.

Yam replies:

As Fay implies, all human endeavors have their dark sides. The point I was making is that the methods of science, which demands open discourse to advance, naturally conflict with government tactics that abuse human rights.

THE POWER OF COMPUTERS

The July article "Trends in Computing: Taking Computers to Task," by W. Wayt Gibbs, suffered from one major omission. The assertion that computers have not helped us "do more work, of increasing value, in less time" may be debatable for commercial and home computing. But it is spectacularly untrue in many areas in science and engineering. The practice of mechanical engineering has changed dramatically toward computer-based design and analysis, yielding increased productivity, better quality, higher safety and faster time to market. Pharmaceutical companies routinely use powerful workstations to discover new drugs far more productively.

People's productivity may be improved or their lives saved by the use of computers that most never buy, use or see.

JOHN R. MASHEY
Portola Valley, Calif.

Letters to the editors should be sent by e-mail to editors@sciam.com or by post to Scientific American, 415 Madison Ave., New York, NY 10017. Letters may be edited for length and clarity.

50, 100 AND 150 YEARS AGO

SCIENTIFIC AMERICAN

NOVEMBER 1947

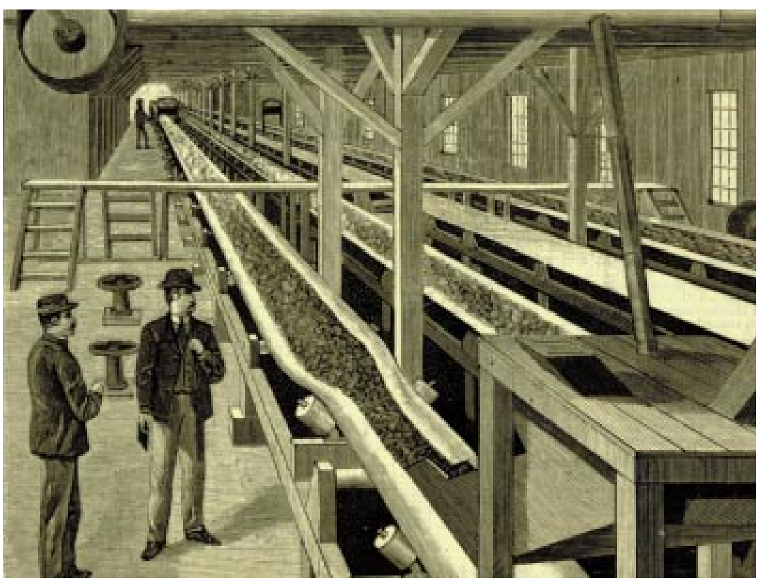
JET THRUST BOOSTED—"Installed downstream from the turbine of a conventional jet engine, a device called an 'after burner' can add more than one third to the power plant's normal propulsive thrust, giving added power for takeoff, during combat conditions, or where extra speed is required. This is accomplished by spraying fuel into the tail-pipe where its combustion adds mass and velocity to the gases of the jet stream. This after burner is in effect a ram-jet engine, where the speed of the air stream in the tail-pipe is well above that needed to make the ram-jet operate. The after burner does not impose any additional stress on the operation of the turbo-jet—a desirable quality since turbo-jet power plants are operating near the critical stress limits of the turbine components."

NOVEMBER 1897

LATEST ON MARCONI—"In Sig. Marconi's recent experiments at Spezia with his 'telegrafo senza fili,' it appears that good telegrams and clear signals were got through at a distance of twelve miles. To the mast of a ship, ninety feet high, a vertical copper wire was attached. Another mast of like height was erected ashore, and the transmitter was attached to its vertical wire. It was also demonstrated that the receiving instruments could be securely placed deep down in the hull of an ironclad war vessel, messages being perfectly intelligible in a cabin eight feet under water, notwithstanding its surroundings of massive iron."

HIGH-ALTITUDE DEATH—"'Alpine misadventure' is a wide word, and includes victims whose sudden fall into a crevasse or mountain torrent is set down to 'loss of balance,' 'misplaced footing,' or one of many mishaps besetting the mountaineer, when syncope—fainting—due to cardiac lesion was the real cause. The hypothesis is strengthened by the death of a burgomeister of a Westphalian town, on the Furka Pass on the Rhone Glacier. The burgomeister, rising in his carriage to get a better view, had barely uttered, 'Oh! C'est magnifique!' when he dropped down dead. The altitude, the rarefied air, the tension—conditions inseparable from Alpine ascents—were too much for a 'chronic sufferer from weak heart.'"

GRAIN SHIPPING—"The phenomenal wheat crop in America for 1897 is estimated at about 500,000,000 bushels. The crops of Europe, however, have been blighted by a disastrous season. Over 200,000,000 bushels of our wheat will be required by the Old World, and the shipment of this vast bulk will materially improve the finances of the companies that carry it across the ocean. The mechanical systems now employed for transshipping grain in the port of New York have proved of great value in reducing time and cost and are capable of handling a vast amount of wheat. Our illustration shows the long belt conveyors that move grain to storage bins or even directly into the holds of waiting ocean steamers."



Mechanical systems for shipping grain

NOVEMBER 1847

TEA IN INDIA—"The Calcutta Gazette informs us that efforts to extend the cultivation of the tea-plant in the northwest of India have been highly successful. The climate and soil in Kemaoon are as suited to the favorable growth of the shrub as the finest Chinese locality. Moreover, the tea-brokers in England have pronounced the Indian tea equal to China tea of a superior class, possessing the flavor of orangepekoe. The price at which tea can be raised is so low as to afford the greatest encouragement for the application of capital. The 100,000 acres available for tea cultivation in the Dhoon alone would yield 7,500,000 pounds, equal to one sixth the entire consumption of England."

ELECTRICITY FROM SUNLIGHT—"Father Maces, Professor of Natural History in the College of La Paix, at Nemour, has just made a discovery of great scientific importance. In a notice in the bulletins of the Royal Academy he has, it is asserted, succeeded in transforming the solar light into electricity. His apparatus, which is extremely simple, spoke several times under the influence of the light, and remained mute without that influence. Even when one witnesses the phenomenon, one scarcely ventures to trust one's own eyes, yet the indications of electricity are evident."

SHIRTS—"A patent has been taken out for dispensing with sewing in the manufacture of shirts, collars, and linen articles. The pieces are fastened together with indissoluble glue. What next?"

NEWS AND ANALYSIS

IN FOCUS

DEATH IN THE DEEP

"Dead zone" in the Gulf of Mexico challenges regulators

FRANKLIN VIOLA

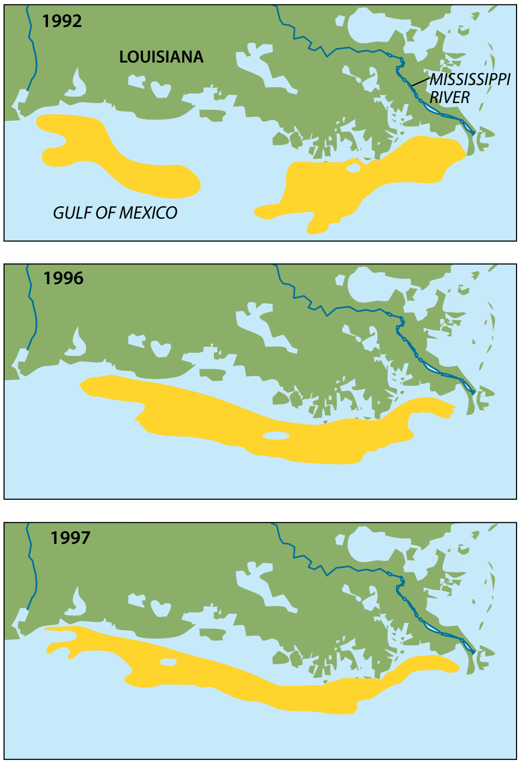
***SUFFOCATED JUVENILE BLUE CRAB** died because of mats of bacteria that thrive in low oxygen levels in the Gulf of Mexico.*

Every spring something goes wrong with the water chemistry in a vast region of the Gulf of Mexico. Oxygen concentrations in the lower part of the water column plummet to a small fraction of normal, sometimes reaching undetectable levels. The suffocating blanket kills or drives away some fish and most bottom dwellers, such as shrimp, snails, crabs and starfish. In the worst-affected areas, the bottom sediment turns black. The so-called hypoxic zone has grown larger in recent years and is now a long tongue the size of Hawaii that licks along the Louisiana coast.

The cause of the phenomenon is no mystery. The Mississippi River, one of the 10 largest in the world, dumps 580 cubic kilometers of water into the Gulf every year; its drainage basin encompasses 40 percent of the land area of the contiguous 48 states. Studies of water samples, sediments from the seafloor and other data show that the amount of dissolved nitrogen in the outflow of the Mississippi and the adjacent Atchafalaya has trebled since 1960. Phosphorus levels have doubled. These elements, present in forms on which single-celled organisms can feed, stimulate the growth of phytoplankton near the sea surface, which provide food for unicellular animals. The planktonic remains and fecal matter then fall to the ocean floor, where bacteria devour them, consuming oxygen as they do so.

The process, known as eutrophication, is familiar to marine and estuarine scientists. Similar episodes have been recorded in partially enclosed seas and basins around the globe: the Chesapeake Bay, the Baltic Sea, the Black Sea and the Adriatic Sea, among others. But the Gulf's eutrophic region is the biggest in the Western Hemisphere. Moreover, it lies in a region that provides the U.S. with more than 40 percent of its commercial fisheries. R. Eugene Turner of Louisiana State University, who together with Nancy N. Rabalais of the Louisiana Universities Marine Consortium pioneered the study of the phenomenon, says fishermen and shrimpers are blaming the hypoxic zone for declines in their catch.

Environmentalists have dubbed the region the "dead zone," a label that overlooks the fact that life is certainly present—but life of the wrong sort. The sea surface may look normal, but the bottom is littered with dead or visibly distressed creatures. In extreme hypoxia it is covered with mats of stinking, sulfur-oxidizing bacteria, according to Rabalais. The hypoxic zone grows more pronounced during the summer but is dissipated by storms and disperses in the fall.



COURTESY OF THE GULF OF MEXICO PROGRAM; LAUREL ROGERS

ZONE OF LOW OXYGEN (yellow) ***in the Gulf of Mexico has grown to extend over 5,500 square miles.***

Rabalais, Turner and others have published detailed papers documenting the association between nitrogen levels in the Mississippi, the rate at which algae called diatoms accumulate on the seafloor and the hypoxic conditions. "We've studied sediment cores," Turner says, "and we have water-quality data from the Gulf for 20 years—good data for 14 years." Good water-quality data for the Mississippi goes back further, to the mid-1950s. Rabalais and Turner have also compared the chemistry of the river with that of other large rivers around the world.

Their work has satisfied most oceanographers that there is indeed a direct link between dissolved nutrients, principally nitrogen, the hypoxia in the lower water column and the ecological changes. "I know the linkages," Rabalais asserts. Few seem inclined to dissent. "They've done a good job," agrees Robert W. Howarth of Cornell University. "The ecological changes are definitely due to hypoxia, and the hypoxia is clearly due to elevated nutrients."

Rabalais and Turner's work pinpoints as a crucial variable the ratio of nitrogen to silicate (from minerals) in the Mississippi outflow. As the amount of nitrogen has increased compared with the amount of silicate, which is slowly declining because of planktonic activity upstream, overall production of plankton in the Gulf has increased. Hypoxia is the result. More alarming changes could be in store. Rabalais suspects the changing nutrient balance might start to benefit noxious flagellate protozoa at the expense of the less harmful diatoms. Toxic algal blooms are indeed becoming more common in the Gulf, as they are in polluted coastal regions around the world. "We are concerned that future nutrient changes could make it worse," Turner says.

The Gulf hypoxic zone represents a grand challenge for environmental policy. The exact geographic origin of the excess nitrogen is a matter of contention. According to the U.S. Geological Survey, most of it—56 percent—is from fertilizer runoff. The biggest contributor, the agency estimates, is the upper Midwest, especially the Illinois basin. Another 25 percent of Mississippi nitrogen is from animal manures. Municipal and domestic wastes, in contrast, account for only 6 percent. "Nitrogen loading has gone up coincidentally with fertilizer use," Turner affirms.

The suggestion that America's breadbasket is the cause of the Gulf's problems has not gone over well with agricultural interests. Turner maintains, however, that the observed effects in the Gulf could be explained by just 20 percent of the fertilizer used in the Mississippi basin draining into the river. New techniques for applying fertilizer hold out the hope of reducing runoff without sacrificing crop yields.

Efforts getting under way to study and perhaps control the hypoxic zone "break new ground," says Don Scavia, head of the coastal ocean program at the National Oceanic and Atmospheric Administration (NOAA) and head of an interagency working group on the hypoxic zone. "The scale of the issue drives it—it is nutrients from 1,000 miles away." NOAA, together with the Environmental Protection Agency, has funded research on hypoxia in the Gulf for several years.

The Mississippi River Basin Alliance and the Gulf Restoration Network, bodies representing users of the land on one hand and of the sea on the other, have joined forces to seek reductions in nitrogen runoff. "Studies won't reduce nutrient loading in the Mississippi River," says Cynthia M. Sarthou of the Gulf Restoration Network. Sarthou states that her organization is looking for ways to encourage voluntary reductions by farmers. The alliance, in contrast, is targeting nonfarm sources. "Some farmers say it's people versus fish," notes Suzi Wilkins of the Mississippi River Basin Alliance. "It's actually farmers versus fishermen."

This past summer agencies launched a far-reaching economic and technical examination of the Gulf hypoxic zone. The aim is to find out about its detailed dynamics, its likely consequences and what remedies might be most effective. The study will adjust for the fact that conventional accounting techniques tend to undervalue the benefits of natural resources, Scavia explains.

The goal is to learn what sacrifices might be worthwhile to restore the region's ecological health. One effort will try to nail down scientifically the question of whether the hypoxia has really caused declines in fish and shrimp catches, as opposed to overfishing, for example. "We should not rely on anecdote," warns Andrew Solow of the Woods Hole Oceanographic Institution. Another segment of the study will use computer modeling to estimate the effects of reductions in nitrogen use. Such reductions are only one possible approach to control, Scavia points out. He suggests that buffer strips of wetland, created to serve as a barrier near the river, might be able to absorb some excess nitrogen.

The scientific assessment is due to be complete in 18 months. But already a management group is looking at measures that could be initiated sooner. "We'll look for win-win solutions within the next two months," Scavia declares. "This can't wait." *—Tim Beardsley in Washington, D.C.*

ASTROPHYSICS

GIRTH OF A STAR

X-ray oscillations help to estimate a neutron star's radius

In a celestial bestiary of oddities, the neutron star holds its own as one of the oddest. Essentially an overblown atomic nucleus, a proverbial spoonful of its substance weighs as much as a mountain. For decades, researchers have been trying to figure out just how large, or rather small, a neutron star is. Now, thanks to a satellite and some luck, they seem to have found a way.

When the dust settles, scientists will have measured a neutron star's size for the first time. As a bonus, they may get to determine just what is inside one: the radius and mass of a neutron star depend sensitively on the nuclear substance contained within. Knowledge of these attributes can thus provide sharp bounds on the nuclear interactions at play.

The breakthrough is owed to the Rossi X-Ray Timing Explorer, a satellite that can measure the arrival time of a photon to within a microsecond. Since late 1996, observers from the University of Amsterdam and the NASA Goddard Space Flight Center have been reporting a curious pattern in x-rays coming from some neutron stars. The photons are arriving in regular pulses of about 1,000 beats per second, when instead a jumble of periodicities had been expected: "As if you go to the piano and lay down your arm," explains Frederick Lamb of the University of Illinois. "Now what we see is like playing a chord, just two or three notes."

The x-ray chord seems to involve material sucked onto the neutron star from a companion. As a clump of gas orbits the neutron star, some material from it streams directly onto the surface, radiating x-rays from the spot where it hits. The patch of radiation follows the orbiting clump around the star (much as the spot thrown on the ground by a police helicopter's searchlight moves with the chopper). When the bright patch goes behind the neutron star, it is hidden, and Rossi sees no x-rays; when in front, the pulse appears.

If this model is right, the clump of gas must be going around the neutron star an incredible 1,000 times per second. Such a high frequency sets a tight bound on the orbit's size. For the most rapid oscillation observed so far, 1,200 hertz, gravitational theory predicts that the orbital radius is a mere 17 kilometers. The star itself must be even smaller. (And in September the Hubble Space Telescope spied a lone neutron star less than 14 kilometers in radius.)

Theorists are still arguing over the exact numbers. The uncertainty hinges on just where the special clump of gas is. William Zhang of the NASA Goddard center and, independently, Philip Kaaret of Columbia University have calculated that the clumps must all be at a "marginally stable" orbit predicted by general relativity: nothing inside this radius can orbit a star but must fall right in. They find that the neutron stars are therefore twice as massive as the sun.

In contrast, Lamb argues that the marginally stable orbit is the least distance at which the clumps can orbit; in actuality they reside farther out, at a so-called sonic point. Beyond *that* radius, the clumps dissipate fast; within it, they last long enough to circle the neutron star a few hundred times. Lamb finds instead an upper bound of 2.2 solar masses for the neutron star. The actual mass, he says, could be much smaller.

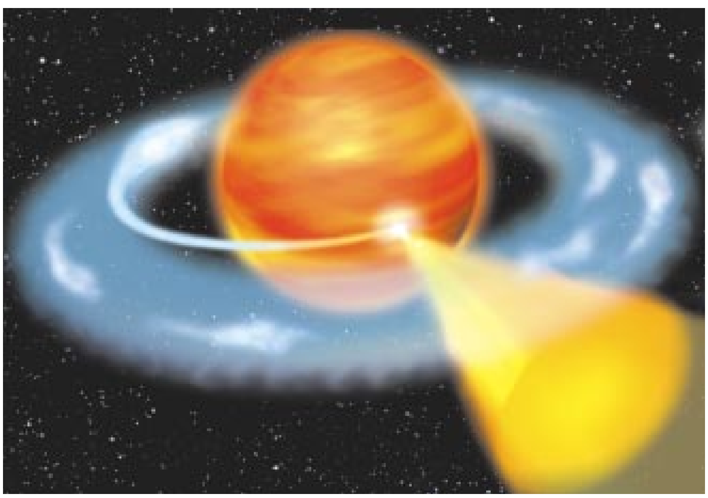
"For the first time, if this interpretation is confirmed, we have accurate limits on radius and mass," Lamb comments. "It begins to limit the possible properties of dense matter." What exactly fills up a neutron star, and how, has never been very clear. That there are neutrons, everyone agrees, but how neutrons interact at such high densities is a mystery. In addition, free quarks, "strange" particles such as kaons and all kinds of weird objects are postulated to show up in massive neutron stars. "Nobody has a completely comprehensive model," muses Robert Wiringa of Argonne National Laboratory, who professes authorship of two of the more "conservative but reliable" ones.

It is not yet clear which of these schemes are endangered by the new observations, but some certainly are. "At any moment detection of a higher frequency would rule out most [models]," Lamb states. His bounds favor "soft" models, in which the nuclear substance is highly compressible. Such material cannot provide much resistance to gravity, so that if enough extra mass falls in from a companion star, the neutron star would readily squeeze into a black hole.

For his part, Zhang feels that the heavy mass he calculates for a neutron star implies that the nuclear matter is "hard": it holds its own against gravity for much longer. His calculations would rule out, for instance, kaons as an essential component of neutron stars: they cannot hold up more than 1.5 solar masses. (Imploding kaon stars would create light black holes, of less than two solar masses. These have never been found, perhaps because black-hole searches only scrutinize objects with at least five times the sun's mass—to make sure that neutron stars are not mistakenly selected.)

As scientists refine their models, Rossi continues to search. Within months, the fine line between neutron stars and black holes may finally be drawn.

—*Madhusree Mukerjee*



SLIM FILMS

***X-RAYS BEAM FROM A NEUTRON STAR** when matter from an orbiting clump falls onto it. As the cloud circles, the beam is seen to blink on and off, allowing the orbital frequency and radius to be measured.*

IN BRIEF

Virus versus Virus

Yale University researchers have redesigned a common cattle virus, called vesicular stomatitis virus, so that it can attack cells infected by HIV, the cause of AIDS. John K. Rose and his colleagues replaced a VSV gene with genes coding for two human proteins. These molecules—normally found on the surface of *T* cells—form a lock of sorts, which the HIV virus picks using one of its own surface proteins, gp120. In this way, HIV enters *T* cells and prompts them to produce more HIV particles. But the cattle virus, armed with the *T* cell lock, blocks this cycle by intercepting HIV particles before they can bud from infected *T* cells. The altered virus is highly specific and lowers the count of HIV particles to undetectable levels in laboratory tests.

Gulf Worms

From the mushroom-shaped mounds of methane ice that seep up through the floor of the Gulf of Mexico, geochemists from Texas A&M University have sampled what appears to be a new species of worm (*head shot at left*). The flat, pinkish, centipedelike creatures, called polychaetes, are one to two inches long and live in dense colonies in the energy-rich ice deposits, some 150 miles south of New Orleans. The researchers speculate that the worms may influence activity within the methane mounds.

ROSEMARY WALSH *Penn State Electron Microscope Facility*

Exotic Mesons

Good news for the Standard Model came from Brookhaven National Laboratory this past summer. Physicists at last tracked the ever elusive exotic meson. A team of 51 researchers from eight institutions spent five years sifting through the mess left when an 18-billion-electron-volt beam of pi mesons hits a liquid hydrogen target. They found that in 500 cases out of 40,000, the collision product did not resemble an ordinary meson, which contains a quark and an antiquark, knotted together by a gluon. Instead the results resembled quark pairs joined by a vibrating gluon string, or gluon-bound quark quartets.

More "In Brief" on page 28

GENETICS

THE FOOD GENOME PROJECT

Sequencing Bessie and her fodder

Aging congressmen have been generous in their support of genomic research that might help what ails them. Now lawmakers are being asked to extend that bounty to crops and farm animals. Spurred by pressure from the National Corn Growers Association (NCGA) for an initiative to sequence corn genes, the U.S. Department of Agriculture is cooking up a $200-million National Food Genome Strategy. That sum, to be spent over four years, would study the DNA of plants, animals and microbes to "enhance the usefulness" of economically important species. A Senate committee has approved the plan in principle.

The proposal still has a long way to go in Congress, but there seems to be strong support for a coordinated attack on the genomes of species that humans rely on for food and fiber. Although the effort to sequence the human genome only recently moved into high gear, early phases of that project, which focused on mapping the locations of genes and different kinds of markers, produced valuable information that promises huge gains for medicine. Boosters of the food genome plan maintain it could lead to comparable leaps forward for agriculture by making it easier to produce genetically altered animals and plants.

Genetically altered soybeans, potatoes, corn, squash and cotton have been widely planted in the past two years, and now rice can be similarly improved [see "Making Rice Disease-Resistant," page 100]. Kellye A. Eversole, an NCGA lobbyist, goes so far as to put numbers on the possible benefits from a food genome project. She foresees a 20 percent increase in production over 10 years.

The USDA plan follows hard on the heels of a Plant Genome Initiative soon likely to be under way at the National Science Foundation. The NSF initiative would focus on a wide range of plants, especially corn, and would continue work on a small mustard plant, *Arabidopsis*, that has already been extensively studied. The Senate Appropriations Committee has allocated $40 million to the NSF plan for next year, although that amount might yet be reduced before legislators sign off on it. The idea of sequencing plants has been endorsed by an interagency task force, which noted in June that Japan has initiated an "extensive" rice genome program. A U.S. plant genome initiative might later be folded into the broader food genome effort that would include farm animals.

Not surprisingly, the prospect of large numbers of federal dollars flowing into new scientific initiatives has prompted some anxieties. Mark E. Sorrells of Cornell University and others have warned against an overemphasis on corn, because its genetic peculiarities make it unlikely that lessons learned from this plant would help improve other crops. The American Society of Plant Physiologists has initiated a letter-writing campaign aimed at ensuring that the NSF initiative does not come at the expense of nongenomic plant research.

More arguments are doubtless in store, but it seems clear that momentum for expanding agricultural genomics is growing. Life down on the farm will soon look very different.

—Tim Beardsley in Washington, D.C.

ALAN LEVENSON *Tony Stone Worldwide*

***FRUITS AND VEGETABLES** are the target of a gene-sequencing effort that could lead to improved crops.*

In Brief, continued from page 24

Is the Black Death Back?

Researchers from the Pasteur Institute in Paris report that a 16-year-old boy in Madagascar contracted a strain of bubonic plague that resisted all modern treatments. Before the advent of antibiotics, the plague claimed masses of victims. In this case, the boy lived, but so did the strain itself, readily introducing its mutated genes into other plague bacteria in a petri dish. Scientists worry that it could spread as easily in nature, either via fleas that have bitten infected rodents or by way of sickly sneezes and coughs.

Totally Random

Lava lamps are not just mesmerizing, they're groovy mathematical tools, too. Robert G. Mende, Jr., Landon Curt Noll and Sanjeev Sisodiya of Silicon Graphics used the familiar retro fixtures to generate truly random numbers—something computers cannot do. They focused a digital camera on six of the liquid-filled cylinders and took periodic shots of their shifting ooze. The camera added its own electronic noise to the resulting image, which was converted into a string of 0s and 1s. Next, the Secure Hash Algorithm (yes, that's its real name), from the National Institute of Standards and Technology, compressed and scrambled the binary string to create a seed value for a standard random-number generator.

LISA KOYUKI SMITH

Guided Gene Therapy

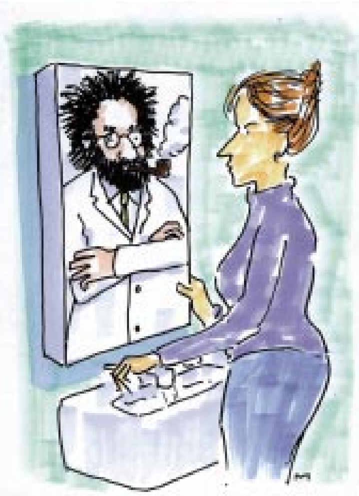
Scientists have struggled to find means for delivering therapeutic genes only to those cells that need them. Often clinicians introduce missing or corrective genes by way of a weakened virus, hoping the virus will infect diseased tissues, express itself and do little harm elsewhere. But the tactic has frequently caused undesirable side effects. Now, however, a group from the University of Chicago has delivered genes, in a viral vehicle, to a specific tissue type in animals. The team, led by Michael Parmacek, attached a therapeutic gene to a newly discovered "on-off" switch, taken from a gene that is activated in smooth muscle. As a result, the therapeutic gene limited its expression to these cells.

More "In Brief" on page 32

ANTI GRAVITY

The Big Picture

Picture a scientist. Now try again, once you erase the image of Einstein from your gedanken blackboard. Since you read this magazine, you may be a scientist, and thus you may have depicted yourself. Unless you're a man in a white lab coat, however, chances are that when most people think of scientists, they're not thinking of you. What they are thinking of was the subject of a study in the August *American Journal of Physics* entitled "Probing Stereotypes through Students' Drawings of Scientists." That article, by Jrène Rahm of the University of Colorado at Boulder and Paul Charbonneau of the National Center for Atmospheric Research, also sums up previous studies on the scientist's image.



MICHAEL CRAWFORD

In 1957 *Science* reported on 35,000 American high school students who were asked to describe a typical scientist. The "average" response: "A man who wears a white coat and works in a laboratory.... He may wear a beard, may be unshaven and unkempt....The sparkling white laboratory is full of sounds: the bubbling of liquids in test tubes ...the muttering voice of the scientist.... He writes neatly in black notebooks." These images obviously represent grand misconceptions—most notebooks would stymie gifted cryptographers and perhaps even pharmacists, the muttering is more likely a grad student wondering if he can sneak away long enough for a game of checkers on Saturday night, and the lab last sparkled when its occupants were developing phlogiston theory.

Nearly 30 years later a 1983 study published in *Science Education* asked more than 4,800 children in grades K through 5 to draw their idea of a scientist. The conceptions were overwhelmingly male, lab-jacketed and adorned with Don King hairdos. The stereotype surfaces in grade 2 and is the image of choice for most fifth graders.

Fearing that public perception is driving students away from science, Rahm and Charbonneau extended the "Draw-a-Scientist" test. They administered it to 49 undergraduates and graduate students enrolled in a teacher certification program: the next generation's teachers. These older, more sophisticated students might be expected to draw a more varied array. The vast majority, however, stayed with the man in the white jacket. Seventy percent of the scientists pictured needed glasses, 58 percent wore lab coats, and 52 percent had facial hair or "extravagant hairdos," a number that may actually be too low to attract the MTV generation. Only 16 percent were clearly female.

A few students went for a reality-based approach. "We had two versions of Einstein," Rahm and Charbonneau write, "and, somewhat more troubling, two of Groucho Marx. Equally troubling, one drawing appeared to be a cross between Konrad Lorenz (in his later years) and Colonel Sanders." (Helpful hint: if the bird is chasing the man, it's Lorenz.)

Although this study doesn't address whether the stereotype drives students away from science, Charbonneau is concerned. "If everybody thinks scientists are crackpots," he says, "they think, 'Hey, I'm not getting into this business.' "

One attempt to buff up scientists' images (male ones, anyway) is the *Studmuffins of Science* calendar, featuring bulging biceps of beefcake Ph.Ds. "I wouldn't do it," Charbonneau says of public flexing, "but it tries to say that scientists can look like actors, the most important people in society." Another attempt, despite its name, is NerdKards, trading cards featuring famous scientists and their stats. The inventor, retired Connecticut teacher Nicholas Georgis, explains that Nerd here stands for Names Earning Respect and Dignity. Unfortunately, the only woman pictured is Marie Curie, and she shares the card with Pierre. Still, it's a first step to a day when a kid wouldn't trade a Harold Varmus for a Ken Griffey, Jr. Smacking homers is cool. Discovering oncogenes is cool and important. —*Steve Mirsky*

In Brief, continued from page 28

Deadly Dinner Date

Entomologists have known that some female fireflies flash their light to attract suitors from another species and then devour those who call. As it turns out, the meal arms the females with a double dose of lucibufagins, chemicals that repulse hungry spiders. Thomas Eisner and his colleagues at Cornell University raised females of the genus *Photuris* in the laboratory and fed *Photinis* males to only some. Although both the males and females produce lucibufagins on their own, spiders ate only those females who had not dined on suitors.

Polar Meltdown

For many years, scientists have warned that global warming will melt away sea ice in the Antarctic, but it has proved hard to demonstrate. Satellite records of sea ice did not exist before the 1970s.

LUIS ROSENDO *FPG International*

New work, though, has confirmed what most feared: by studying whaling records, William de la Mare of the Australian Antarctic Division of the Department of the Environment, Sport and Territories has found that between the mid-1950s and the early 1970s the sea ice edge in the Antarctic most likely receded some 2.8 degrees in latitude—representing a 25 percent reduction. Because whales are most often caught near the sea ice edge, records of their capture—logged by the Bureau of International Whaling Statistics since 1931—implicitly contain information about the extent of sea ice in the region.

Welcome to Mars

In September, after a 300-day cruise, the National Aeronautics and Space Administration's Surveyor spacecraft at last entered orbit around Mars. Now it will take another four months before the 2,000-pound probe produces any results. Surveyor must first spiral in closer to the red landscape it is there to map, using an innovative "aerobraking" tactic: with each pass of the planet, Surveyor dips lower into the atmosphere. The resulting air resistance slows the craft, which then covers less ground on its next go-round. Once Surveyor is finished mapping Mars, it will serve as a communications satellite. *—Kristin Leutwyler*

SA

FIELD NOTES

A REAL DIVE

The U.S. Navy's Experimental Diving Unit is a diver's heaven—and hell

I've got 10 kids out in the Gulf today diving," says Lt. Commander Robert Mazzone, his outstretched arm indicating the blue Gulf of Mexico framed by his office window. "They're in 87-degree water, using state-of-the-art equipment. And they're getting paid to do it," he adds, not quite believing it himself.

It's all in a day's work at the U.S. Navy's Experimental Diving Unit in Panama City, Fla. The NEDU's daunting main mission: to make sure that the equipment and especially the often complex breathing gear used by navy divers—including the exotic, $45,000 "rebreathers" used by the navy's elite Explosive Ordnance Disposal teams—do not, well, let them down.

The NEDU is officially responsible only to the U.S. military's diving community. Yet the "Authorized for Navy Use" list (http://www.navsea.navy.mil/sea00c/pdf/anu.pdf), which is compiled from the NEDU's labors, has become a kind of consumer guide for recreational divers throughout North America and Europe. Then, too, recent regulations have made it impossible to sell diving equipment anywhere in the European Union unless it has made it through the NEDU's rigorous testing procedures. "The NEDU is probably the only organization to have developed rigorous, mathematically based procedures for testing underwater equipment," notes John R. Clarke, the unit's scientific director.

Besides testing gear, the NEDU does occasional studies on physiological aspects of diving. During my visit in August, researcher Marie E. Knafelc was studying how the human ear works underwater, in hopes of coming up with better regulations to protect the hearing of divers who work with power tools. "Divers seem to have more hearing loss than nondivers," she explains and discounts the possibility that the loss is pressure-related. As she speaks, pairs of navy divers enter a large outdoor test pool and are exposed to the noise of underwater power tools.

For its main mission, the NEDU puts equipment through a battery of tests, beginning with ones that do not put human beings at risk. If the gear passes those trials, it makes it to "the monster," the largest hyperbaric chamber in the U.S. that can be compressed to deep depths and the centerpiece of the NEDU's testing facilities. Sealed in the chamber, navy divers test equipment at high pressure in any of the five subchambers full of breathing gas or underwater in a large "wet pot" below the subchambers.

The chamber can be pressurized to a depth of 610 meters (2,000 feet). But only one or two of the 600 dives a year done in the chamber get down to 300 meters or more. Such deep dives take about 30 days. For physiological reasons, at least seven different gas mixtures are required at those pressures, to keep the divers from suffering the toxic effects of oxygen or the narcotic effects of nitrogen. Different gases are used at

GLENN ZORPETTE

"WET POT" surrounds experimental diver David Junkers.

different pressures, and by the time the divers reach 300 meters, they are breathing 3 percent oxygen and 97 percent helium (for comparison, air is 21 percent oxygen and nearly 79 percent nitrogen).

At such pressures, and with the helium gas, speech is utterly unintelligible. The divers speak into microphones that relay their voices to descramblers and then on to headphones so that they can understand one another. For reasons that are not entirely understood, the senses of smell and taste are significantly diminished, so food for divers is invariably loaded with spices. Bread and muffins take on the consistency and texture of a rubber ball.

Not just the barometric pressures are extreme. According to Master Chief Diver David Junkers, a veteran of 1,000 experimental dives, divers toil continuously from 6 A.M. to 5 P.M., with occasional after-dinner chores as well. "We have had occasional problems," Junkers notes, including a fistfight now and then at high pressure. But careful screening of dive teams keeps such flare-ups to a minimum. Navy divers are also notorious for finding creative ways of blowing off steam. "You get guys who are exhibitionists," Junkers explains. "And some guys are pretty good artists; they'll draw cartoons about the guys outside locking them in."

"It gets pretty rude and crude in there sometimes," Junkers adds with a shrug. "You could be eating a meal, and the guy next to you is [going to the bathroom]. You can't be too squeamish."

—Glenn Zorpette

PEST MANAGEMENT

BIOLOGICAL NONCONTROL

My enemy's enemy may be no friend

For decades, biologists have been fighting fire with fire by releasing exotic organisms, often insects, to attack pests and weeds that threaten crops and ruin rangeland. New research has shown that a weevil brought to North America to devour an invader called musk thistle is also damaging relatively harmless thistles belonging to a different genus. The finding has prompted investigators to put on hold experimental releases of another exotic insect that they were hoping would join the fight against musk thistle.

Musk thistle arrived in North America in the mid-19th century. The Eurasian weevil *Rhinocyllus conicus* was first released to combat it in 1968, and releases continue. The insect's larvae eat into the thistle's flower heads and feed on the seeds there. Paul E. Boldt of the U.S. Department of Agriculture's Grassland, Soil and Water Research Laboratory in Temple, Tex., estimates that *Rhinocyllus* saves farmers hundreds of millions of dollars every year because it allows them to use less herbicide.

But in what Peter B. McEvoy of Oregon State University terms a "dogged" piece of research, Svata M. Louda of the University of Nebraska–Lincoln and her colleagues have found that *Rhinocyllus* larvae are also feeding on flower heads of five native thistles, comparatively innocent bystanders belonging to the genus *Cirsium*. At one site the weevil reduced seed production in a *Cirsium* species by 86 percent. Louda, who published her findings in August in *Science,* suggests the Eurasian weevil might next attack a related and ecologically very similar North American thistle that is officially listed as threatened. The weevil has spread rapidly during this decade and is apparently now also outcompeting populations of a native insect that feeds on thistles.

Louda's results play into a long-running controversy. In 1995 the now defunct U.S. Office of Technology Assessment said in a report that any untoward ecological effects of biological-control programs "have probably gone unnoticed" because nobody systematically searches for them. Yet despite the lack of follow-up investigations, multiple exotic species are often introduced, one after another, to fight the same target organism. "There is no theory to indicate that this is wise," says Donald R. Strong of the University of California at Davis. "The situation is becoming serious because the rate of approvals requested for biological control is going up rapidly."

Researchers in the 1960s showed in tests that *Rhinocyllus* preferred the target musk thistle to several native thistles. But "the weevil was known to feed outside its intended target species," says James Nechols of Kansas State University. Boldt adds that today researchers are more cautious about preventing damage to native species than they were 30 years ago. The USDA proposed strengthening its regulations on biological-control schemes three years ago but ran into opposition from proponents who feared burdensome additional requirements.

This past spring, after gaining USDA approval, Boldt started to release experimentally a new exotic organism to control musk thistle—the flea beetle *Psylliodes chalcomera*. This flea beetle's breadth of diet was tested earlier in cages on at least 55 plant species, including some native *Cirsium* thistles, Boldt notes. These tests showed that flea beetle adults ate and oviposited in one *Cirsium* species, but their larvae, which are generally more damaging, indulged in only "a little nibbling." Reassured by these results, Boldt released several hundred in Texas, and Nechols may have accidentally allowed some to escape in Kansas when a storm blew over testing cages.

Nechols thinks the insect would most

SVATA M. LOUDA; GAETANO CAMPOBASSO *USDA-ARS European Biological Control Laboratory, Rome (inset)*

NATIVE THISTLE
Cirsium canescens *is being threatened by weevils imported to control musk thistle. Tests for using the flea beetle from Europe* **(inset)** *for similar biological control were put on hold.*

likely cause less damage to nontarget thistles than *Rhinocyllus* does. But Strong has doubts about the assessment process that gave the thumbs-up to the flea beetle project, saying the process is susceptible to political influence. "The data in the original literature and on the final approval don't look like the same insect," he states.

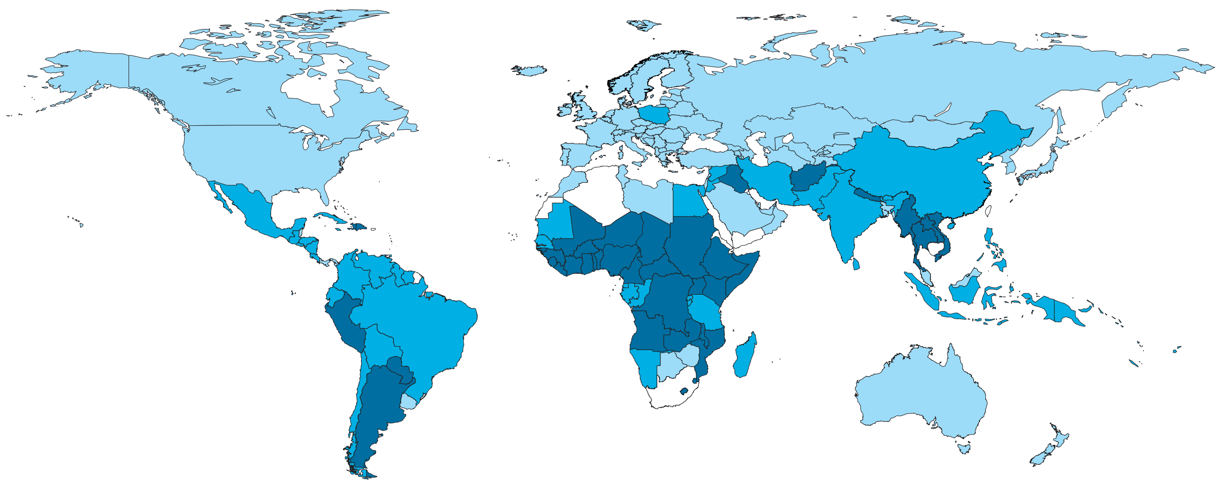
In any event, with the publication of Louda's results, Boldt and Nechols have voluntarily suspended further flea beetle releases until they have better information. The insect was not tested on rare thistles, Boldt explains, because their seeds, needed for experiments in enclosed cages, are hard to come by.

Louda's findings will probably be thoroughly studied at the USDA, where efforts are now under way to craft new compromise regulations on introduced biological-control organisms. Strong believes carnivorous insects, in particular, at present get an easy ride: he suggests some ladybugs introduced to kill other insects may have eliminated local native ladybug populations. "It's chilling," Strong observes, "and there is no public dialogue."

—Tim Beardsley in Washington, D.C.

BY THE NUMBERS

Access to Safe Drinking Water



PERCENT OF POPULATION IN URBAN AREAS HAVING ACCESS TO CLEAN WATER

LESS THAN 75 | 75 TO 94.9 | 95 OR MORE | NO DATA

SOURCE: The World Resources Institute. Data are based on surveys of national governments in 1980, 1983, 1985, 1988 and 1990.

In 1848 and 1849 up to a million people in Russia and 150,000 in France died of cholera, the classic disease of contaminated water. Typhoid fever, another disease transmitted through water, was most likely responsible for the deaths of 6,500 out of 7,500 colonists in Jamestown, Va., early in the 17th century; during the Spanish-American War, it disabled one fifth of the American army.

Today waterborne disease is no longer a major problem in developed countries, thanks to water-purification methods such as filtration and chlorination and to the widespread availability of sanitary facilities. But in developing countries, waterborne and sanitation-related diseases kill well over three million annually and disable hundreds of millions more, most of them younger than five years of age.

Bacterial and viral diseases contracted by drinking contaminated water include, in addition to cholera and typhoid, childhood diarrheal ailments, infectious hepatitis and poliomyelitis. Drinking water may also be contaminated with parasites, such as those that cause ascariasis, a disease in which large worms settle in the small intestines, and dracunculiasis (guinea worm), in which ingested larvae mature internally and eventually burst through the skin. Water-related illnesses are also spread through food, hand-to-mouth contact or person-to-person contact. Some are transmitted primarily when skin and nematode come together in unsanitary waters; examples are schistosomiasis, which causes anemia and enlargement of the liver and spleen; trachoma, the leading cause of blindness in humans; and hookworm, which causes anemia, gastrointestinal disturbance and other problems.

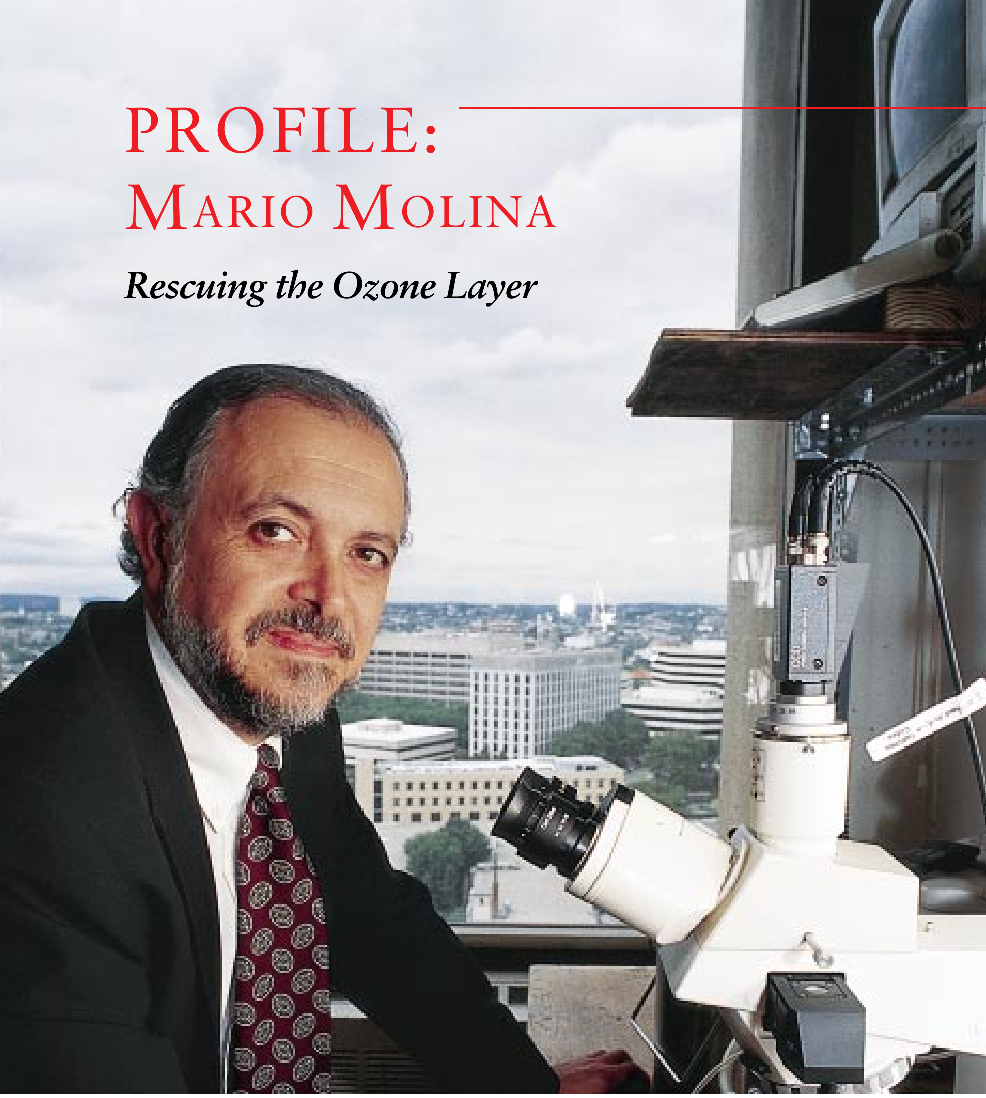
The map shows the percent of urban populations with access to safe drinking water. (Those in urban areas, particularly in developing countries, have better access than rural residents.) Of all developing regions, sub-Saharan Africa has the lowest access to safe water and the highest mortality rate from water-related disease. In Abidjan, Ivory Coast, for example, 38 percent of the city's population of almost three million have no access to piped water, and 15 percent have no toilets and so must defecate in the open. China, on the other hand, has a high level of access to safe water and one of the lowest mortality rates from these diseases in the developing world. Mortality rates from water-related disease are high in India and the Middle East, somewhat lower in some of the non-Chinese parts of Asia, and lower still in South America.

Around the world a billion people lack access to safe water, and 1.8 billion do not have adequate sanitary facilities. According to one estimate, providing safe water and decent sanitation facilities for all human beings would cost $68 billion over the next 10 years—an enormous sum, but equivalent to only 1 percent of the world's military expenditures for the same period. *—Rodger Doyle (rdoyle2@aol.com)*

RODGER DOYLE

PROFILE: MARIO MOLINA

Rescuing the Ozone Layer



JESSICA BOYATT

Mario Molina is walking me through his laboratory at the Massachusetts Institute of Technology, which is overflowing with exotic equipment. He makes his way to a small room in the back of the lab where he points out one of his latest toys, a powerful microscope hooked up to a video camera. He details how he and his students designed this high-tech setup to watch the formation of cloud particles. Despite his enthusiastic description, my mind wanders—I'm distracted by the dazzling clouds visible (without magnification) through the large window over Molina's shoulder. Somehow I did not expect that the man who suggested that chlorofluorocarbons (CFCs) were destroying the ozone layer, some 20 kilometers above our heads, would use a microscope to probe the vast expanses of the atmosphere.

But within the confines of his laboratory, the Nobel Prize–winning Molina has seen quite a bit—much of it troubling. Molina is not an alarmist by temperament: "I've never claimed the world was coming to an end," he chuckles, yet a hint of seriousness remains in his gentle voice. When Molina and his colleague F. Sherwood Rowland of the University of California at Irvine announced their CFC findings in 1974, it seemed to many people that, in fact, the sky was falling.

Damage to the protective ozone layer, which shields the earth's surface from harmful ultraviolet radiation, would mean outbreaks of skin cancer and cataracts as well as the loss of crops and wildlife. So great was the concern that 10 years ago this fall, governments around the globe outlawed CFCs by signing the Montreal Protocol on Substances That Deplete the Ozone Layer.

The reluctant Cassandra of the chemistry world started out just having fun. As a young boy, he showed an interest in chemistry, so his indulgent parents allowed him to convert one bathroom in the spacious family home in Mexico City into a private laboratory.

After boarding school in Switzerland and graduate schools in Germany and France, Molina made his way to the University of California at Berkeley to complete his Ph.D. in physical chemistry. When he arrived in 1968, the campus was embroiled in student unrest about the Vietnam War. His time at Berkeley served as an awakening for him about the significance of science and technology to society. (Molina's time there had a personal significance as well: fellow graduate student Luisa Tan would later become his wife and frequent research collaborator.) Molina's project was rather academic: using lasers to study how molecules behave during chemical reactions. But because laser technology also can be used in weapons, the work was unpopular with student activists.

"We had to think of these issues: Why are we doing what we are doing? Would the resources be better spent in some other way? Is science good or bad?" Molina asks, waxing philosophical. "I came to the conclusion that science itself is neither good nor bad." Technology—what people do with science—was another story.

A desire to understand the implications of technology led Molina to study CFCs during a postdoctoral fellowship under Rowland. "All we knew is that these industrial compounds were unusually stable. We could measure them everywhere in the atmosphere," Molina says. "We wondered: What happens to them? Should we worry?"

The irony of CFCs is that years ago they were initially valued precisely because there seemed to be no need to worry. At a 1930 meeting the inventor of the compounds inhaled CFC vapors and then blew out a candle to show that the chemicals were neither toxic nor flammable. Over the next 50 years, CFCs made an array of new technologies possible: modern refrigerators, household and automobile air conditioners, aerosol spray cans, Styrofoam, cleaning techniques for microchips and other electronic parts.

Most emissions, such as exhaust from cars and smokestacks, actually never get very high in the air—the pollutants react with the hydroxyl radical (OH), which is essentially an atmospheric detergent that makes compounds soluble in rainwater. Molina checked to see how fast CFCs would react with hydroxyl radicals. The answer: zip. "It seemed that maybe nothing whatsoever interesting would happen to them," he says.

If chemicals could not break down CFCs, perhaps sunlight would. Based on their laboratory observations, Rowland and Molina realized that in the stratosphere, ultraviolet radiation is sufficiently energetic to break apart CFC molecules, releasing, among other substances, highly reactive chlorine atoms. Small amounts of chlorine can destroy ozone by acting as a catalyst (that is, the chlorine is not used up in the process of breaking down ozone).

In June 1974 Rowland and Molina published their paper in the journal *Nature* proposing a connection between CFCs and destruction of the ozone layer. Much to their surprise, the article received little notice. A few months later the two held a press conference at a chemistry meeting. "Eventually, we caught people's attention," Molina says.

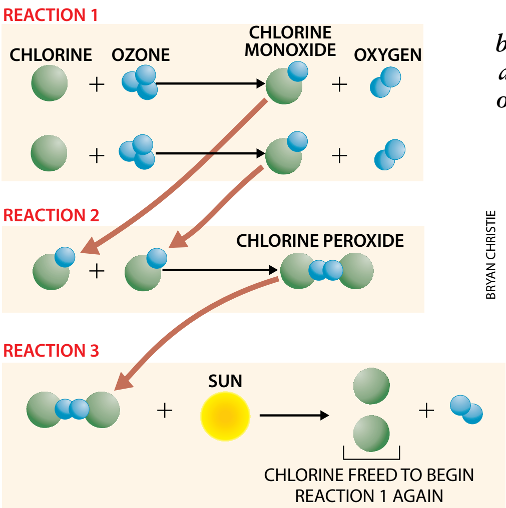
Indeed. Over the next few years, letters about CFCs poured into Congress—the final tally is second only to the number received about the Vietnam War. The government responded quickly, passing amendments to the Clean Air Act in 1977 that called for the regulation of any substance "reasonably anticipated to affect the stratosphere." Soon the use of CFCs as propellants in spray cans was banned in the U.S. Chemical companies began to seek alternatives to CFCs; compounds known as hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs) are the most common choices. (Although HCFCs still contribute to ozone depletion because they contain chlorine, they are not as hazardous as CFCs, because they typically fall apart before reaching the stratosphere. The HFCs pose no threat to the ozone layer.)

Significantly, this flurry of action took place despite the fact that no one had ever observed any loss of stratospheric ozone. The famous hole in the ozone layer above Antarctica was not even detected until 1985. Molina commends this "important precedent in the use of precautionary principles" and suggests that the need to "do something even though the evidence is not there [is] very typical of environmental issues."

A more comprehensive international treaty regulating CFCs took longer to negotiate. But in September 1987 more than two dozen countries signed the Montreal Protocol. The agreement imposed an immediate reduction in the production and use of CFCs; subsequent amendments led to a total phaseout of CFCs in developed countries in 1995 (developing countries have until 2010).

Although the Montreal Protocol was signed after the discovery of the Antarctic ozone hole, many scientists and policymakers at the time were still unsure whether the ozone hole had been caused by CFCs or whether it was just part of a natural cycle. Molina himself remembers that when he first heard news of the ozone hole he "had no idea" whether CFCs were truly to blame. To prove the connection between CFCs and the Antarctic ozone hole, Molina and his wife proposed a new series of chemical reactions in 1987 that measurements confirmed in 1991.

That satisfied most science and policy experts, although a few critics still persist. As late as 1995 (ironically, the same year Molina won the Nobel Prize for Chemistry, along with Rowland and Paul J. Crutzen of the Max Planck Institute for Chemistry in Mainz, Germany), Congress held hearings questioning whether the ozone hole was real and, if so, whether CFCs were really the culprit. The state of Arizona declared the Montreal Protocol invalid within its



CHLORINE DESTROYS OZONE ***but is not consumed in the process. Mario and Luisa T. Molina proposed this series of reactions to explain how CFCs caused the Antarctic ozone hole.***

BRYAN CHRISTIE

boundaries. Molina's patience is clearly tried by these suggestions. "You can go to the stratosphere and see how much chlorine there is and convince yourself that it's coming from CFCs," he says, his voice rising.

In the scientific community, the ozone problem is basically settled. Today the challenges lie more in the area of enforcing the Montreal Protocol. (The latest concern: a burgeoning black market in CFC trade.) Molina and his research group have moved on as well, investigating a wide range of reactions that occur in the atmosphere, including some that are important in urban air pollution. And Molina now spends less time in the lab and more time speaking to government officials on policy questions. In 1994 President Bill Clinton appointed him a science and technology adviser to the administration.

Molina also encourages students from developing countries, particularly in Latin America, to study environmental sciences. (He is the first Mexican-American to win a Nobel Prize and the first person born in Mexico to win in the sciences.) Part of his prize money has gone to create a fellowship for these students to study in the U.S. Given the environmental problems faced by developing nations, including deforestation, desertification, and worsening water and air pollution, Molina considers it crucial to involve people from these regions when crafting solutions.

Molina's smog-choked hometown offers a poignant tale. "When I was a kid in Mexico City, [pollution] was not a problem," he recalls. Over the past 50 years, of course, that has changed. Molina finds it puzzling that more is not done to combat pollution in cities, which is so plainly obvious compared with CFC pollution in the stratosphere. "You can already see it and smell it and breathe it," he comments.

Molina hopes this argument will convince policymakers, specifically in the developing world, to reduce emissions of fossil fuels now, a move that should also help alleviate global warming. Although Molina sees the evidence linking fossil fuels and climate change as still somewhat tentative, the connection between fossil fuels and urban pollution is unequivocal—and thereby on much firmer footing than the CFC-ozone depletion connection was when controls on CFCs were established. "If we take a look at the whole picture, it is much clearer to me that some strong action needs to be taken on the energy issue." Interesting what shows up in Molina's microscope. —*Sasha Nemecek*

GEORGE LEPP *Tony Stone Images*

***DRUG FACTORY OF THE FUTURE?* Corn can be mutated to make human anticancer proteins.**

BIOTECHNOLOGY

PLANTIBODIES

Human antibodies produced by field crops enter clinical trials

Down a country road in southern Wisconsin lies a cornfield with ears of gold. The kernels growing on these few acres could be worth millions—not to grocers or ranchers but to drug companies. This corn is no Silver Queen, bred for sweetness, but a strain genetically engineered by Agracetus in Middleton, Wis., to secrete human antibodies. This autumn a pharmaceutical partner of Agracetus's plans to begin injecting cancer patients with doses of up to 250 milligrams of antibodies purified from mutant corn seeds. If the treatment works as intended, the antibodies will stick to tumor cells and deliver radioisotopes to kill them.

Using antibodies as drugs is not new, but manufacturing them in plants is, and the technique could be a real boon to the many biotechnology firms that have spent years and hundreds of millions of dollars trying to bring these promising medicines to market. So far most have failed, for two reasons.

First, many early antibody drugs either did not work or provoked severe allergic reactions. They were not human but mouse antibodies produced in vats of cloned mouse cells. In recent years, geneticists have bred cell lines that churn out antibodies that are mostly or completely human. These chimeras seem to work better: this past July one made by IDEC Pharmaceuticals passed scientific review by the Food and Drug Administration. The compound, a treatment for non-Hodgkin's lymphoma, will be only the third therapeutic antibody to go on sale in the U.S.

The new drug may be effective, but it will not be cheap; cost is the second barrier these medicines face. Cloned animal cells make inefficient factories: 10,000 liters of them eke out only a kilogram or two of usable antibodies. So some antibody therapies, which typically require a gram or more of drug for each patient, may cost more than insurance companies will cover. Low yields also raise the expense and risk of developing antibody drugs.

This, Agracetus scientist Vikram M. Paradkar says, is where "plantibodies" come in. By transplanting a human gene into corn reproductive cells and adding other DNA that cranks up the cells' production of the foreign protein, Agracetus has created a strain that it claims yields about 1.5 kilograms of pharmaceutical-quality antibodies per acre of corn. "We could grow enough antibodies to supply the entire U.S. market for our cancer drug—tens of thousands of patients—on just 30 acres," Paradkar predicts. The development process takes about a year longer in plants than in mammal cells, he concedes. "But start-up costs are far lower, and in full-scale production we can make proteins for orders of magnitude less cost," he adds.

Plantibodies might reduce another risk as well. The billions of cells in fermentation tanks can catch human diseases; plants don't. So although Agracetus must ensure that its plantibodies are free from pesticides and other kinds of contaminants, it can forgo expensive screening for viruses and bacterial toxins.

Corn is not the only crop that can mimic human cells. Agracetus is also cultivating soybeans that contain human antibodies against herpes simplex virus 2, a culprit in venereal disease, in the hope of producing a drug cheap enough to add to contraceptives. Planet Biotechnology in Mountain View, Calif., is testing an anti-tooth-decay mouthwash made with antibodies extracted from transgenic tobacco plants. CropTech in Blacksburg, Va., has modified tobacco to manufacture an enzyme called glucocerebrosidase in its leaves. People with Gaucher's disease pay up to $160,000 a year for a supply of this crucial protein, which their bodies cannot make.

"It's rather astounding how accurately transgenic plants can translate the subtle signals that control human protein processing," says CropTech founder Carole L. Cramer. But, she cautions, there are important differences as well. Human cells adorn some antibodies with special carbohydrate molecules. Plant cells can stick the wrong carbohydrates onto a human antibody. If that happens, says Douglas A. Russell, a molecular biologist at Agracetus, the maladjusted antibodies cannot stimulate the body into producing its own immune response, and they are rapidly filtered from the bloodstream. Until that discrepancy is solved, Russell says, Agracetus will focus on plantibodies that don't need the carbohydrates. Next spring the company's clinical trial results may reveal other differences as well.

—W. Wayt Gibbs in San Francisco

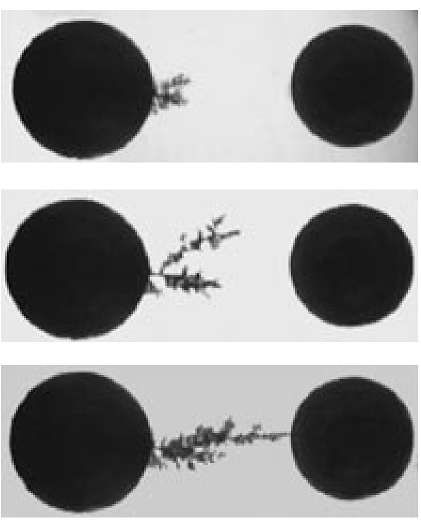
NANOFABRICATION

FROM CHIPS TO CUBES

Chemists make self-growing microcircuits

If you want to pack more circuitry into an electronic gadget—and in the world of electronic gadgets, more is almost always better—you have to use smaller wires. Engineers have two tools to do this, microsoldering and photolithography, both of which have proved phenomenally successful. But both are also pressing against known limits. To keep computer sophistication racing forward at its rocket sled pace, semiconductor outfits will need a fundamentally new way to build ever denser microcircuitry. Jean-Claude Bradley, a chemist at Drexel University in Philadelphia, thinks he is on to one. If his technique works as hoped, it might be used, decades from now, to make microprocessors that look more like cubes than chips.

The first step, however, is a much more modest one. Bradley and his colleagues created two copper wires to make an exceedingly simple circuit that lights up a tiny bulb. What is interesting is not so much what they did but what they did not do: they did not use any of the standard and experimental techniques for building circuitry. No robot-controlled soldering pens. No ultraviolet lamps or light-sensitive acid washes to etch micron-size wires. No marvelously detailed printing plates to stamp out a circuit pattern.



COURTESY OF JEAN-CLAUDE BRADLEY

***COPPER BEADS* bathed in water and an electrical field extend tendrils to form a connection.**

Bradley used only decidedly low-tech gear. "We start off with a project board just like you'd buy at Radio Shack," he says. The board is covered with a grid of holes, each hole capped by a copper ring. Bradley covered two adjacent rings with a single drop of water, then stuck platinum electrodes into the bottom of the holes so that they were close to, but did not contact, the rings. He plugged the electrodes into the rough equivalent of two nine-volt batteries. Almost immediately, a branch of copper began growing from one ring toward the other. Within 45 seconds, the wire completed the circuit.

"This is the first example of constructing circuitry simply by controlling an electrical field," Bradley asserts. "You don't need to touch the copper rings in any way." Indeed, in a paper published in the September 18 issue of *Nature,* Bradley reported that his lab has grown finer wires less than a micron thick—nearly as thin as the wires in computer chips—between copper particles floating freely in a solvent. But it will take much more work to create complex microcircuits using electrodeposition.

Bradley says electrochemists understand in rough terms why this process works. The voltage applied to the platinum electrodes creates an electrical field that surrounds the two copper rings. The field polarizes the copper: it forces positive charges to one side and negative charges to the other. The same thing happens to both rings, so if the two are side by side, the positive edge of one ring will face the negative edge of the other. Opposites attract, and in a strong field, the opposite edges can attract so strongly that the electrical force will rip copper atoms off one ring and dump them into the water-filled gulf between the two. Once enough copper atoms are in the water, they begin to coalesce into a solid wire, which grows until it contacts the other ring and creates a conduit that nullifies the voltage difference between the two rings.

That explains why the wires grow, but Bradley admits that many mysteries about the phenomenon will have to be solved before electrodeposition will yield useful circuits. The wires form branching, treelike structures, for example. Smooth wires conduct higher currents and higher frequency signals more readily. And computer logic is made from semiconductors such as silicon, as well as conductors such as aluminum. Bradley thinks he can probably make smooth semiconductor circuitry by using different materials and solvents and by strengthening the electrical field. But he has yet to prove this.

Perhaps more important, chemists still need to demonstrate what Bradley claims is "the technique's real potential: to construct truly three-dimensional circuits." Acid etches, soldering guns and printing plates work well only on flat surfaces; that is why microchips are so thin. But if metal particles are suspended within a porous cube, Bradley speculates, one could then use a mesh of electrodes or beams of polarized light to generate minute electrical fields and in this way to grow wires that run up and down as well as to and fro. Now that Drexel has applied for provisional patents, Bradley has begun looking for industrial partners to bankroll the next step in his research: to make circuits that are as tall as they are broad.

—W. Wayt Gibbs in San Francisco

APPLIED GEOLOGY

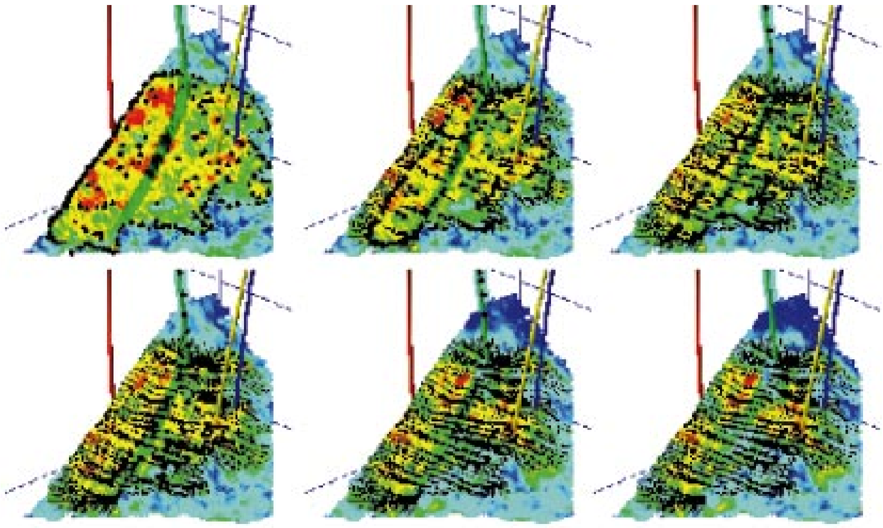
OIL IN 4-D

Time-lapse software boosts oil recovery

Digging for oil used to be like mining for gold. The treasure had to be haphazardly pried from sheets of rocks, pools of water and heaps of debris. Until the 1980s, only one barrel of oil could be removed for every two that lay below. Then, with a technique that mapped oil fields three-dimensionally, an extra half barrel could be recovered. Now, by organizing those 3-D images over time, engineers hope to extract two barrels out of every three. Their technique, called time-lapse imaging, helps to locate hidden oil reserves and complements new methods for hitting lost oil. These advances come at a good time—experts estimate that in 45 years the world's remaining one trillion gallons of oil will have been depleted.

Researchers at the Columbia University Lamont-Doherty Earth Observatory were the first to think of applying the fourth dimension—that is, time—to oil production. As often occurs in scientific breakthroughs, an unsolved mystery drew Roger Anderson's lab workers to the Eugene Island field, in the Gulf of Mexico, in 1991. After 20 years of pumping, the field had yielded twice what it should have based on standard expectations. Perplexed, the scientists lobbied for money from the Department of Energy and several oil companies to study the nine-square-mile basin. By combining maps from 1985 to 1994, they charted a visual history of the site and eventually found oil trickling from deep reservoirs below. In the process they caught a glimpse of the complex forces driving oil upward. “It was one of those serendipitous discoveries. We went in looking to see how an oil field charges itself, and instead we found out how it was draining,” Anderson says.

That information, coupled with dramatic advances in computer power, made the old idea of incorporating temporal data into flow models viable. Indeed, Lamont’s program, called Lamont-Doherty 4-D Software, is changing oil exploration the same way time-lapse imaging revolutionized weather forecasting and medical imaging. With 4-D, geoscientists can simulate drainage with different drill placements and find bypassed reserves by observing oil and gas flows over time.



LAMONT-DOHERTY EARTH OBSERVATORY

DRAINAGE SIMULATION OF UNDERGROUND OIL *shows how oil* (black dots) *trickles toward a well over time. Red spots are oil deposits that could be tapped with new wells.*

The 4-D images, which can show clusters of oil and gas wobbling like Jell-O against water pockets, rock slabs and salt pillars, are derived from low-frequency sound waves. Taken successively, echoes from the waves map the features of an oil field over time. Oil companies then plug the seismic data into the software. Tapping those secret stores then requires the help of another recent innovation: the flexible drill pipe, or well. Unlike traditional wells, these can snake across long swaths of oil and mud.

The 4-D software, which is now being tested in the North Sea and the Gulf of Mexico, came about after Lamont teamed up in 1995 with Western Atlas International, an oil-field service company. In what represents a growing trend, Western Atlas funded the software's development in exchange for exclusive rights to the end product. "Now that the cold war is over, places like Columbia are thinking more practically," says Anderson, who leads the project. "Unlike government funding of science, industry pays for value rather than cost. It removes some of the practicality from science and replaces it with past productivity and performance."

Unlike its major competitors in the time-lapse business—Schlumberger and Petroleum Geo Services—Lamont-Doherty processes and analyzes the data in one application, a more qualitative but less costly solution. Companies can also buy the program (for about $100,000) and interpret the information themselves, saving millions of dollars, as well as adapt it to in-house strategies. "We can mix and match ideas from Lamont with our own internal work," says James Robinson, a scientist at Shell who uses the program. "It's good at seeing where things have moved, quickly." And the software can be used to enhance other techniques that pull more oil out of a field, such as adding carbon dioxide, microbacteria, heat or water to fields.

Although 4-D and related technologies will allow on average 65 percent of a field to be drained, Lamont-Doherty plans to hit the 75 percent mark by making its program interactive. This would do for Exxon's oil rigs what CAD-CAM, or computer-automated design and manufacturing, did for Boeing's 777. Scientists could go from simulated drilling to actual pumping with a keystroke.

To get there, scientists still need to understand how the incomplete vacuum of a well interacts with pockets of fluid and gas, which vary in density. "To make a really good flight simulator, you have to have a model of how the plane works. In the oil field, the model's missing. Right now we're just observing the drainage—we don't really know the physics," Anderson remarks. With that knowledge, the program will be able to predict oil flows and revise drainage information in real time—and stave off the inevitable depletion of the earth's black gold. *—Kimberly Martineau*

MACHINATIONS

"PLEASE, NO DOUBLE-STICKY TAPE"

Death and destruction—with sportsmanship—in Robot Wars

The technofunk fades, and the competitors lie sprawled across the floor, their bodies still. Some have been slammed by the swatting arms attached to the wall of the arena, some have been punctured by an evil-looking spike that periodically lowers to feast on the contestants, and others have simply been battered senseless during the matches. In a somber voice the announcer probes for signs of life: "Ziggy, can you move? Razor Back, can you move? Gator, can you move?"

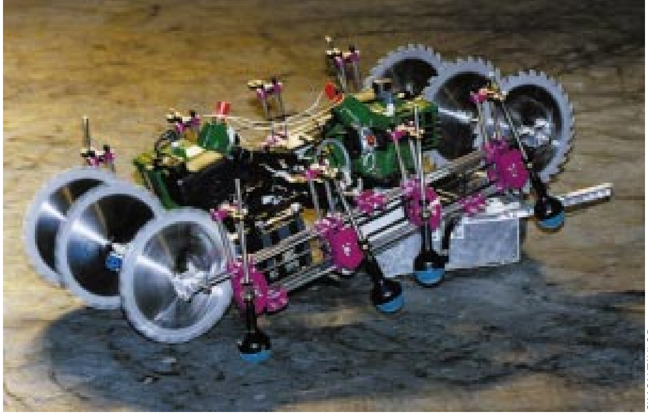
It's the aftermath of the lightweight-class melee on the final day of the Fourth Annual Robot Wars. The Herbst Pavilion at Fort Mason in San Francisco is overflowing with robot devotees and the proud parents of the destructive critters; the latter can be identified by the obsessive glint in their eyes as they crowd around tables in the "pit," where they minister to their magnificent fighting machines.

Robot Wars is a form of metallic cockfight: no guts, but plenty of glory. "It's a bloodless blood sport, and for that reason it's PC," says Marc Thorpe, creator of the event and self-declared opponent of political correctness. (He did, after all, win a controversial National Endowment for the Arts grant in 1974 to teach two dolphins to swim synchronously.) The 80 or so participating robots do have to adhere to a form of TC, or technological correctness. All of them—from heavyweights Vlad the Impaler and Mash-N-Go to middleweight Melga the Dental Hygienist and featherweights Fishstick from Guam and the Beast beneath Your Bed—cannot indulge in unsportsmanlike tactics. They cannot use powerful lasers, untethered projectiles, acids, explosives, flames, stun guns, heat guns, nets, ropes, irons, expandable foam, tape, water or glue.

"The 'no liquids' has to do with the fun quotient," explains Thorpe, formerly chief model maker at Industrial Light and Magic. "If liquids are permitted, the arena can become soupy" and interfere with the battles. Tape was banned after last year's wars, when SimCity creator Will Wright entered a clusterbot that fragmented into other robots that dispensed double-sided tape. "It just tied everybody up," Thorpe describes. "It was clever, but it makes for a very boring competition." The no-fibers policy emerged after a robot draped a net over an opponent's saw and immediately jammed it. "The nature and the spirit of the event is destruction and survival. It would undermine the whole event if there were no saw," he declares.

The most common limiting factor, however, seems all too human: "Overweight robots," Thorpe says, "are probably the biggest single problem discovered during the tech inspection."

—Marguerite Holloway in San Francisco



SIMON TAYLOR

DEATH ON WHEELS:
Razor Back and others conform to technological correctness.

APPLIED PHYSICS

BOOM

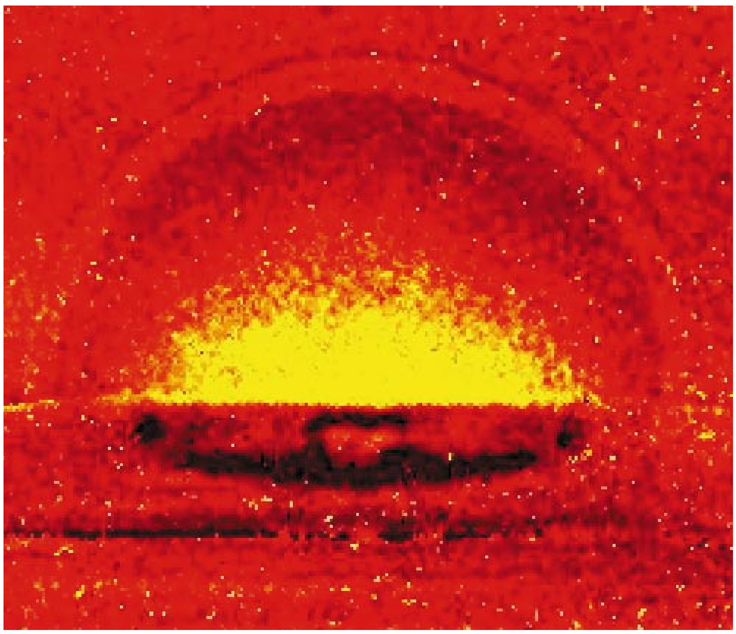
Ever wonder what the inside of a nuclear bomb looks like a microsecond after it detonates? Physicists at Los Alamos National Laboratory stay up nights thinking about such things, and a group of them recently demonstrated a clever new way to film the burn fronts that determine whether a warhead booms or fizzles. The technique may, ironically, one day reduce the damage that radiation treatment inflicts on some cancer patients.

The experiment did not require the researchers to obliterate a chunk of New Mexico. It actually takes two detonations for a nuclear weapon to execute its dreaded function. An initial blast of conventional high explosive is painstakingly tailored to implode a plutonium core into a critical mass. If it works, a chain reaction then takes over to produce a second, much bigger explosion. But thanks to the Comprehensive Test Ban Treaty, that would be illegal.

Rather than risk what would undoubtedly be a hefty fine, the Los Alamos team, led by John McClelland, substituted ordinary metal for plutonium. Then the researchers set off their half bomb inside a four-foot-diameter sphere made of steel. The idea, explains Christopher Morris, the project's chief scientist, is to make movies of the burning explosive, then to use those pictures to check the accuracy of supercomputer models.

Superman might be able to watch a shock front moving at more than 15,000 miles per hour behind two inches of steel, but for mere mortals, even x-rays aren't up to the job. "There is no technology for making an x-ray movie," Morris says, and even the fastest photographs suffer pronounced motion blur.

So the scientists hooked their blast chamber up to the lab's particle accelerator and made what Morris claims is the world's first movie recorded using matter rather than light (*above*). About 325 nanoseconds after detonation, the accelerator peppered the sphere with rapid-fire bursts of protons. A special camera on the other side translated the protons into an image showing the high explosive (*black-outlined block*) and the burning plasma (*yellow* and *dark red*) that it hurled outward.



C. M. RIEDEL

"This might even be exciting to people who don't care about the evil weapons stuff we do here," Morris speculates. "This technique should be able to deliver radiation more accurately to tumors with less damage to surrounding tissue," because the protons can be focused more tightly than x-rays. Preliminary tests of proton therapy for eye cancer have already begun, he says.

—W. Wayt Gibbs in San Francisco

We Don't Need No Regulation

Politicians will meddle as they have for generations. Now that the Internet is front-page news, what politician doesn't want to appear to be leading the leaders? The problem is, they don't know enough about technology to grasp which wave of public sentiment to get in front of.

An example is the debate over the regulation of encryption. This issue has created a wildly vacillating Congress, judiciary and executive within the U.S. (and consternation among governing bodies worldwide). First, the U.S. adopted a heavy-handed, controlling attitude on encryption. Now it apparently prefers a laissez-faire policy. But maybe not: a plethora of regulatory bills is pending before Congress. This erratic course points out the folly of sluggish governments attempting to keep up with Internet Time.

"The Internet should be a global free-trade zone," President Bill Clinton said in reversing his administration's stance on the export of encrypted computer products. That change led to "A Framework for Global Electronic Commerce" (www.iitf.nist.gov). The report aims to create a uniform code for electronic commerce, to delegate privacy regulation to industry and consumer groups, to let security standards and management be driven by market forces, to address Internet copyright protection issues, and to promise not to tax goods and services delivered by the Internet. Most dramatically, it takes a hands-off stance on content—no restrictions on pornography. The framework's primary author, Ira Magaziner, has been propelled into the limelight as a consequence of this enlightened policy.

So far so good, but the battle is not over. Spanning all nations, the Internet is the biggest machine in history. It is not clear that any single government can control it. Few politicians understand that. The Clinton administration may have shifted, but Congress still doesn't get it. This year no fewer than four bills regarding encryption either went or are scheduled to go before the legislature.

The most liberal proposal went down this past spring. Called the Promotion of Commerce Online in the Digital Era, or ProCODE Act, it was killed by the Senate Commerce Committee, which believed Clinton would have vetoed it. The ProCODE Act was exactly what the civil cyberians wanted—absolutely no export ban on encryption software.

A compromise of sorts is the Secure Public Networks Act, which passed the Senate Commerce Committee on June 19 (now it waits for a House vote and more committee meetings). It restricts export of strong encryption except when manufacturers require "key recovery." (Using more than 56 bits to encrypt a message is considered "strong," but in reality, 1,024 bits are needed to assure secrecy.) Think of an encoded message as a treasure chest with a lock that can be unlocked by only two keys: the one that the originator used to encode the message and the one that the receiver needs to decode it.


DAVID SUTER

This bill would force consumers to store their secret keys in a safe place—in a "key escrow account"—where the government can get the keys and unlock the messages. Of course, the government would need a court order to do that, but even so, the computer industry opposes the interference. Thus, the fight has centered on key recovery—what some have colorfully called the "back door."

In the end, Congress may have to yield to the freewheelers, especially in light of the shenanigans of Phil Zimmerman. He's the cyberhero who a few years ago wrote PGP (for "Pretty Good Privacy"), a very strong encryption software that was posted on the Internet. Now it is all over the world producing strong encryption—up to 2,048 bits—for free.

For a while, Zimmerman was accused of illegally exporting munitions. The feds eventually gave up on him: technically, Zimmerman had not violated the law, because a friend posted the software on the Internet, not him. With similar legal finesse, Zimmerman's company, PGP, Inc., worked out a deal with a non-U.S. company that also sidestepped the embargo on strong encryption.

The Clinton administration's change of heart stems in part from Zimmerman's and PGP's end runs around the rules. Whether such tactics have similarly influenced Congress should become clear soon. A proposal is in the works: the Safety and Freedom through Encryption Act, or SAFE Act. Barring last-minute amendments, this bill may be the best hope for individual freedom in cyberspace. It would lift controls on commercial and personal transactions alike. At press time, Congress was expected to vote on it this fall; it has 134 out of 218 votes needed to pass. This bill stands in stark contrast to the restrictive Encrypted Communications Privacy Act of 1997, which remains bottled up in committee and will probably die.

So it seems that SAFE is the leading candidate for passage, and the battle tilts toward noninterference and free-enterprisers such as Zimmerman. Already PGP, Inc., has secured Commerce Department permission to ship its 128-bit cryptography to a preapproved list of U.S. subsidiaries outside the country. Likewise, VeriFone got the go-ahead to ship overseas its software for secure on-line credit-card transactions.

If this trend continues, everyone will be able to export secure software. Not only will banks and credit-card companies enjoy security, but you and I will be able to send messages to friends and business associates without concern about invasion of privacy. Zimmerman's PGP has traveled from outlaw to pin-striped suit in Internet Time. Let's hope enlightened governments around the world keep up. —*Ted Lewis*

TED LEWIS is author of The Friction-Free Economy: Marketing Strategies for a Wired World, *published in October by HarperCollins.*

Mercury: The Forgotten Planet

Although one of Earth's nearest neighbors, this strange world remains, for the most part, unknown

by Robert M. Nelson

The planet closest to the sun, Mercury is a world of extremes. Of all the objects that condensed from the presolar nebula, it formed at the highest temperatures. The planet's dawn-to-dusk day, equal to 176 Earth-days, is the longest in the solar system, longer in fact than its own year. When Mercury is at perihelion (the point in its orbit closest to the sun), it moves so swiftly that, from the vantage of someone on the surface, the sun would appear to stop in the sky and go backward—until the planet's rotation catches up and makes the sun go forward again. During daytime, its ground temperature reaches 700 kelvins, the highest of any planetary surface (and more than enough to melt lead); at night, it plunges to a mere 100 kelvins (enough to freeze krypton).

Such oddities make Mercury exceptionally intriguing to astronomers. The planet, in fact, poses special challenges to scientific investigation. Its extreme properties make Mercury difficult to fit into any general scheme for the evolution of the solar system. In a sense, Mercury's unusual attributes provide an exacting and sensitive test for astronomers' theories. Yet even though Mercury ranks after Mars and Venus as one of Earth's nearest neighbors, distant Pluto is the only planet we know less about. Much about Mercury—its origins and evolution, its puzzling magnetic field, its tenuous atmosphere, its

DON DIXON

DAWN ON MERCURY, 10 times more brilliant than on Earth, is heralded by flares from the sun's corona snaking over the horizon. They light up the slopes of Discovery scarp (*cliffs at right*). In the sky, a blue planet and its moon are visible. (This artist's conception is based on data from the Mariner 10 mission.)

possibly liquid core and its remarkably high density—remains obscure.

Mercury shines brightly, but it is so far away that early astronomers could not discern any details of its terrain; they could map only its motion in the sky. As the innermost planet, Mercury (as seen from Earth) never wanders more than 27 degrees from the sun. This angle is less than that made by the hands on a watch at one o'clock. It can thus be observed only during the day, but scattered sunlight makes it difficult to see, or shortly before sunrise and after sunset, with the sun hanging just over the horizon. At dawn or dusk, however, Mercury is very low in the sky, and the light from it must pass through up to 10 times as much turbulent air as when it is directly overhead. The best Earth-based telescopes can see only those features on Mercury that are a few hundred kilometers across or wider—a resolution far worse than that for the moon seen with the unaided eye.

Despite these obstacles, terrestrial observation has yielded some interesting results. In 1955 astronomers were able to bounce radar waves off Mercury's surface. By measuring the so-called Doppler shift in the frequency of the reflections, they learned of Mercury's 59-day rotational period. Until then, Mercury had been thought to have an 88-day period, identical to its year, so that one side of the planet always faced the sun. The simple two-to-three ratio between the planet's day and year is striking. Mercury, which initially rotated much faster, probably dissipated energy through tidal flexing and slowed down, becoming locked into this ratio by an obscure process.

The new space-based observatories, such as the Hubble Space Telescope, are not limited by the problems of atmospheric distortion, and one might think them ideal tools for studying Mercury. Unfortunately, the Hubble, like many other sensors in space, cannot point at Mercury, because the rays of the nearby sun might accidentally damage sensitive optical instruments on board.

The only other way to investigate Mercury is to send a spacecraft to examine it up close. Only once has a probe made the trip: Mariner 10 flew by in the 1970s as part of a larger mission to explore the inner solar system. Getting the spacecraft there was not a trivial task. Falling directly into the gravitational potential well of the sun was impossible; the spacecraft had to ricochet around Venus to relinquish gravitational energy and thus slow down for a Mercury encounter. Mariner's orbit around the sun provided three close flybys of Mercury: on March 29, 1974; September 21, 1974; and March 16, 1975. The spacecraft returned images of about 40 percent of Mercury, showing a heavily cratered surface that, at first glance, appeared similar to that of the moon.

The pictures, sadly, led to the mistaken impression that Mercury differs very little from the moon and just happens to occupy a different region of the solar system. As a result, Mercury has become the neglected planet of the American space program. There have been more than 40 missions to the moon, 20 to Venus and more than 15 to Mars. By the end of the next decade, an armada of spacecraft will be in orbit about Venus, Mars, Jupiter and Saturn, returning detailed information about these planets and their environs for many years to come. But Mercury will remain largely unexplored.

The Iron Question

It was the Mariner mission that elevated scientific understanding of Mercury from almost nothing to most of what we currently know. The ensemble of instruments carried on that probe sent back about 2,000 images, with an effective resolution of about 1.5 kilometers, comparable to shots of the moon taken from Earth through a large telescope. Yet those many pictures captured only one face of Mercury; the other side has never been seen.

By measuring the acceleration of Mariner in Mercury's surprisingly strong gravitational field, astronomers confirmed one of the planet's most unusual characteristics: its high density. The other terrestrial (that is, nongaseous) bodies—Venus, the moon, Mars and Earth—exhibit a fairly linear relation between density and size. The largest, Earth and Venus, are quite dense, whereas the smaller ones, the moon and Mars, have lower densities.

Vital Statistics

Mercury is the innermost planet and has a highly inclined and eccentric orbit. It rotates about its own axis very slowly, so that one Mercury-day equals 176 Earth-days—longer than its year of 88 Earth-days. Proximity to the sun combined with elongated days gives Mercury the highest daytime temperatures in the solar system.

The planet has a rocky and cratered surface and is somewhat larger than the Earth's moon. It is exceptionally dense for its size, implying a large iron core. In addition, it has a strong magnetic field, which suggests that parts of the core are liquid. Because the small planet should have cooled fast enough to have entirely solidified, these findings raise questions about the planet's origins—and even about the birth of the solar system.

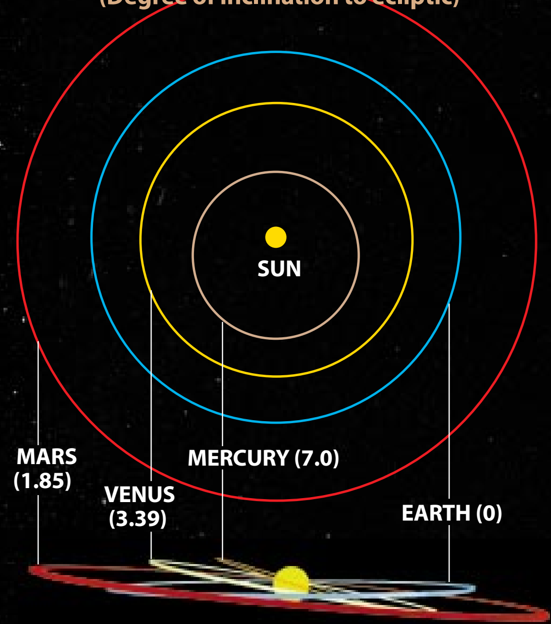
Mercury's magnetic field forms a magnetosphere around the planet, which partially shields the surface from the powerful wind of protons emanating from the sun. Its tenuous atmosphere consists of particles recycled from the solar wind or ejected from the surface.

Despite the planet's puzzling nature, only one spacecraft, Mariner 10, has ever flown by Mercury.

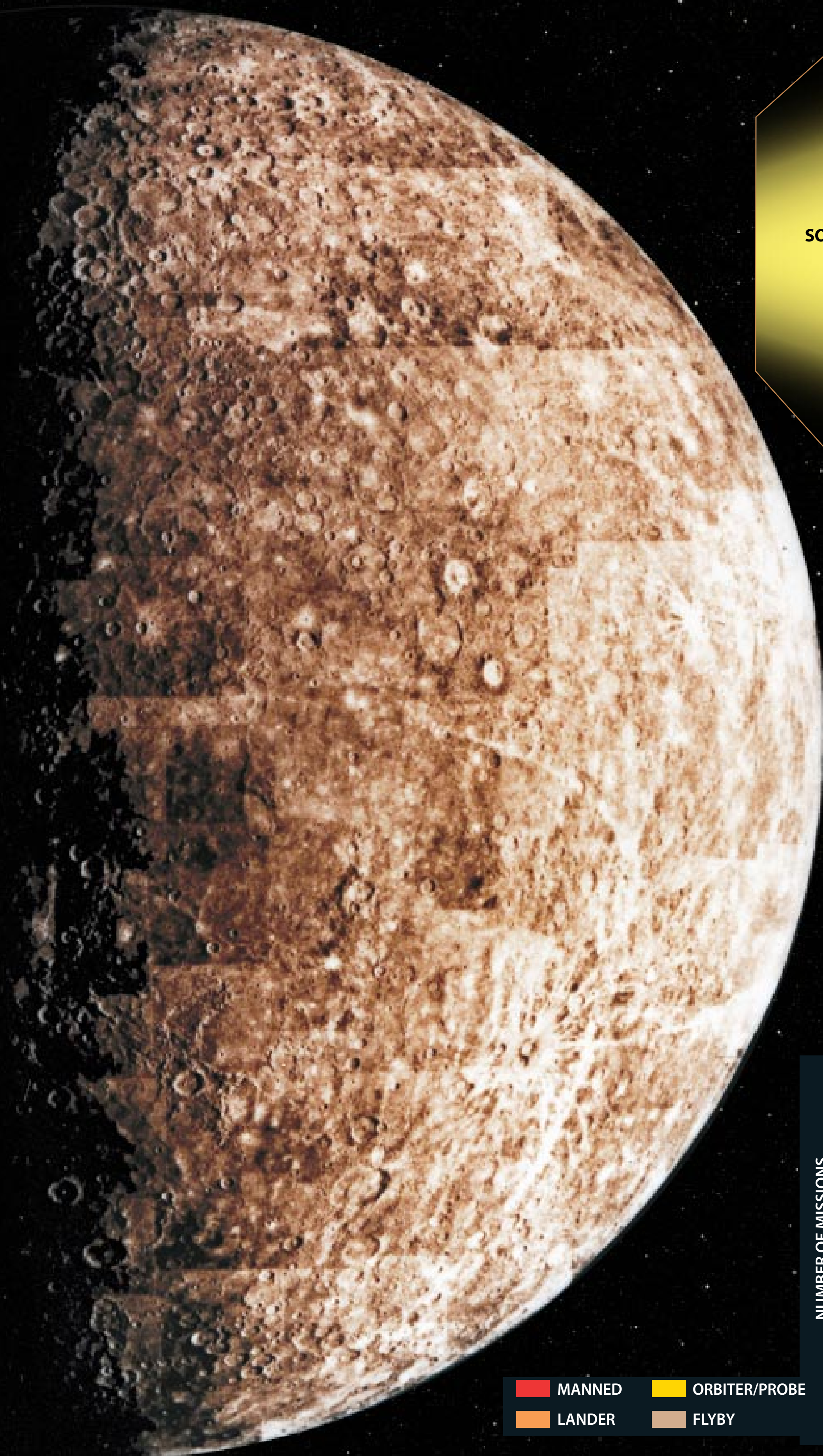
—R.M.N.

RELATIVE SIZES OF TERRESTRIAL BODIES

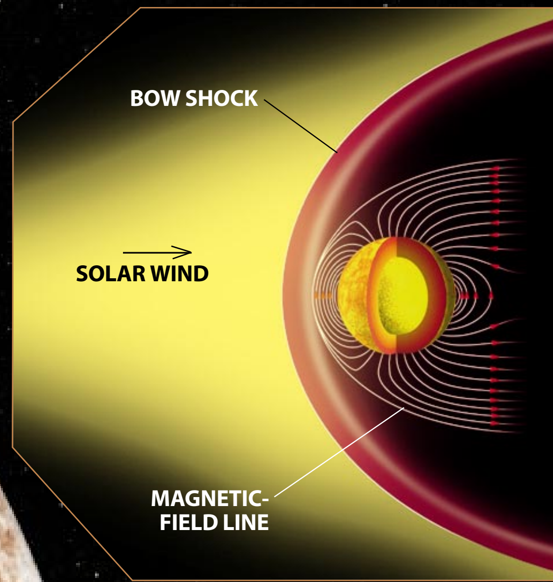
RELATIVE ORBITS OF TERRESTRIAL BODIES (Degree of inclination to ecliptic)



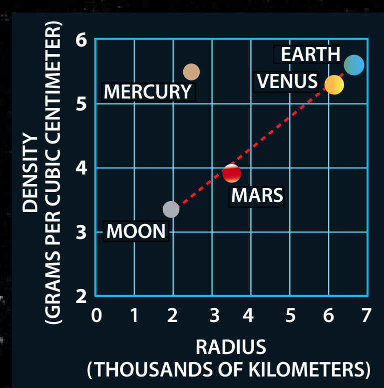
PHOTOGRAPHS BY NATIONAL AERONAUTICS AND SPACE ADMINISTRATION; ILLUSTRATIONS BY SLIM FILMS



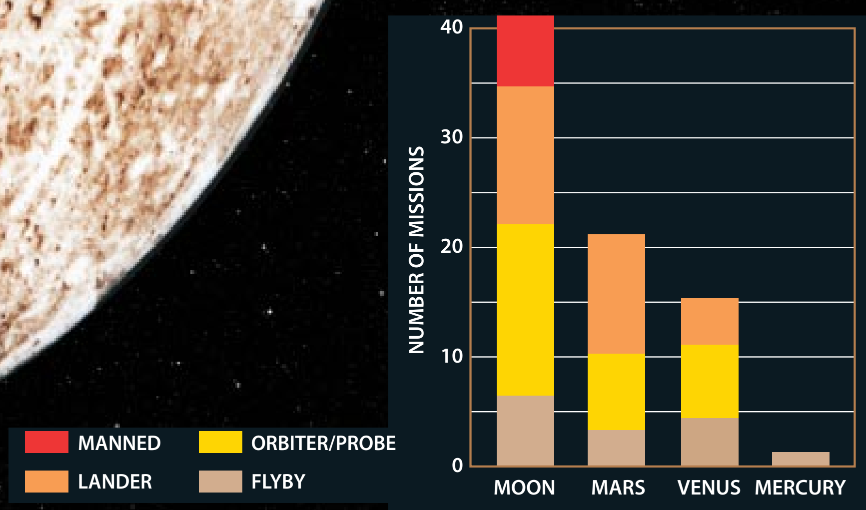
MERCURY'S MAGNETOSPHERE



DENSITY OF TERRESTRIAL BODIES



MISSIONS TO TERRESTRIAL BODIES



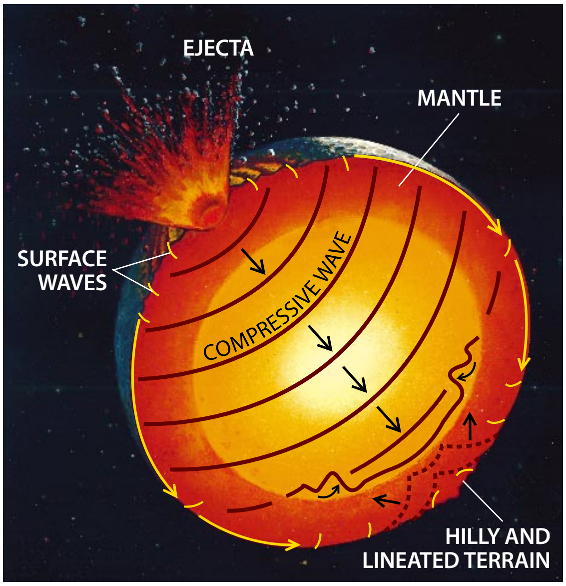
Mercury is not much bigger than the moon, but its density is typical of a far larger planet such as Earth.

This observation provides a fundamental clue about Mercury's interior. The outer layers of a terrestrial planet consist of lighter materials such as silicate rocks. With depth, the density increases, because of compression by the overlying rock layers and the different composition of the interior materials. The high-density cores of the terrestrial planets are probably made largely of iron.

Mercury may therefore have the largest metallic core, relative to its size, of all the terrestrial planets. This finding has stimulated a lively debate on the origin and evolution of the solar system. Astronomers assume that all the planets condensed from the solar nebula at about the same time. If this premise is true, then one of three possible circumstances may explain why Mercury is so special. First, the composition of the solar nebula might have been dramatically different in the vicinity of Mercury's orbit—much more so than theoretical models would predict. Or, second, the sun may have been so energetic early in the life of the solar system that the more volatile, low-density elements on Mercury were vaporized and driven off. Or, third, a very massive object may have collided with Mercury soon after its formation, vaporizing the less dense materials. The current body of evidence is not sufficient to discriminate among these possibilities.

Oddly enough, careful analysis of the Mariner findings, along with laborious spectroscopic observations from Earth, has failed to detect even trace amounts of iron in Mercury's crustal rocks. The apparent dearth of iron on the surface contrasts sharply with its presumed abundance in Mercury's interior. Iron occurs on Earth's crust and has been detected by spectroscopy on the rocks of the moon and Mars. So Mercury may be the only planet in the inner solar system with all its high-density iron concentrated in the interior and only low-density silicates in the crust. It may be that Mercury was molten for so long that the heavy substances settled at the center, just as iron drops below slag in a smelter.

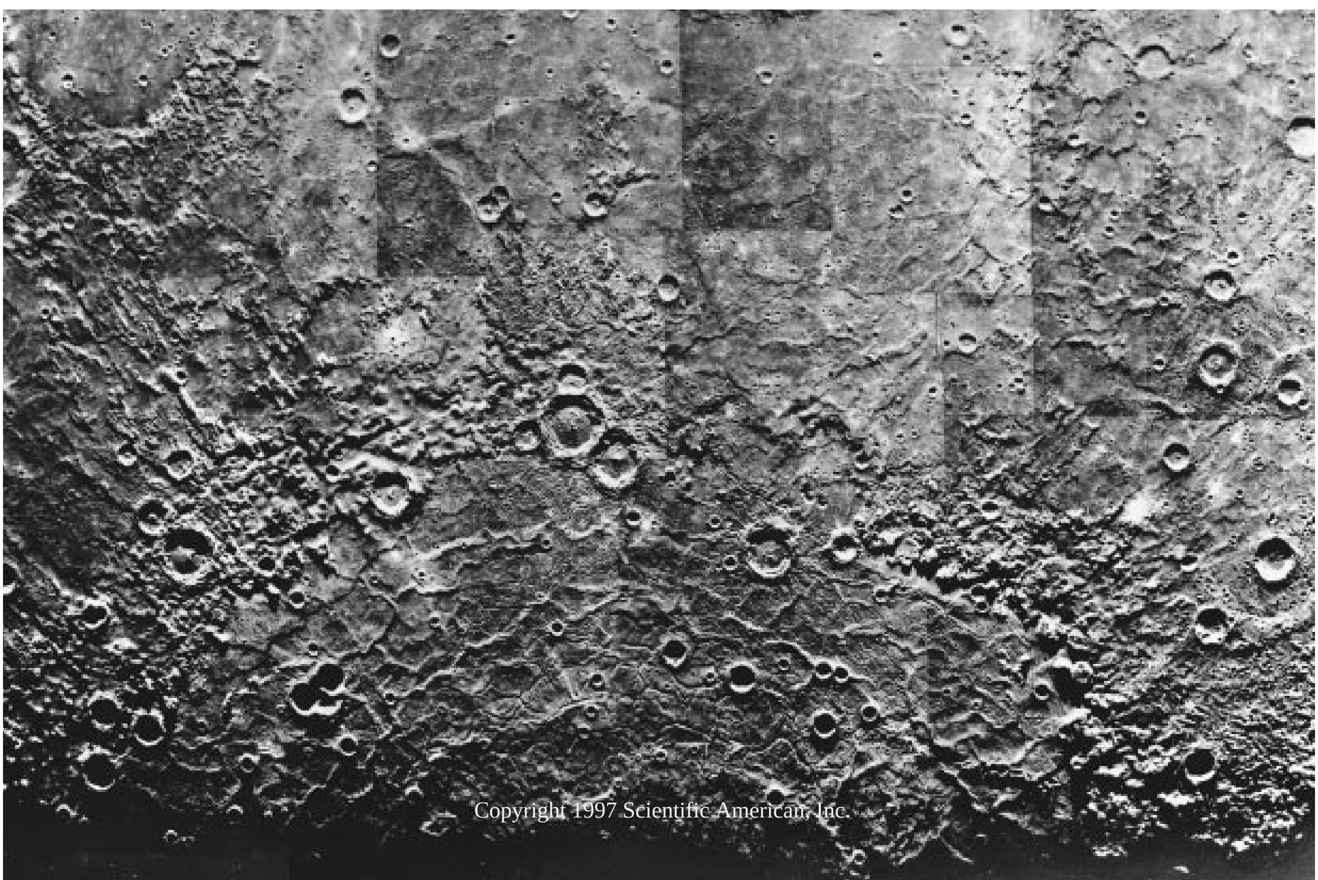
Mariner 10 also found that Mercury has a relatively strong magnetic field—the most powerful of all the terrestrial planets except Earth. The magnetic field of Earth is generated by electrically conductive molten metals circulating in the core, through a process called the self-sustaining dynamo. If Mercury's magnetic field has a similar source, then that planet must have a liquid interior.



ALFRED T. KAMAJIAN; COURTESY OF P. H. SCHULTZ AND D. E. GAULT

CALORIS CRATER
was formed when a giant projectile hit Mercury 3.6 billion years ago (*above*). Shock waves radiated through the planet, creating hilly and lineated terrain on the opposite side. The rim of Caloris itself (*below*) consists of concentric waves that froze in place after the impact. The flattened bed of the crater, 1,300 kilometers across, has since been covered with smaller craters.

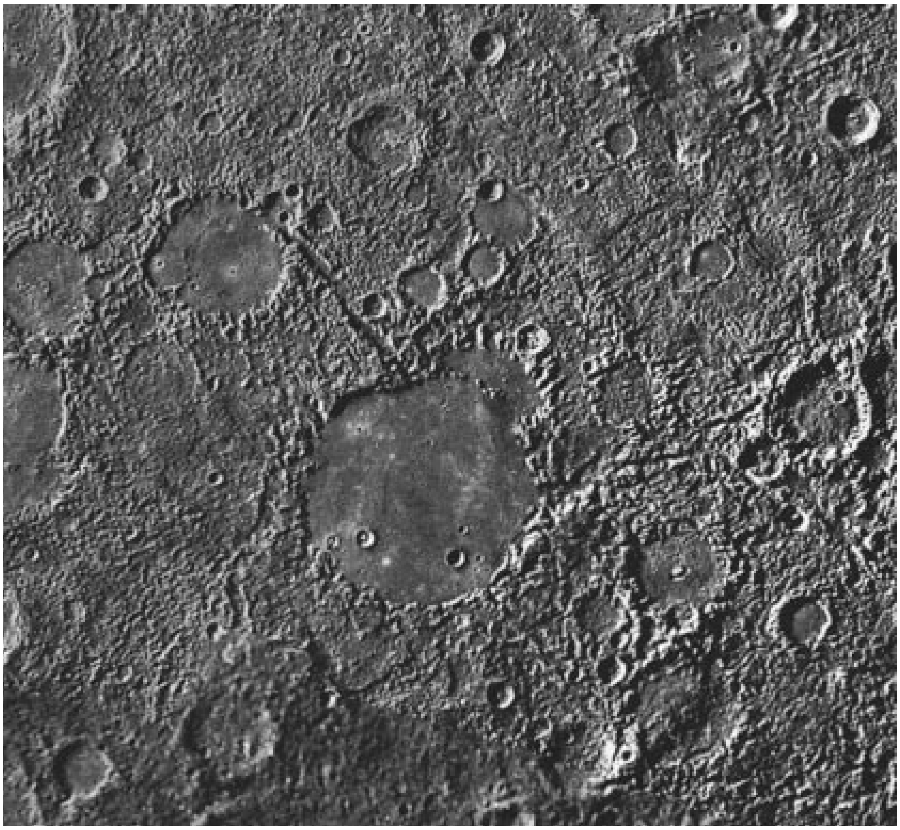
NASA



But there is a problem with this hypothesis. Small objects like Mercury have a high proportion of surface area compared with volume. Therefore, other factors being equal, smaller bodies radiate their energy to space faster. If Mercury has a purely iron core, as its large density and strong magnetic field imply, then the core should have cooled and solidified eons ago. But a solid core cannot support a self-sustaining magnetic dynamo.

This contradiction suggests that other materials are present in the core. These additives may depress the freezing point of iron, so that it remains liquid even at relatively low temperatures. Sulfur, a cosmically abundant element, is a possible candidate. Recent models, in fact, assume Mercury's core to be made of solid iron but surrounded by a liquid shell of iron and sulfur, at 1,300 kelvins. This solution to the paradox, however, remains a surmise.

Once a planetary surface solidifies sufficiently, it may bend when stress is applied steadily over long periods, or it may crack like a piece of glass on sudden impact. After Mercury was born four billion years ago, it was bombarded with huge meteorites that broke through its fragile outer skin and released torrents of lava. More recently, smaller collisions have caused lava to flow. These impacts must have either released enough energy to melt the surface or tapped deeper, liquid layers. Mercury's surface is stamped with events that occurred after its outer layer solidified.



ANTIPODE OF CALORIS
contains highly chaotic terrain, with hills and fractures that resulted from the impact on the other side of the planet. Petrarch crater (*at center*) was created by a far more recent impact, as evinced by the paucity of smaller craters on its smooth bed. But that collision was violent enough to melt rock, which flowed through a 100-kilometer-long channel and flooded a neighboring crater.

Planetary geologists have tried to sketch Mercury's history using these features—and without accurate knowledge of the rocks that constitute its surface. The only way to determine absolute age is by radiometric dating of returned samples (which so far are lacking). But geologists have ingenious ways of assigning relative ages, mostly based on the principle of superposition: any feature that overlies or cuts across another is the younger. This principle is particularly helpful in establishing the relative ages of craters.

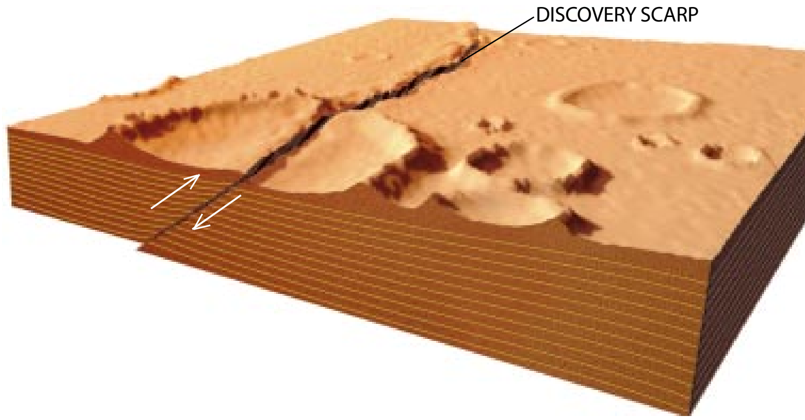
A Fractured History

Mercury has several large craters that are surrounded by multiple concentric rings of hills and valleys. The rings probably originated when a meteorite hit, causing shock waves to ripple outward like waves from a stone dropped into a pond, and then froze in place. Caloris, a behemoth 1,300 kilometers in diameter, is the largest of these craters. The impact that created it established a flat basin—wiping the slate clean, so to speak—on which a fresh record of smaller impacts has built up. Given an estimate of the rate at which projectiles hit the planet, the size distribution of these craters indicates that the Caloris impact probably occurred around 3.6 billion years ago; it serves as a reference point in time. The collision was so violent that it disrupted the surface on the opposite side of Mercury: the antipode of Caloris shows many cracks and faults.

Mercury's surface is also crosscut by linear features of unknown origin that are preferentially oriented north-south, northeast-southwest and northwest-southeast. These lineaments are called the Mercurian grid. One explanation for the checkered pattern is that the crust solidified when the planet was rotating much faster, perhaps with a day of only 20 hours. Because of its rapid spin, the planet would have had an equatorial bulge; after it slowed to its present period, gravity pulled it into a more spherical shape. The lineaments likely arose as the surface accommodated this change. The wrinkles do not cut across the Caloris crater, indicating that they were established before that impact occurred.

While Mercury's rotation was slowing, the planet was also cooling, so that the outer regions of the core solidified. The accompanying shrinkage probably reduced the planet's surface area by about a million square kilometers, producing a network of faults that are evident as a series of curved scarps, or cliffs, crisscrossing Mercury's surface.

Compared with Earth, where erosion has smoothed out most craters, Mercury, Mars and the moon have heavily cratered surfaces. The craters on these three planets also show a similar distribution of sizes, except that Mercury's craters tend to be somewhat larger. The objects striking Mercury most likely had higher velocity than those hitting the other planets. Such a pattern is to be expected if the projectiles were in elliptical orbits about the sun: they would have been moving faster in the region of Mercury's orbit than they were farther out. So these rocks may have been all from the same family, one that probably originated in the asteroid belt. In contrast, the moons of Jupiter have a different distribution of crater sizes, indicating that they collided with a different group of objects.



DISCOVERY SCARP (*crooked line seen in inset above and on opposite page*) stretches for 500 kilometers and in places is two kilometers high. It is a thrust fault, one of many riddling the surface of Mercury. These faults were probably created when parts of Mercury's core solidified and shrank. In consequence, the crust had to squeeze in to cover a smaller area. This compression is achieved when one section of crust slides over another—generating a thrust fault.

A Tenuous Atmosphere

Mercury's magnetic field is strong enough to trap charged particles, such as those blowing in with the solar wind (a stream of protons ejected from the sun). The magnetic field forms a shield, or magnetosphere, that is a miniaturized version of the one surrounding Earth. Magnetospheres change constantly in response to the sun's activity; Mercury's magnetic shield, because of its smaller size, can change much faster than Earth's. Thus, it responds quickly to the solar wind, which is 10 times denser at Mercury than at Earth.

The fierce solar wind steadily bombards Mercury on its illuminated side. The magnetic field is just strong enough to prevent the wind from reaching the planet's surface, except when the sun is very active or when Mercury is at perihelion. At these times, the solar wind reaches all the way down to the surface, and its energetic protons knock material off the crust. The particles thus ejected can then get trapped by the magnetosphere.

Objects as hot as Mercury do not, however, retain appreciable atmospheres around them, because gas molecules tend to move faster than the escape velocity of the planet. Any significant amount of volatile material on Mercury should soon be lost to space. For this reason, it had long been thought that Mercury did not have an atmosphere. But the ultraviolet spectrometer on Mariner 10 detected small amounts of hydrogen, helium and oxygen, and subsequent Earth-based observations have found traces of sodium and potassium.

The source and ultimate fate of this atmospheric material is a subject of animated argument. Unlike Earth's gaseous cloak, Mercury's atmosphere is constantly evaporating and being replenished. Much of the atmosphere is probably created, directly or indirectly, by the solar wind. Some components of the thin atmosphere may come from the magnetosphere or from the direct infall of cometary material. And once an atom is "sputtered" off the surface by the solar wind, it also adds to the tenuous atmosphere. It is even possible that the planet is still outgassing the last remnants of its primordial inventory of volatile substances.

Recently a team of astronomers from the California Institute of Technology and the Jet Propulsion Laboratory (JPL), both in Pasadena, Calif., observed the circular polarization of a radar beam reflected from near Mercury's poles. Those results suggest the presence of water ice. The prospect of a planet as hot as Mercury having ice caps—or any water at all—is intriguing. It may be that the ice resides in permanently shaded regions near Mercury's poles and is left over from primordial water that condensed on the planet when it formed.

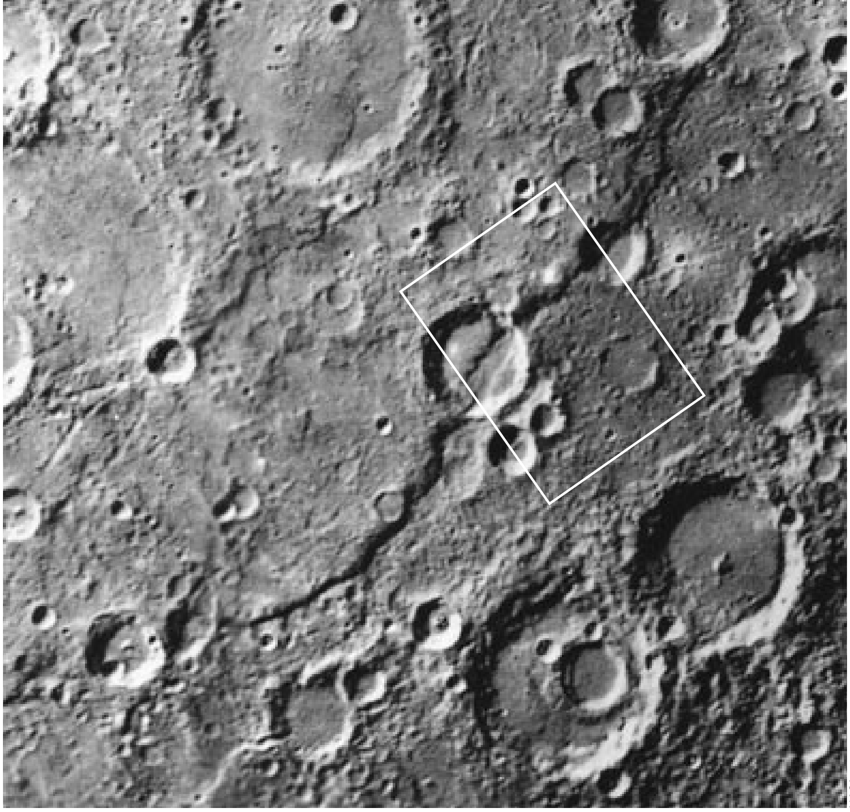
If so, Mercury must have stayed in a remarkably stable orientation for the entire age of the solar system, never tipping either pole to the sun—despite devastating events such as the Caloris impact. Such stability would be highly remarkable. Another possible source of water might be the comets that are continually falling into Mercury. Ice landing at a pole may remain in the shade, evaporating very slowly; such water deposits may be a source of Mercury's atmospheric oxygen and hydrogen. On the other hand, astronomers at the University of Arizona have suggested that the shaded polar regions may contain other volatile species such as sulfur, which mimics the radar reflectivity of ice but has a higher melting point.

Obstacles to Exploration

Why has Mercury been left out of the efforts to explore the solar system for nearly a quarter century? One possibility, as mentioned, is the superficial similarity between Mercury and the moon. Another, more subtle factor arises from the way planetary missions are devised. The members of peer-review panels for the National Aeronautics and Space Administration have generally been involved in NASA's most recent missions. The preponderance of missions has been to other planets, so that these planetary scientists have developed a highly specialized body of expertise and interests. In contrast to the planets thus favored, Mercury has a small advocacy group.

Another consideration is economics. The top levels of NASA are demanding that scientists propose missions that are "faster, better, cheaper," that focus on a limited set of objectives and that trade the science value against the total cost. In the present constrained budgetary environment, the largest deep-space exploration proposals that NASA is able to consider from individuals are those to its Discovery program. Interested scientists team up with industry to propose missions, some of which are selected and funded by NASA for further study. (Four of these missions have so far been undertaken.) The Discovery proposals are supposed to constrain the cost of a mission to $226 million or less. By comparison, NASA's Galileo mission to Jupiter and its Cassini mission to Saturn will both cost more than $1 billion.

A mission to orbit Mercury poses a special technical hurdle. The spacecraft must be protected against the intense en-



NASA

ergy radiating from the sun and even against the solar energy reflected off Mercury. Because the spacecraft will be close to the planet, at times "Mercury-light" can become a greater threat than the direct sun itself. Despite all the challenges, NASA received one Discovery mission proposal for a Mercury orbiter in 1994 and two in 1996.

The 1994 proposal, called Hermes '94, employed a traditional hydrazine–nitrogen tetroxide propulsion system, requiring as much as 1,145 kilograms of propellants. Much of this fuel is needed to slow the spacecraft as it falls toward the sun. The mission's planners, who include myself, could have reduced the fuel mass only by increasing the number of planetary encounters (to remove gravitational energy). Unfortunately, these maneuvers would have increased the time spent in space, where exposure to radiation limits the lifetime of critical solid-state components.

The instrument complement would have permitted Mercury's entire surface to be mapped at a resolution of one kilometer or better. These topographic maps could be correlated with charts of Mercury's magnetic and gravitational fields. NASA initially selected the mission as a candidate for study but ultimately rejected it because of the high cost and risk.

In 1996 the Hermes team, JPL and Spectrum Astro Corporation in Gilbert, Ariz., proposed a new technology that permitted the same payload while slashing the fuel mass, cost and time spent in interplanetary cruise. Their design called for a solar-powered ion thruster engine, requiring only 295 kilograms of fuel. This revolutionary engine would propel the spacecraft by using the sun's energy to ionize atoms of xenon and accelerate them to high velocity using an electrical field directed out of the rear of the spacecraft. This innovation would have made the interplanetary cruise time of Hermes '96 a year shorter than that for Hermes '94. Yet NASA did not consider Hermes '96 for further study, because it regarded solar electric propulsion without full backup from chemical propellant to be too experimental.

NASA has, however, selected one proposal for a Mercury orbiter for intensive consideration in the 1996 cycle of Discovery missions. This design, called Messenger, was developed by engineers at the Applied Physics Laboratory in Maryland. Like Hermes '94, it would rely on traditional chemical propulsion and carry similar sensors. Moreover, it would have two devices that could determine the proportions of the most abundant elements of the crustal rocks. Although these two instruments are scientifically attractive, their additional mass requires that the spacecraft swoop by Venus twice and Mercury three times before it goes into orbit. This trajectory will lengthen the journey to Mercury to more than four years (about twice that of Hermes '96). Messenger is also the most costly Discovery mission under consideration, with a current price tag of $211 million.

Officials awarding contracts for Discovery missions emphasize that they rely strongly on advice from reviewers outside NASA. When making decisions, these panels strive for consensus, a process that causes them to favor proved technologies and remain unreceptive to new ones. Fortunately, NASA has instituted a separate program that embraces futuristic ideas. The mission now planned under this program, called New Millennium Deep Space One, is designed to demonstrate in space all the groundbreaking technologies that have been previously proposed. In July 1998 Deep Space One, powered by a solar ion drive, will begin a three-year journey to fly by asteroid McAuliffe (named after *Challenger* astronaut Christa McAuliffe), the planet Mars and Comet West-Kohoutek-Ikamura. Deep Space One may prove that solar electric propulsion works as well as its supporters now expect. If so, then during the first part of the next century, solar engines should power many flights around the inner solar system—and will surely help solve the long-neglected mysteries of Mercury. SA

The Author

ROBERT M. NELSON has been a research scientist at the Jet Propulsion Laboratory in Pasadena, Calif., since 1979. He received his Ph.D. in planetary astronomy from the University of Pittsburgh in 1977. Nelson was co-investigator for the Voyager spacecraft's photopolarimeter and is on the science team for the Visual and Infrared Mapping Spectrometer of the Cassini Saturn Orbiter mission. He was also the principal investigator on the Hermes '94 and '96 proposals for a Mercury orbiter and is the flight scientist for the experimental New Millennium Deep Space One mission, to be launched in 1998. The author expresses his gratitude to the Hermes team members for their enlightening contributions.

Further Reading

Atlas of Mercury. Edited by M. E. Davies, D. E. Gault, S. E. Dwornik and R. G. Strom. NASA Scientific and Technical Information Office, Washington, D.C., 1978.

Mercury. Edited by F. Vilas, C. R. Chapman and M.S. Matthews. University of Arizona Press, 1988.

The New Solar System. Edited by J. K. Beatty and A. Chaikin. Cambridge University Press and Sky Publishing Corporation, 1990.

Fermat's Last Stand

His most notorious theorem baffled the greatest minds for more than three centuries. But after 10 years of work, one mathematician cracked it

by Simon Singh and Kenneth A. Ribet

PIERRE DE FERMAT, a 17th-century master of number theory, often wrote to other mathematicians, asking if they had the ingenuity to match his solutions. He devised his most famous challenge, his so-called last theorem, while studying *Arithmetica,* by Diophantus of Alexandria. Fermat asserted that there are no nontrivial solutions for the equation $a^n + b^n = c^n$, where n represents any whole number greater than 2. In the margin of *Arithmetica,* Fermat jotted a comment that tormented three centuries of mathematicians: "I have a truly marvelous demonstration of this proposition, which this margin is too narrow to contain."

GRANGER COLLECTION; BROWN UNIVERSITY LIBRARY; DIGITAL COMPOSITION BY SLIM FILMS

This past June, 500 mathematicians gathered in the Great Hall of Göttingen University in Germany to watch Andrew J. Wiles of Princeton University collect the prestigious Wolfskehl Prize. The reward—established in 1908 for whoever proved Pierre de Fermat's famed last theorem—was originally worth $2 million (in today's dollars). By the summer of 1997, hyperinflation and the devaluation of the mark had reduced it to a mere $50,000. But no one cared. For Wiles, proving Fermat's 17th-century conundrum had realized a childhood dream and ended a decade of intense effort. For the assembled guests, Wiles's proof promised to revolutionize the future of mathematics.

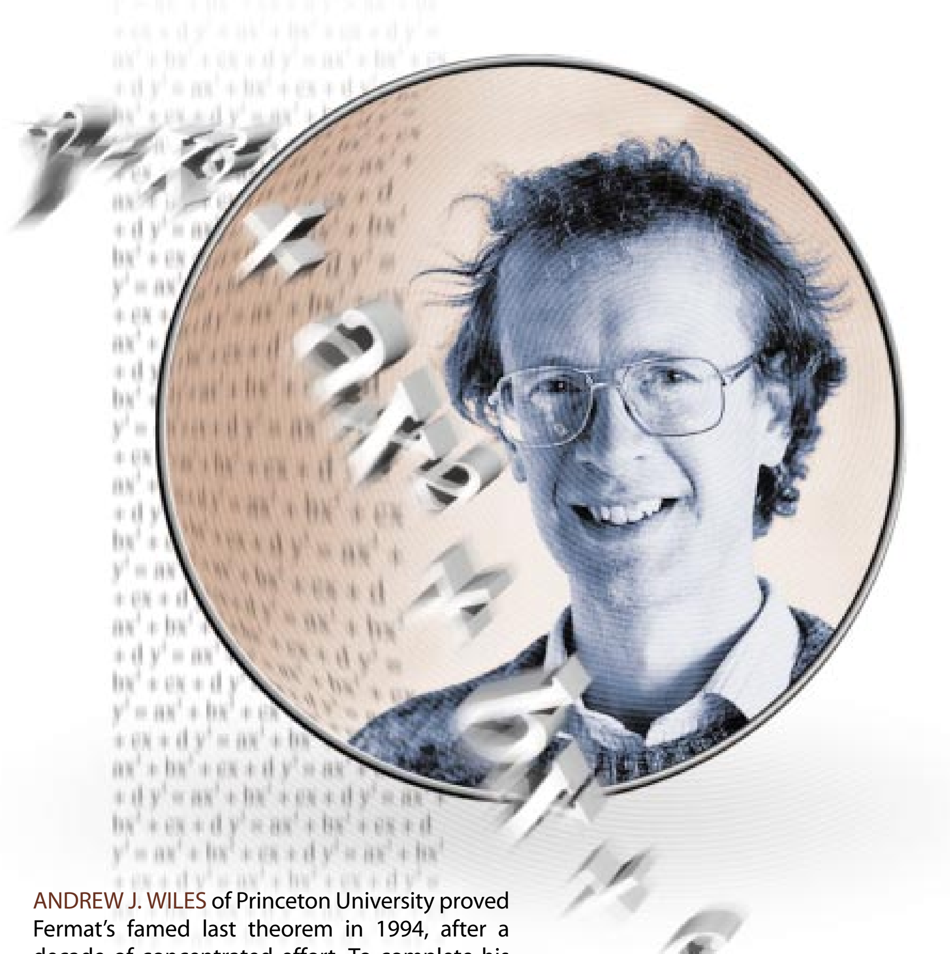
Indeed, to complete his 100-page calculation, Wiles needed to draw on and further develop many modern ideas in mathematics. In particular, he had to tackle the Shimura-Taniyama conjecture, an important 20th-century insight into both algebraic geometry and complex analysis. In doing so, Wiles forged a link between these major branches of mathematics. Henceforth, insights from either field are certain to inspire new results in the other. Moreover, now that this bridge has been built, other connections between distant mathematical realms may emerge.

The Prince of Amateurs

Pierre de Fermat was born on August 20, 1601, in Beaumont-de-Lomagne, a small town in southwest France. He pursued a career in local government and the judiciary. To ensure impartiality, judges were discouraged from socializing, and so each evening Fermat would retreat to his study and concentrate on his hobby, mathematics. Although an amateur, Fermat was highly accomplished and was largely responsible for probability theory and the foundations of calculus. Isaac Newton, the father of modern calculus, stated that he had based his work on "Monsieur Fermat's method of drawing tangents."

Above all, Fermat was a master of number theory—the study of whole numbers and their relationships. He would often write to other mathematicians about his work on a particular problem and ask if they had the ingenuity to match his solution. These challenges, and the fact that he would never reveal his own calculations, caused others a great deal of frustration. René Descartes, perhaps most noted for inventing coordinate geometry, called Fermat a braggart, and the English mathematician John Wallis once referred to him as "that damned Frenchman."

Fermat penned his most famous challenge, his so-called last theorem, while studying the ancient Greek mathematical text *Arithmetica,* by Diophantus of Alexandria. The book discussed positive whole-number solutions to the equation $a^2 + b^2 = c^2$, Pythagoras's formula de-



GAMMA LIAISON; SLIM FILMS

ANDREW J. WILES of Princeton University proved Fermat's famed last theorem in 1994, after a decade of concentrated effort. To complete his 100-page calculation, Wiles needed to draw on and further develop many modern ideas in mathematics. In particular, he had to prove the Shimura-Taniyama conjecture for a subset of elliptic curves, objects described by cubic equations such as $y^2 = x^3 + ax^2 + bx + c$.

scribing the relation between the sides of a right triangle. This equation has infinitely many sets of integer solutions, such as $a = 3$, $b = 4$, $c = 5$, which are known as Pythagorean triples. Fermat took the formula one step further and concluded that there are no nontrivial solutions for a whole family of similar equations, $a^n + b^n = c^n$, where n represents any whole number greater than 2.

It seems remarkable that although there are infinitely many Pythagorean triples, there are no Fermat triples. Even so, Fermat believed he could support his claim with a rigorous proof. In the margin of *Arithmetica,* the mischievous genius jotted a comment that taunted generations of mathematicians: "I have a truly marvelous demonstration of this proposition, which this margin is too narrow to contain." Fermat made many such infuriating notes, and after his death his son published an edition of *Arithmetica* that included these teases. All the theorems were proved, one by one, until only Fermat's last remained.

Numerous mathematicians battled the last theorem and failed. In 1742 Leonhard Euler, the greatest number theorist of the 18th century, became so frustrated by his inability to prove the last theorem that he asked a friend to search Fermat's house in case some vital scrap of paper was left behind. In the 19th century Sophie Germain—who, because of prejudice against women mathematicians, pursued her studies under the name of Monsieur Leblanc—made the first significant breakthrough. Germain proved a general theorem that went a long way toward solving Fermat's equation for values of n that are prime numbers greater than 2 and for which $2n + 1$ is also prime. (Recall that a prime number is divisible only by 1 and itself.) But a complete proof for these exponents, or any others, remained out of her reach.

At the start of the 20th century Paul Wolfskehl, a German industrialist, bequeathed 100,000 marks to whoever could meet Fermat's challenge. According to some historians, Wolfskehl was at one time almost at the point of suicide, but he became so obsessed with trying to prove the last theorem that his death wish disappeared. In light of what had happened, Wolfskehl rewrote his will. The prize was his way of repaying a debt to the puzzle that saved his life.

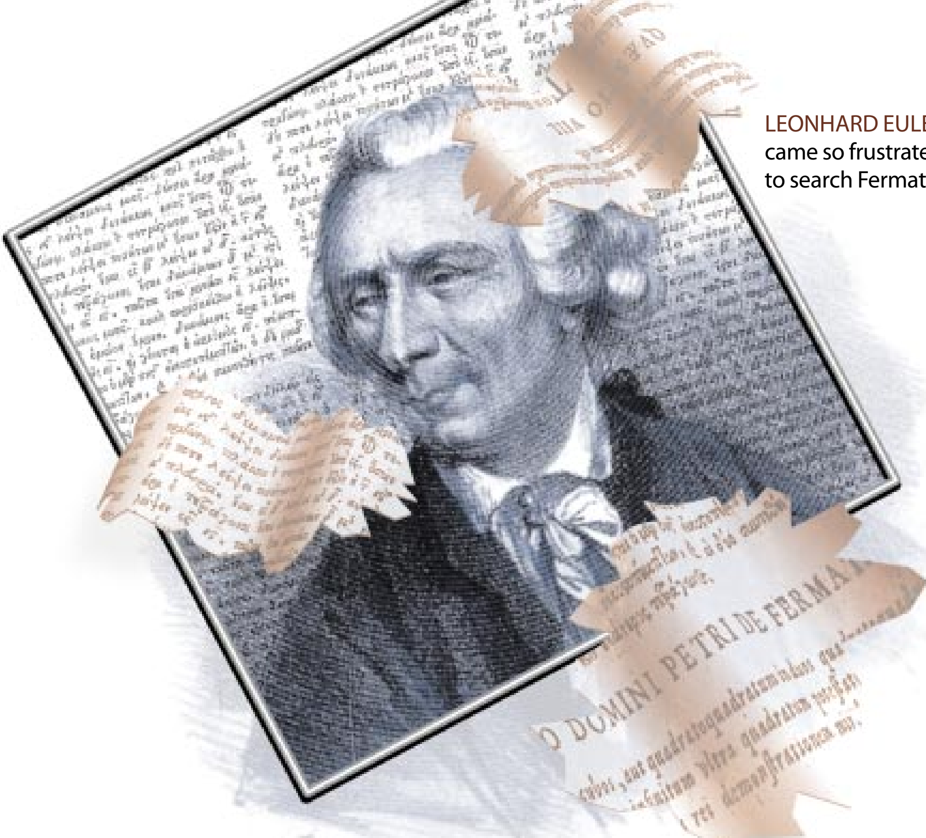
Ironically, just as the Wolfskehl Prize was encouraging enthusiastic amateurs to attempt a proof, professional mathematicians were losing hope. When the great German logician David Hilbert was asked why he never attempted a proof of Fermat's last theorem, he replied, "Before beginning I should have to put in three years of intensive study, and I haven't that much time to squander on a probable failure." The problem still held a special place in the hearts of number theorists, but they regarded Fermat's last theorem in the same way that chemists regarded alchemy. It was a foolish romantic dream from a past age.

The Childhood Dream

Children, of course, love romantic dreams. And in 1963, at age 10, Wiles became enamored with Fermat's last theorem. He read about it in his local library in Cambridge, England, and promised himself that he would find a proof. His schoolteachers discouraged him from wasting time on the impossible. His college lecturers also tried to dissuade him. Eventually his graduate supervisor at the University of Cambridge steered him toward more mainstream mathematics, namely into the fruitful research area surrounding objects called elliptic curves. The ancient Greeks originally studied elliptic curves, and they appear in *Arithmetica.* Little did Wiles know that this training would lead him back to Fermat's last theorem.

Elliptic curves are not ellipses. Instead they are named as such because they are described by cubic equations, like those used for calculating the perimeter of an ellipse. In general, cubic equations for elliptical curves take the form $y^2 = x^3 + ax^2 + bx + c$, where a, b and c are whole numbers that satisfy some simple conditions. Such equations are said to be of degree 3, because the highest exponent they contain is a cube.

Number theorists regularly try to ascertain the number of so-called rational solutions, those that are whole numbers or fractions, for various equations. Linear or quadratic equations, of degree 1 and 2, respectively, have either no rational solutions or infinitely many, and it is simple to decide which is the case. For complicated equations, typically of degree 4 or higher, the number of solutions is always finite—a fact called Mordell's conjecture, which the German



LEONHARD EULER, the greatest number theorist of the 18th century, became so frustrated by Fermat's last theorem that in 1742 he asked a friend to search Fermat's house for any scrap of paper left behind.

GRANGER COLLECTION; BROWN UNIVERSITY LIBRARY; SLIM FILMS

mathematician Gerd Faltings proved in 1983. But elliptic curves present a unique challenge. They may have a finite or infinite number of solutions, and there is no easy way of telling.

To simplify problems concerning elliptic curves, mathematicians often reexamine them using modular arithmetic. They divide x and y in the cubic equation by a prime number p and keep only the remainder. This modified version of the equation is its "mod p" equivalent. Next, they repeat these divisions with another prime number, then another, and as they go, they note the number of solutions for each prime modulus. Eventually these calculations generate a series of simpler problems that are analogous to the original.

The great advantage of modular arithmetic is that the maximum values of x and y are effectively limited to p, and so the problem is reduced to something finite. To grasp some understanding of the original infinite problem, mathematicians observe how the number of solutions changes as p varies. And using that information, they generate a so-called L-series for the elliptic curve. In essence, an L-series is an infinite series in powers, where the value of the coefficient for each pth power is determined by the number of solutions in modulo p.

In fact, other mathematical objects, called modular forms, also have L-series. Modular forms should not be confused with modular arithmetic. They are a certain kind of function that deals with complex numbers of the form $(x + iy)$, where x and y are real numbers, and i is the imaginary number (equal to the square root of -1).

What makes modular forms special is that one can transform a complex number in many ways, and yet the function yields virtually the same result. In this respect, modular forms are quite remarkable. Trigonometric functions are similar inasmuch as an angle, q, can be transformed by adding π, and yet the answer is constant: $\sin q = \sin (q + \pi)$. This property is termed symmetry, and trigonometric functions display it to a limited extent. In contrast, modular forms exhibit an immense level of symmetry. So much so that when the French polymath Henri Poincaré discovered the first modular forms in the late 19th century, he struggled to come to terms with their symmetry. He described to his colleagues how every day for two weeks he would wake up and search for an error in his calculations. On the 15th day he finally gave up, accepting that modular forms are symmetrical in the extreme.

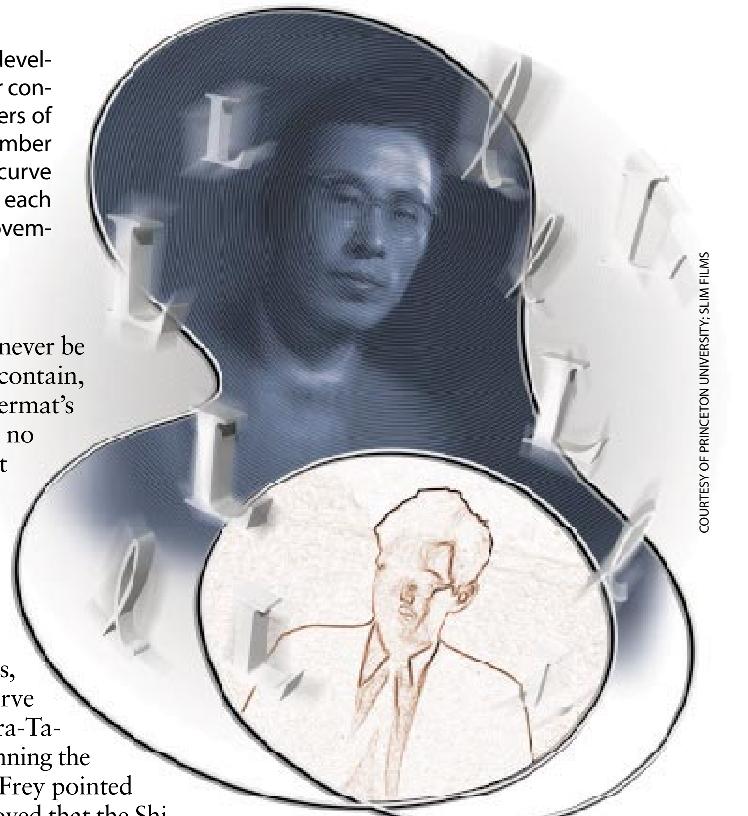
A decade or so before Wiles learned about Fermat, two young Japanese mathematicians, Goro Shimura and Yutaka Taniyama, developed an idea involving modular forms that would ultimately serve as a cornerstone in Wiles's proof. They believed that modular forms and elliptic curves were fundamentally related—even though elliptic curves apparently belonged to a totally different area of mathematics. In particular, because modular forms have an L-series—although derived by a different prescription than that for elliptic curves—the two men proposed that every elliptic curve could be paired with a modular form, such that the two L-series would match.

Shimura and Taniyama knew that if they were right, the consequences would be extraordinary. First, mathematicians generally know more about the L-series of a modular form than that of an elliptic curve. Hence, it would be unnecessary to compile the L-series for an elliptic curve, because it would be identical to that of the corresponding modular form. More generally, building such a bridge between two hitherto unrelated branches of mathematics could benefit both: potentially each discipline could become enriched by knowledge already gathered in the other.

The Shimura-Taniyama conjecture, as it was formulated by Shimura in the early 1960s, states that every elliptic curve can be paired with a modular form; in other words, all elliptic curves are modular. Even though no one could find a way to prove it, as the decades passed the hypothesis became increasingly influential. By the 1970s, for instance, mathematicians would often assume that the Shimura-Taniyama conjecture was true and then derive some new result from it. In due course, many major findings came to rely on the conjecture, although few scholars expected it would be proved in this century. Tragically, one of the men who inspired it did not live to see its ultimate importance. On November 17, 1958, Yutaka Taniyama committed suicide.

The Missing Link

In the fall of 1984, at a symposium in Oberwolfach, Germany, Gerhard Frey of the University of Saarland gave a lecture that hinted at a new strategy for attacking Fermat's last theorem. The theorem asserts that Fermat's equation has no positive whole-number solutions. To test a statement of this type, mathematicians frequently assume that it is false and then explore the consequences. To



GORO SHIMURA AND YUTAKA TANIYAMA (*top* and *bottom,* respectively) developed an idea during the 1950s that ultimately served in Wiles's proof. Their conjecture involved modular forms—functions that deal with complex numbers of the form ($x + iy$), where x and y are real numbers, and i is the imaginary number (equal to the square root of -1). The two men proposed that every elliptic curve could be paired with a modular form, such that the L-series associated with each would match. Tragically, Taniyama did not live to see Wiles's success. On November 17, 1958, he killed himself.

COURTESY OF PRINCETON UNIVERSITY; SLIM FILMS

say that Fermat's last theorem is false is to say that there are two perfect nth powers whose sum is a third nth power.

Frey's idea proceeded as follows: Suppose that A and B are perfect nth powers of two numbers such that $A + B$ is again an nth power—that is, they are a solution to Fermat's equation. A and B can then be used as coefficients in a special elliptic curve: $y^2 = x(x - A)(x + B)$. A quantity that is routinely calculated whenever one studies elliptic curves is the "discriminant" of the elliptic curve, $A^2B^2(A + B)^2$. Because A and B are solutions to the Fermat equation, the discriminant is a perfect nth power.

The crucial point in Frey's tactic is that if Fermat's last theorem is false, then whole-number solutions such as A and B can be used to construct an elliptic curve whose discriminant is a perfect nth power. So a proof that the discriminant of an elliptic curve can never be an nth power would contain, implicitly, a proof of Fermat's last theorem. Frey saw no way to construct that proof. He did, however, suspect that an elliptic curve whose discriminant was a perfect nth power—if it existed—could not be modular. In other words, such an elliptic curve would defy the Shimura-Taniyama conjecture. Running the argument backwards, Frey pointed out that if someone proved that the Shimura-Taniyama conjecture is true and that the elliptic equation $y^2 = x(x - A)(x + B)$ is not modular, then they would have shown that the elliptic equation cannot exist. In that case, the solution to Fermat's equation cannot exist, and Fermat's last theorem is proved true.

Many mathematicians explored this link between Fermat and Shimura-Taniyama. Their first goal was to show that the Frey elliptic curve, $y^2 = x(x - A)(x + B)$, was in fact not modular. Jean-Pierre Serre of the College of France and Barry Mazur of Harvard University made important contributions in this direction. And in June 1986 one of us (Ribet) at last constructed a complete proof of the assertion. It is not possible to describe the full argument in this article, but we will give a few hints.

To begin, Ribet's proof depends on a geometric method for "adding" two points on an elliptic curve [*see bottom illustration on next page*]. Visually, the idea is that if you project a line through a pair of distinct solutions, P_1 and P_2, the line cuts the curve at a third point, which we might provisionally call the sum of P_1 and P_2. A slightly more complicated but more valuable version of this addition is as follows: first add two points and derive a new point, P_3, as already described, and then reflect this point through the x axis to get the final sum, Q.

This special form of addition can be applied to any pair of points within the infinite set of all points on an elliptic curve, but this operation is particularly interesting because there are finite sets of points having the crucial property that the sum of any two points in the set is again in the set. These finite sets of points form a group: a set of points that obeys a handful of simple axioms. It turns out that if the elliptic curve is modular, so are the points in each finite group of the elliptic curve. What Ribet proved is that a specific finite group of Frey's curve cannot be modular, ruling out the modularity of the whole curve.

For three and half centuries, the last theorem had been an isolated problem, a curious and impossible riddle on the edge of mathematics. In 1986 Ribet, building on Frey's work, had brought it

STEVE MUREZ *Black Star*; SLIM FILMS

SOPHIE GERMAIN pursued her studies under the name of Monsieur Leblanc because of prejudice against women mathematicians. She made the first significant breakthrough in the 19th century, proving a theorem that went a long way toward solving Fermat's equation for values of n that are prime numbers greater than 2 and for which $2n + 1$ is also prime.

COURTESY OF GERHARD FREY; SLIM FILMS

GERHARD FREY suggested a new strategy for attacking Fermat's last theorem in 1984: Suppose that A and B are perfect nth powers such that $A + B$ is again an nth power—that is, they are a solution to Fermat's equation. A and B can then be used as coefficients in a special elliptic curve: $y^2 = x(x - A)(x + B)$; the "discriminant" of this elliptic curve, $A^2B^2(A + B)^2$, is also a perfect nth power. Frey suspected that such an elliptic curve could not be modular. In other words, Frey pointed out that if someone proved that the Shimura-Taniyama conjecture is true or that all elliptic curves are modular, then they might be able to show that the elliptic equation $y^2 = x(x - A)(x + B)$ cannot exist—in which case, the solution to Fermat's equation cannot exist, and Fermat's last theorem is proved true.

center stage. It was possible to prove Fermat's last theorem by proving the Shimura-Taniyama conjecture. Wiles, who was by now a professor at Princeton, wasted no time. For seven years, he worked in complete secrecy. Not only did he want to avoid the pressure of public attention, but he hoped to keep others from copying his ideas. During this period, only his wife learned of his obsession—on their honeymoon.

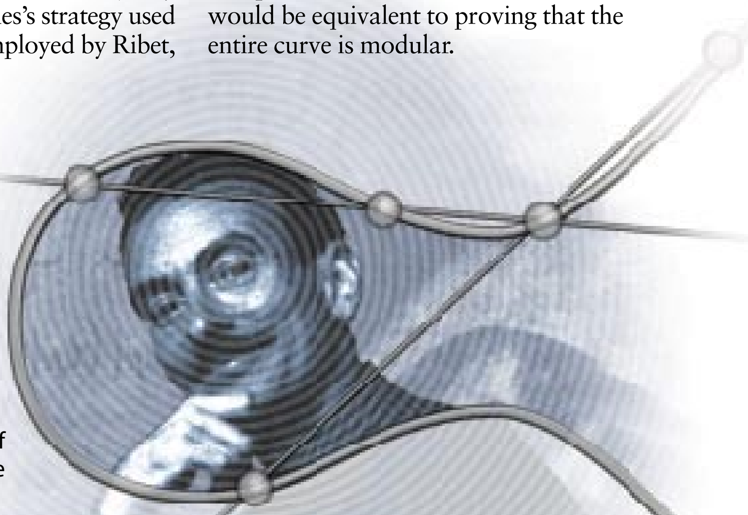
Seven Years of Secrecy

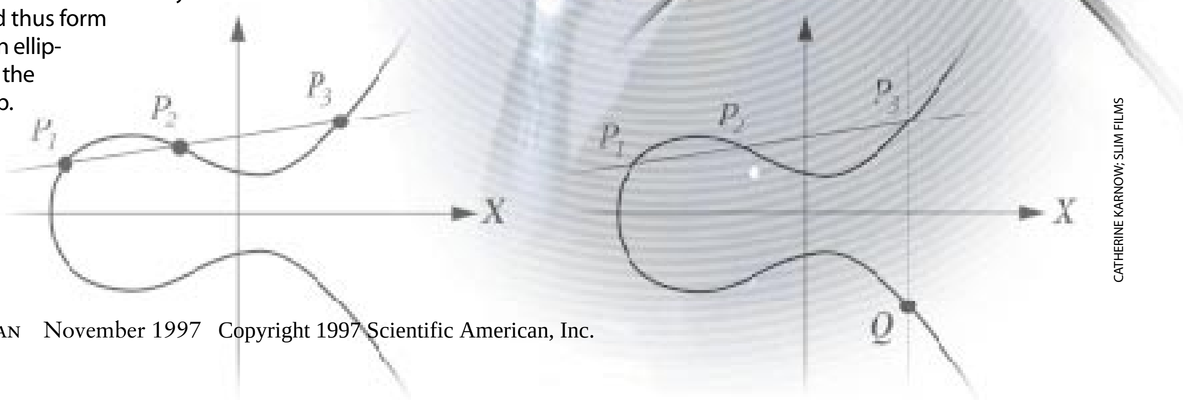
Wiles had to pull together many of the major findings of 20th-century number theory. When those ideas were inadequate, he was forced to create other tools and techniques. He describes his experience of doing mathematics as a journey through a dark, unexplored mansion: "You enter the first room of the mansion, and it's completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch. You turn it on, and suddenly it's all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark. So each of these breakthroughs, while sometimes they're momentary, sometimes over a period of a day or two, they are the culmination of, and couldn't exist without, the many months of stumbling around in the dark that precede them."

As it turned out, Wiles did not have to prove the full Shimura-Taniyama conjecture. Instead he had to show only that a particular subset of elliptic curves—one that would include the hypothetical elliptic curve Frey proposed, should it exist—is modular. It wasn't really much of a simplification. This subset is still infinite in size and includes the majority of interesting cases. Wiles's strategy used the same techniques employed by Ribet, plus many more. And as with Ribet's argument, it is possible to give only a hint of the main points involved.

The difficulty was to show that every elliptic curve in Wiles's subset is modular. To do so, Wiles exploited the group property of points on the elliptic curves and applied a theorem of Robert P. Langlands of the Institute for Advanced Study in Princeton, N.J., and Jerrold Tunnell of Rutgers University. The theorem shows, for each elliptic curve in Wiles's set, that a specific group of points inside the elliptic curve is modular. This requirement is necessary but not sufficient to demonstrate that the elliptic curve as a whole is modular.

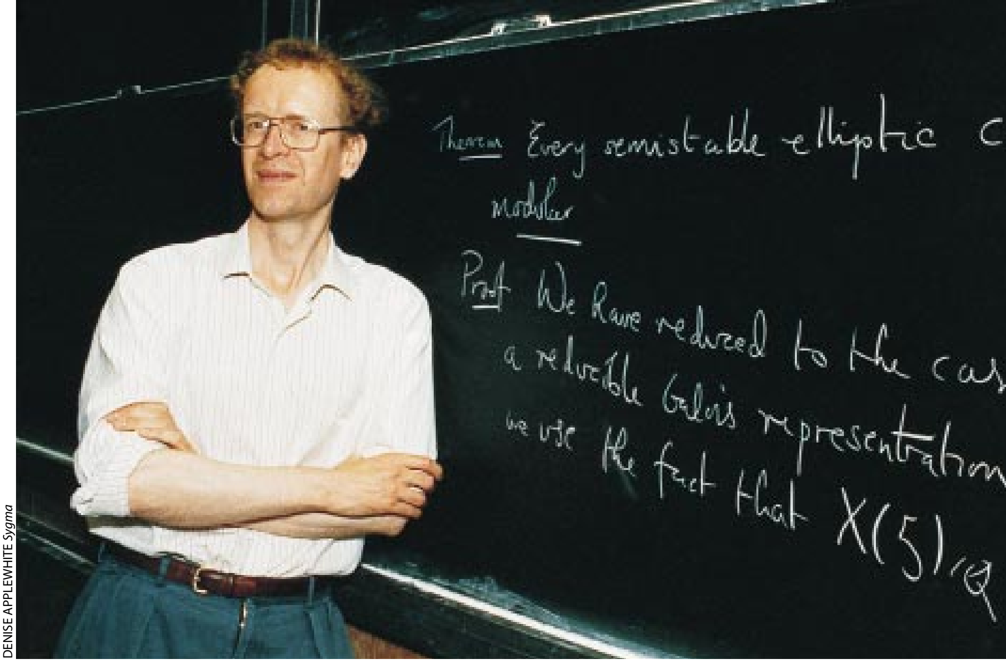
The group in question has only nine elements, so one might imagine that its modularity represents an extremely small first step toward complete modularity. To close this gap, Wiles wanted to examine increasingly larger groups, stepping from groups of size 9 to 9^2, or 81, then to 9^3, or 729, and so on. If he could reach an infinitely large group and prove that it, too, is modular, that would be equivalent to proving that the entire curve is modular.





KENNETH A. RIBET followed Frey's lead and in June 1986 proved that any elliptic curve could not be modular if its discriminant were a perfect nth power. Ribet's proof depends on a geometric method for "adding" points on an elliptic curve. Visually the idea is that it is possible to project a line through a pair of points on the elliptic curve, P_1 and P_2, to obtain a third point, P_3. This new point is then reflected in the x axis to obtain Q, which is said to be the sum of P_1 and P_2. Whereas the set of all points on an elliptic curve is infinite, there are finite sets of points having the crucial property that the sum of any two points in the set is again in the set. Such finite sets obey certain special axioms and thus form so-called finite groups. If an elliptic curve is modular, so are the points in each finite group. Ribet proved that a specific finite group of Frey's curve cannot be modular, ruling out the modularity of the whole curve.

CATHERINE KARNOW; SLIM FILMS



DENISE APPLEWHITE *Sygma*

"EUREKA!" read a *New York Times* headline after Wiles revealed his first proof of Fermat's last theorem at a lecture in June 1993. Soon thereafter, though, reviewers found a serious flaw. Wiles discussed the error only with his former student Richard Taylor. Together they tried to patch up the method Wiles had used and applied tools that he had previously rejected. At last, on September 19, 1994, they found the vital fix.

Wiles accomplished this task via a process loosely based on induction. He had to show that if one group was modular, then so must be the next larger group. This approach is similar to toppling dominoes: to knock down an infinite number of dominoes, one merely has to ensure that knocking down any one domino will always topple the next. Eventually Wiles felt confident that his proof was complete, and on June 23, 1993, he announced his result at a conference at the Isaac Newton Mathematical Sciences Institute in Cambridge. His secret research program had been a success, and the mathematical community and the world's press were surprised and delighted by his proof. The front page of the *New York Times* exclaimed, "At Last, Shout of 'Eureka!' in Age-Old Math Mystery."

As the media circus intensified, the official peer-review process began. Almost immediately, Nicholas M. Katz of Princeton uncovered a fundamental and devastating flaw in one stage of Wiles's argument. In his induction process, Wiles had borrowed a method from Victor A. Kolyvagin of Johns Hopkins University and Matthias Flach of the California Institute of Technology to show that the group is modular. But it now seemed that this method could not be relied on in this particular instance. Wiles's childhood dream had turned into a nightmare.

Finding the Fix

For the next 14 months, Wiles hid himself away, discussing the error only with his former student Richard Taylor. Together they wrestled with the problem, trying to patch up the method Wiles had already used and applying other tools that he had previously rejected. They were at the point of admitting defeat and releasing the flawed proof so that others could try to correct it, when, on September 19, 1994, they found the vital fix. Many years earlier Wiles had considered using an alternative approach based on so-called Iwasawa theory, but it floundered, and he abandoned it. Now he realized that what was causing the Kolyvagin-Flach method to fail was exactly what would make the Iwasawa theory approach succeed.

Wiles recalls his reaction to the discovery: "It was so indescribably beautiful; it was so simple and so elegant. The first night I went back home and slept on it. I checked through it again the next morning, and I went down and told my wife, 'I've got it. I think I've found it.' And it was so unexpected that she thought I was talking about a children's toy or something, and she said, 'Got what?' I said, 'I've fixed my proof. I've got it.'"

For Wiles, the award of the Wolfskehl Prize marks the end of an obsession that lasted more than 30 years: "Having solved this problem, there's certainly a sense of freedom. I was so obsessed by this problem that for eight years I was thinking about it all of the time—when I woke up in the morning to when I went to sleep at night. That particular odyssey is now over. My mind is at rest." For other mathematicians, though, major questions remain. In particular, all agree that Wiles's proof is far too complicated and modern to be the one that Fermat had in mind when he wrote his marginal note. Either Fermat was mistaken, and his proof, if it existed, was flawed, or a simple and cunning proof awaits discovery. SA

The Authors

SIMON SINGH and KENNETH A. RIBET share a keen interest in Fermat's last theorem. Singh is a particle physicist turned television science journalist, who wrote *Fermat's Enigma* and co-produced a documentary on the subject. Ribet is a professor of mathematics at the University of California, Berkeley, where his work focuses on number theory and arithmetic algebraic geometry. For his proof that the Shimura-Taniyama conjecture implies Fermat's last theorem, Ribet and his colleague Abbas Bahri won the first Prix Fermat.

Further Reading

Yutaka Taniyama and His Time: Very Personal Recollections from Shimura. Goro Shimura in *Bulletin of the London Mathematical Society,* Vol. 21, pages 186–196; 1989.

From the Taniyama-Shimura Conjecture to Fermat's Last Theorem. Kenneth A. Ribet in *Annales de la Faculté des Sciences de L'Université de Toulouse,* Vol. 11, No. 1, pages 115–139; 1990.

Modular Elliptic Curves and Fermat's Last Theorem. Andrew Wiles in *Annals of Mathematics,* Vol. 141, No. 3, pages 443–551; May 1995.

Ring Theoretic Properties of Certain Hecke Algebras. Richard Taylor and Andrew Wiles in *Annals of Mathematics,* Vol. 141, No. 3, pages 553–572; May 1995.

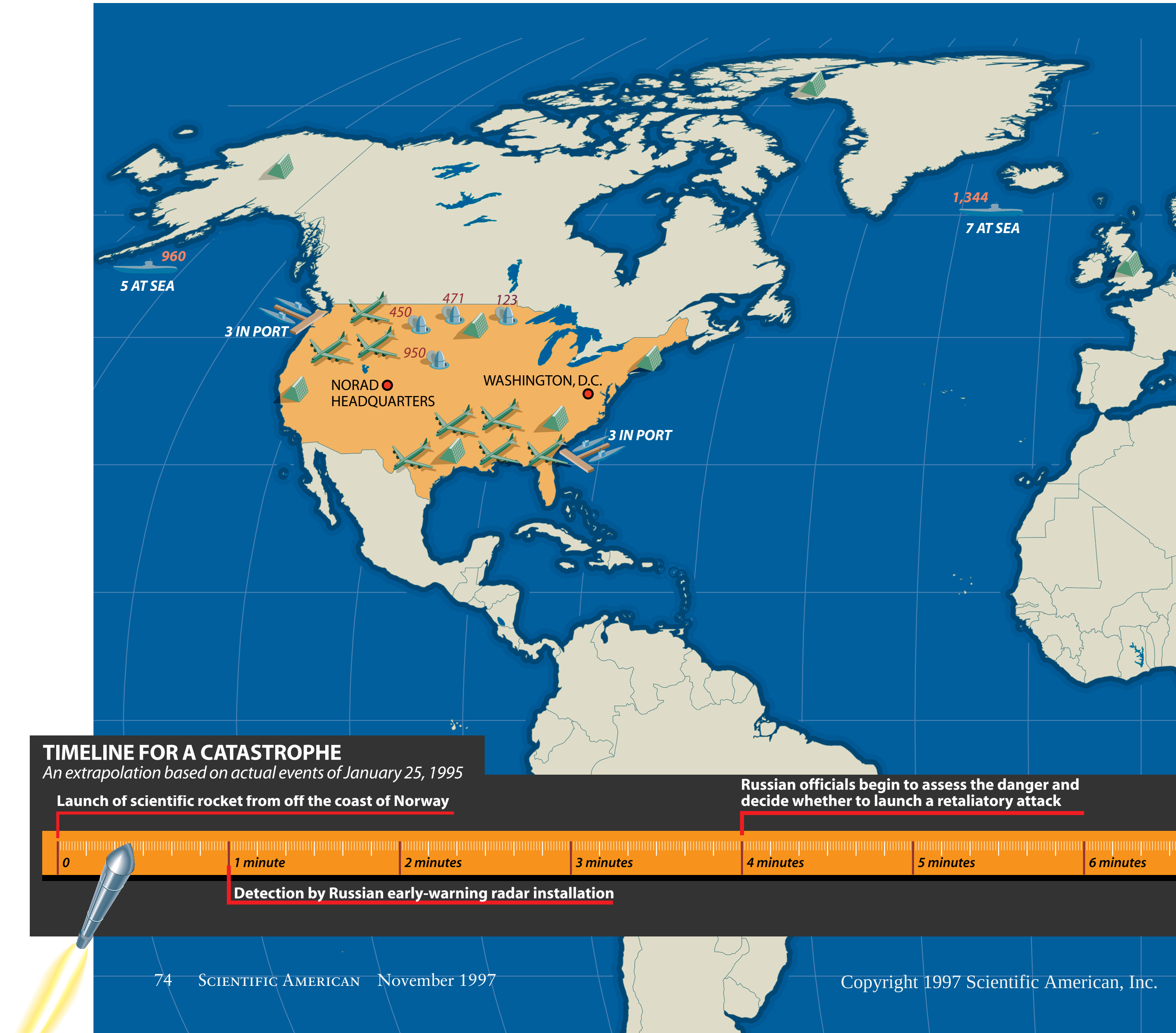
Notes on Fermat's Last Theorem. A. J. van der Poorten. Wiley Interscience, 1996.

Fermat's Enigma. Simon Singh. Walker and Company, 1997.

Taking Nuclear Weapons off Hair-Trigger Alert

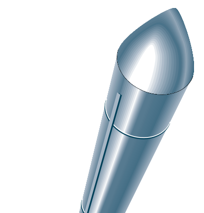
It is time to end the practice of keeping nuclear missiles constantly ready to fire. This change would greatly reduce the possibility of a mistaken launch

by Bruce G. Blair, Harold A. Feiveson and Frank N. von Hippel

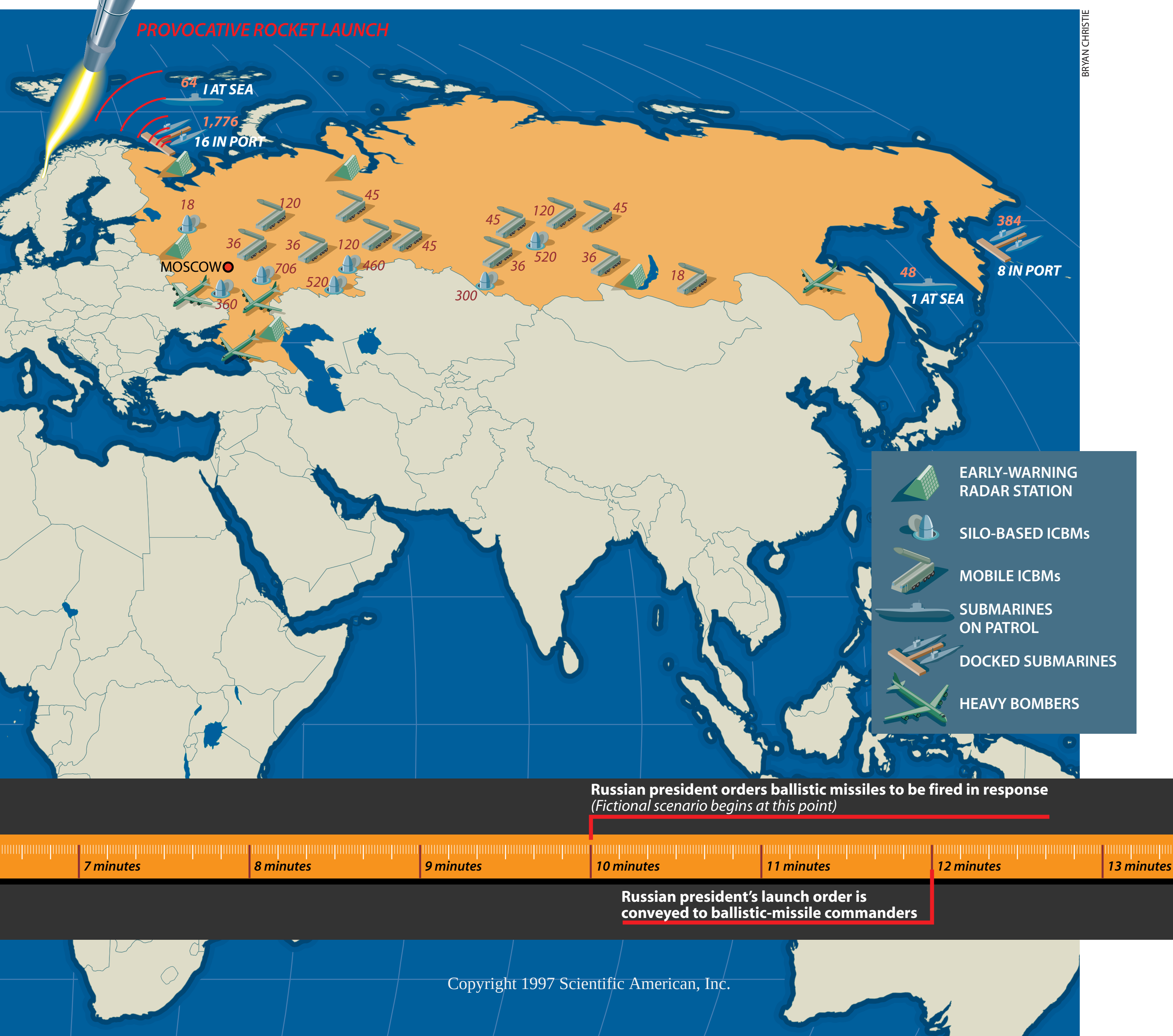


On January 25, 1995, military technicians at a handful of radar stations across northern Russia saw a troubling blip suddenly appear on their screens. A rocket, launched from somewhere off the coast of Norway, was rising rapidly through the night sky. Well aware that a single missile from a U.S. submarine plying those waters could scatter eight nuclear bombs over Moscow within 15 minutes, the radar operators immediately alerted their superiors. The message passed swiftly from Russian military authorities to President Boris Yeltsin, who, holding the electronic case that could order the firing of nuclear missiles in response, hurriedly conferred by telephone with his top advisers. For the first time ever, that "nuclear briefcase" was activated for emergency use.

For a few tense minutes, the trajectory of the mysterious rocket remained unknown to the worried Russian officials. Anxiety mounted when the separation of multiple rocket stages created an impression of a possible attack by several missiles. But the radar crews continued to track their targets, and after about eight minutes (just a few minutes short of the procedural deadline to respond to an impending nuclear attack), senior military officers determined that the rocket was headed far out to sea and posed no threat to Russia. The unidentified rocket in this case turned out to



EQUIPMENT FOR NUCLEAR WAR maintained by the U.S. and Russia includes long-range bombers, ballistic-missile submarines, land-based intercontinental ballistic missiles (ICBMs), early-warning radars and satellites. Despite the conclusion of the cold war, these two former adversaries remain ready to launch thousands of nuclear warheads (*numbers indicated on map*) at each other on minutes' notice.



be a U.S. scientific probe, sent up to investigate the northern lights. Weeks earlier the Norwegians had duly informed Russian authorities of the planned launch from the offshore island of Andoya, but somehow word of the high-altitude experiment had not reached the right ears.

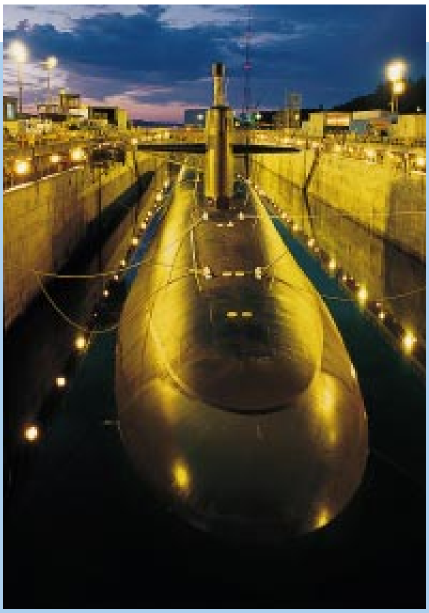
That frightening incident (like some previous false alarms that activated U.S. strategic forces) aptly demonstrates the danger of maintaining nuclear arsenals in a state of hair-trigger alert. Doing so heightens the possibility that one day someone will mistakenly launch nuclear-tipped missiles, either because of a technical failure or a human error—a mistake made, perhaps, in the rush to respond to false indications of an attack.

Both the U.S. and Russian military have long instituted procedures to prevent such a calamity from happening. Designers of command systems in Russia have gone to extraordinary lengths to ensure strict central control over nuclear weapons. But their equipment is not foolproof, and Russia's early-warning and nuclear command systems are deteriorating. This past February the institute responsible for designing the sophisticated control systems for the Strategic Rocket Forces (the military unit that operates Russian intercontinental ballistic missiles) staged a one-day strike to protest pay arrears and the lack of resources to upgrade their equipment. Three days later Russia's defense minister, Igor Rodionov, asserted that "if the shortage of funds persists . . . Russia may soon approach a threshold beyond which its missiles and nuclear systems become uncontrollable."

Rodionov's warning may have been, in part, a maneuver to muster political support for greater defense spending. But recent reports by the U.S. Central Intelligence Agency confirm that Russia's Strategic Rocket Forces have indeed fallen on hard times. Local utility managers have repeatedly shut off the power to various nuclear weapons installations after the military authorities there failed to pay their electric bills. Worse yet, the equipment that controls nuclear weapons frequently malfunctions, and critical electronic devices and computers sometimes switch to a combat mode for no apparent reason. On seven occasions during the fall of 1996, operations at some nuclear weapons centers were severely disrupted when thieves tried to "mine" critical communications cables for their copper.

Many of the radars constructed by the former Soviet Union to detect a ballistic-missile attack no longer operate, so information provided by these installations is becoming increasingly unreliable. Even the nuclear suitcases that accompany the president, defense minister and chief of the General Staff are reportedly falling into disrepair. In short, the systems built to control Russian nuclear weapons are now crumbling.

In addition to these many technical difficulties, Russia's nuclear weapons establishment suffers from a host of human and organizational problems. Crews receive less training than they did formerly and are consequently less proficient in the safe handling of nuclear weapons. And despite President Yeltsin's promises to improve conditions, endemic housing and food shortages have led to demoralization and disaffection within the elite Strategic Rocket Forces, the strategic submarine fleet and the custodians of Russia's stock-



U.S. BALLISTIC-MISSILE SUBS such as this vessel carry 24 multiwarhead missiles.

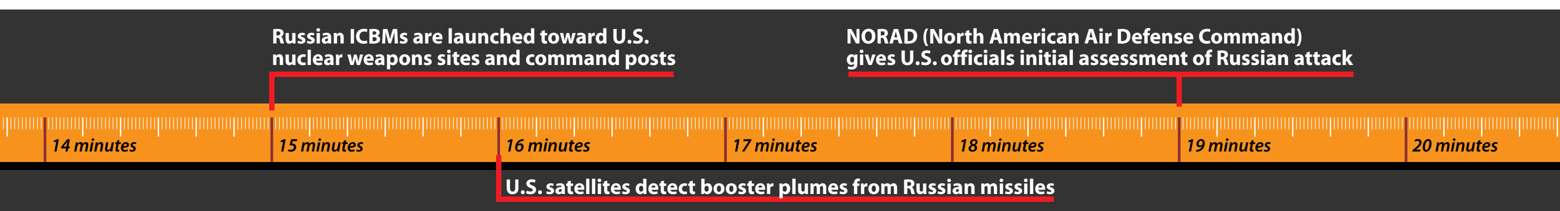
ROBERT GENAT *Arms Communications*

Submarine-Launched Missiles

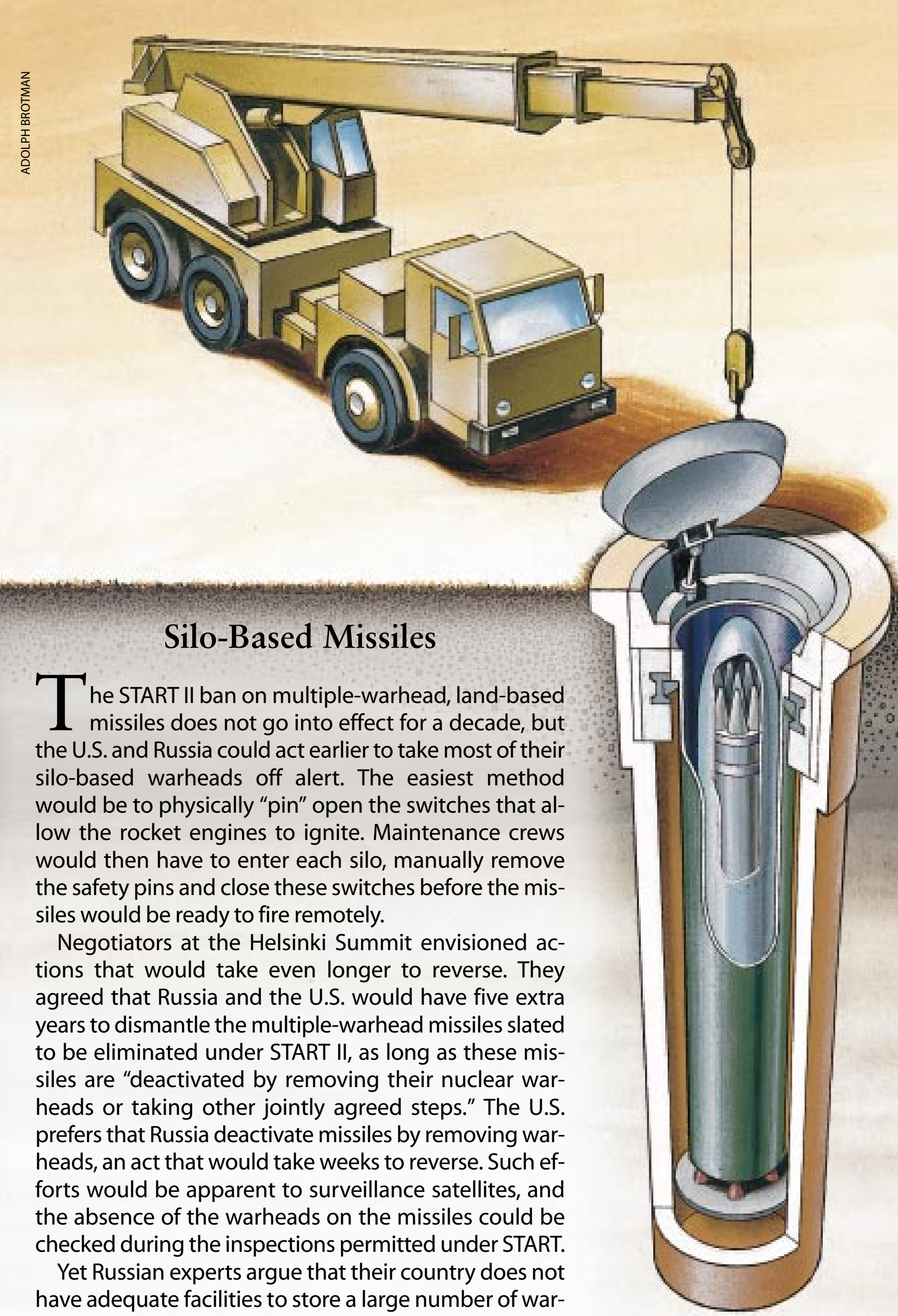
To achieve START II limits, the U.S. plans to eliminate four of its 18 ballistic-missile submarines and to reduce the count of warheads on submarine-launched missiles from eight to five. Later, to meet the START III goals, the U.S. would most likely eliminate an additional four submarines and reduce the number of warheads on each missile to four. All these actions should be taken at once. Russia could then immediately remove the warheads from the submarines it plans to eliminate under the START agreements.

Without rather elaborate verification arrangements, neither country could determine the status of the other's submarines at sea. Both nations, however, should lower launch readiness. Approximately half the submarines that the U.S. has at sea today are traveling to their launch stations in a state of modified alert: the crew needs about 18 hours to perform the procedures, such as removing the flood plates from the launch tubes, that bring a submarine to full alert. Most U.S. submarines at sea could simply stay on modified alert. Their readiness could be reduced further by removing their missiles' guidance systems and storing them on board. Russian submarines lack this option; their missiles are not accessible from inside the vessel.

Russia should also pledge to keep its missiles on submarines in port off launch-ready alert. (The U.S. does not maintain submarines in port on alert.) The U.S. may be able to monitor the alert condition of these Russian submarines, but Russia should make their status obvious. *—B.G.B., H.A.F. and F.N. von H.*



ADOLPH BROTMAN



Silo-Based Missiles

The START II ban on multiple-warhead, land-based missiles does not go into effect for a decade, but the U.S. and Russia could act earlier to take most of their silo-based warheads off alert. The easiest method would be to physically "pin" open the switches that allow the rocket engines to ignite. Maintenance crews would then have to enter each silo, manually remove the safety pins and close these switches before the missiles would be ready to fire remotely.

Negotiators at the Helsinki Summit envisioned actions that would take even longer to reverse. They agreed that Russia and the U.S. would have five extra years to dismantle the multiple-warhead missiles slated to be eliminated under START II, as long as these missiles are "deactivated by removing their nuclear warheads or taking other jointly agreed steps." The U.S. prefers that Russia deactivate missiles by removing warheads, an act that would take weeks to reverse. Such efforts would be apparent to surveillance satellites, and the absence of the warheads on the missiles could be checked during the inspections permitted under START.

Yet Russian experts argue that their country does not have adequate facilities to store a large number of warheads taken from missiles. They are now considering other options: immobilizing the massive silo lids so that heavy equipment would be required to open them, or removing the battery that operates the missile-guidance system during flight. A third possibility would be to replace the aerodynamic missile nose cones with flat-faced covers, which would shelter the warheads but not allow the missiles to fly. *—B.G.B., H.A.F. and F.N. von H.*

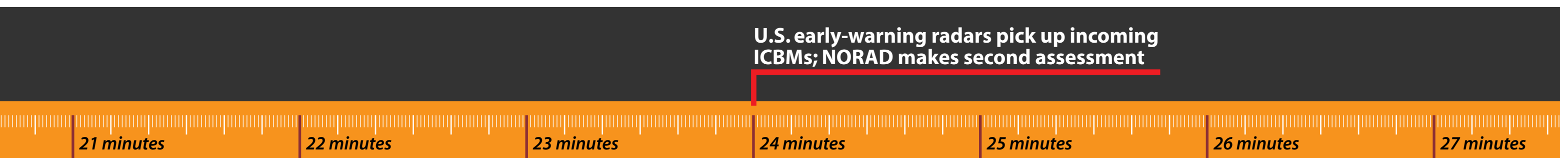
RUSSIAN SILO LID would require a large crane to tilt upward if the device that generates high-pressure gas for its pneumatically operated hinge were purposefully removed.

piles of nuclear warheads. As a result, the likelihood increases that desperate low-level commanders might disregard safety rules or, worse still, that they might take unauthorized control of nuclear weapons—something a deteriorating central command might be unable to prevent or counter. Although most Russian launch crews would need to receive special codes held by the General Staff before they could fire their missiles, one recent CIA report warned that some submarine crews may be able to launch the ballistic missiles on board their vessels without having to obtain such information first.

Even at the top, control over nuclear weapons could splinter along various political fault lines. Relations between politicians and military leaders in Russia are strained, and physical control of the launch codes remains in the hands of the military. Thus, the authority to fire ballistic missiles could be usurped by military commanders during an internal crisis. In fact, during the August 1991 coup against President Mikhail S. Gorbachev, top-level allegiances suddenly shifted, and the normal chain of command for Russia's nuclear weapons was broken. For three days, the power to launch nuclear weapons rested in the hands of Defense Minister Dmitri Yazov and the chief of the General Staff, Mikhail Moiseyev. Given the dire conditions in Russia, something similar could happen again.

The Nuclear Hair Trigger

Although international relations have changed drastically since the end of the cold war, both Russia and the U.S. continue to keep the bulk of their nuclear missiles on high-level alert. So within just a few minutes of receiving instructions to fire, a large fraction of the U.S. and Russian land-based rockets (which are armed with about 2,000 and 3,500 warheads, respectively) could begin their 25-minute flights over the North Pole to their wartime targets.



Less than 15 minutes after receiving the order to attack, six U.S. Trident submarines at sea could loft roughly 1,000 warheads, and several Russian ballistic-missile submarines could dispatch between 300 and 400. In sum, the two nuclear superpowers remain ready to fire a total of more than 5,000 nuclear weapons at each other within half an hour.

Why do two countries at peace retain such aggressive postures, ones that perpetuate the danger of a mistaken or unauthorized launch? Because military planners on both sides remain fixated on the remote specter of a deliberate nuclear surprise attack from their former adversary. They assume that such a "first strike" would be aimed against their own strategic nuclear weapons and the command centers that direct them. To deter such an assault, each country strives to ensure that it could respond with a forceful counterattack against the full spectrum of military targets on its opponent's territory, including all nuclear weapons installations. This requirement saddles military planners with a task virtually identical in scope to mounting a first strike: they must be able to guarantee the rapid destruction of thousands of targets spread across a distant continent.

In order to meet this demand, both the U.S. and Russia rely on a launch-on-warning strategy—that is, each side is poised to release a massive retaliatory missile salvo after detecting an enemy missile attack but before the incoming warheads arrive (which might take just 15 minutes if they were fired from submarines nearby). Although it has thousands of warheads securely deployed at sea, the U.S. adheres to this quick-draw stance because of the vulnerability of its missile silos and command apparatus, including its political and military leadership in Washington, D.C.

Russian officials perceive an even greater need to launch their missiles on warning. The General Staff evidently fears that if its nuclear missiles are not launched immediately, then only tens of them would be able to respond after absorbing a systematic U.S. attack. Russian command posts and missile silos are as vulnerable as those of the U.S. to a massive assault.

Russia's current inability to deploy many of its most survivable forces—submarines at sea and mobile land-based rockets—amplifies this worry. A lack of resources and qualified personnel has forced the Russian navy to cut back operations considerably. At present, the Russian navy typically keeps only two of its 26 ballistic missile submarines at sea on combat patrol at any one time. Similar constraints prevent Russia from hiding more than one or two regiments of its truck-mounted mobile missiles by dispersing them in the field. The remaining 40 or so regiments, each controlling nine single-warhead missiles, keep their trucks parked in garages. These missiles are more exposed to attack than those housed in underground silos. Russia also has 36 10-warhead nuclear missiles carried on railway cars, which were designed to be hidden along Russia's vast rail network. But these railcars remain confined to fixed garrisons in keeping with a decision made by President Gorbachev in 1991.

These vulnerabilities have led Russia to ready some of its submarines in port and mobile missiles in garages to launch on warning, along with the missiles in silos. The time available for deciding to launch these weapons is shortened by the presence of American, British and French submarines cruising in the North Atlantic, only about 2,000 miles (3,200 kilometers) from Moscow. This proximity means that the nuclear-release procedures in Russia require a response time of less than 15 minutes: a few minutes for detecting an attack, another few minutes for top-level decision making and a few minutes for disseminating the launch order. Russian leaders and missile controllers are geared to work within this brief time frame and practice regularly with drills. U.S. nuclear forces operate with a similarly short fuse.

It is obvious that the rushed nature of this process, from warning to decision to action, risks causing a catastrophic

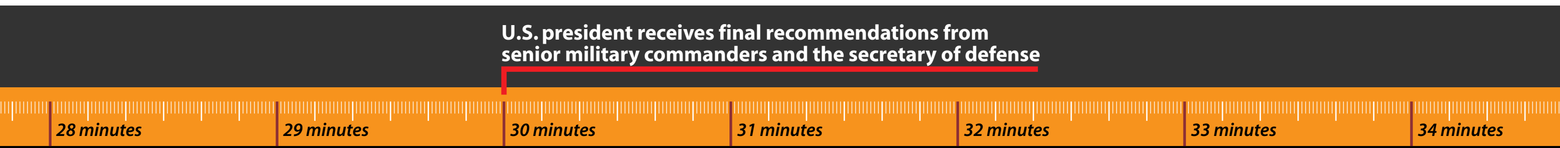
Land-Mobile Missiles

De-alerting" Russia's mobile land-based missiles (the U.S. has none) could begin with removing warheads from the 36 rail-mobile missiles to be eliminated under START II. For the truck-mobile missiles, one possibility might be to alter their garages. Currently the roofs of these shelters are designed to slide open, allowing the launcher inside to tilt upright and fire the missile. Other measures might incapacitate the launcher itself in ways that would take at least some hours to restore.

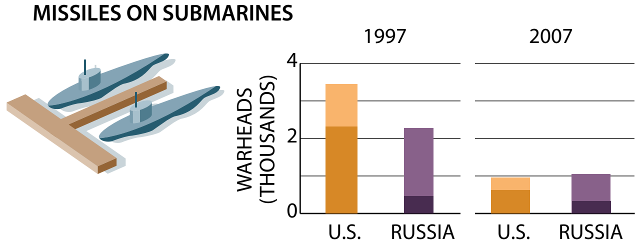
—B.G.B., H.A.F. and F.N. von H.

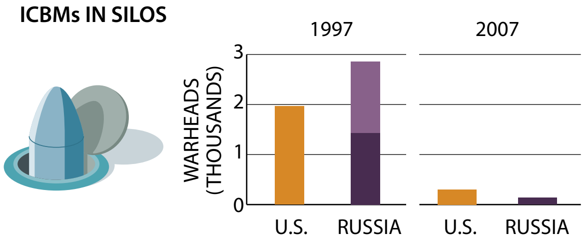
SOVFOTO/EASTFOTO

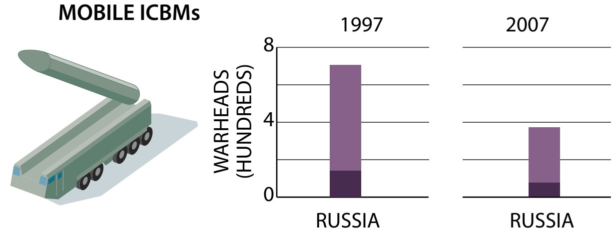
RUSSIAN SS-25 MISSILE can be fired from a truck.



STRATEGIC MISSILE TOTALS for the U.S. and Russia should shrink over the next decade in compliance with the Strategic Arms Reduction Treaties (START). Still, each country could hold from 500 to 1,000 warheads ready to fire under these agreements. (Shaded bars indicate the number of warheads kept on constant alert.)







BRYAN CHRISTIE

mistake. The danger is compounded by the erosion of Russia's ability to distinguish reliably between natural phenomena or peaceful ventures into space and a true missile attack. Only one third of its modern early-warning radars are working at all, and at least two of the nine slots in its constellation of missile-warning satellites are empty.

The dangers stemming from this decline in Russia's technical capabilities are offset, to some extent, by the relaxation of tensions that has come with the end of the cold war. Given the milder political climate, decision makers on both sides should be more inclined to question the validity of any reports they receive of an impending missile attack. Nevertheless, the coupling of two arsenals geared for rapid response carries the inherent danger of producing a mistaken launch and an escalating volley of missiles in return. The possibility of such an apocalyptic accident cannot be ruled out even under normal conditions. And if the control of Russian nuclear weapons were to be stressed by an internal or international political crisis, the danger could suddenly become much more acute.

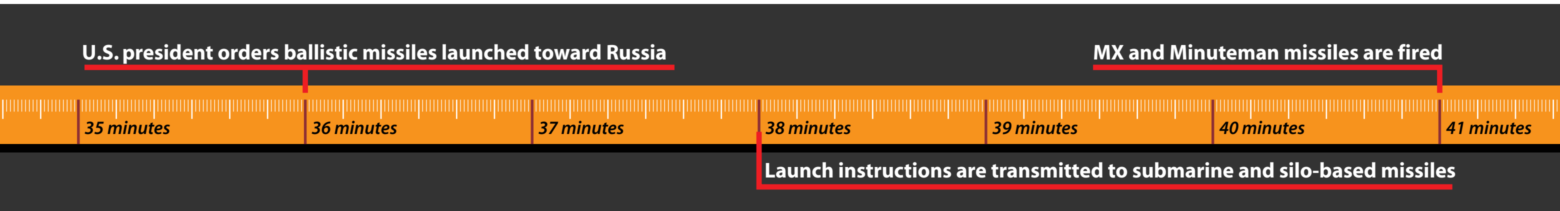
During the cold war, such risks were subordinated to the overriding requirement to deter an enemy believed to be willing to launch a nuclear attack. This rationalization is no longer defensible, if ever it was. Today, when both countries seek normal economic relations and cooperative security arrangements, perpetuating the readiness to launch nuclear weapons on the mere warning of an attack constitutes reckless behavior. Yet this thinking is so entrenched that it will yield only to steady pressure from the public on political leaders—especially presidents—to replace it with a safer policy.

"De-alerting" Missiles

The cuts in nuclear arms set by the Strategic Arms Reduction Treaties (START) should lessen the threat of an accidental nuclear exchange, but those changes will come only gradually. Under the START III framework, endorsed in Helsinki this past spring by President Yeltsin and President Bill Clinton, the U.S. and Russian strategic arsenals would shrink to about 2,000 warheads on each side by the year 2007. But if current practices are not revised, 10 years from now half of those nuclear weapons could still remain ready to launch on a few minutes' notice.

The chance of an accidental launch could be reduced much more rapidly by "de-alerting" the missiles—increasing the amount of time needed to prepare them for launch. The U.S. and Russia should move independently down this path to a safer world, preferably taking quick strides in parallel. Two prominent proponents of this approach are former senator Sam Nunn of Georgia and retired general George L. Butler, commander in chief of the U.S. Strategic Command from 1991 to 1994. This proposal is also gaining support in the community of nongovernmental organizations involved in nuclear security and from some members of the U.S. Congress. In Russia, the Ministry of Defense is seriously studying such an alteration.

President George Bush set a notable precedent for de-alerting nuclear weapons at the end of September 1991, when the Soviet Union began to split apart in the wake of the August coup attempt. On the advice of General Butler, President Bush ordered an immediate stand-down of the many U.S. strategic bombers that had remained ready for decades to take off with only a few minutes' warning. Soon afterward, air force personnel unloaded and stored the many nuclear weapons carried on these planes. In addition, President Bush ended the alert for the strategic missiles destined to be eliminated under START I, a set composed of 450 silo-based Minuteman II rockets, along with the missiles on 10 Poseidon submarines. These im-



portant actions took only a few days.

President Gorbachev reciprocated a week later by ordering the deactivation of more than 500 land-based rockets and six strategic submarines, by promising to keep his strategic bombers at a low level of readiness and by putting the rail-based missiles in garrison. In the subsequent months, both countries also withdrew many thousands of shorter-range tactical nuclear warheads that had been deployed with their armies and navies and placed these weapons in central storage depots.

Presidents Clinton and Yeltsin took a further step together in 1994, when they agreed to stop aiming strategic missiles at each other's country. This change, though a welcome gesture, has little military significance. Missile commanders can reload target coordinates into guidance computers within seconds. In fact, the 1994 pact does not even alleviate the concern about an accidental Russian launch, because an unprogrammed missile would automatically switch back to its primary wartime target, which might be a Minuteman silo in Montana or a command center in Washington, London, Paris or Beijing. And Russian missiles, like their American counterparts, cannot be ordered to self-destruct once they are launched.

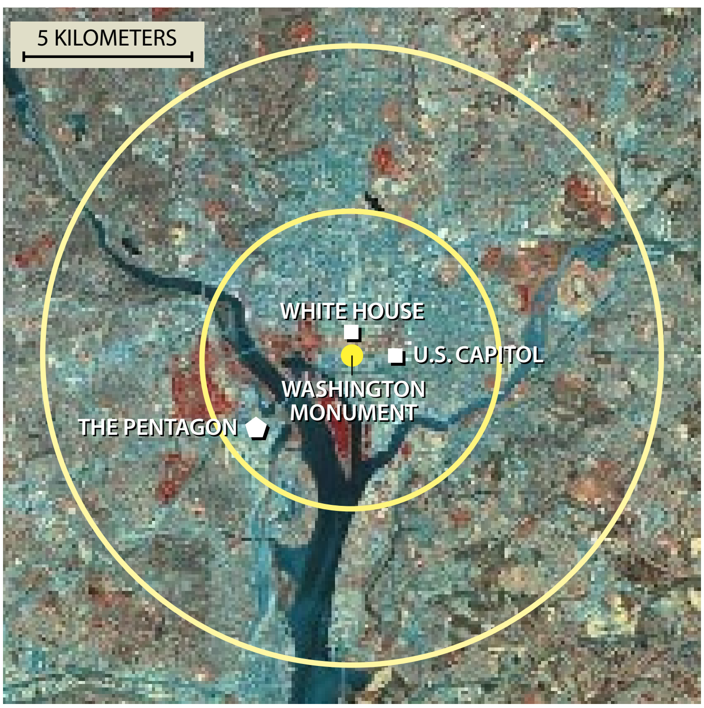
Possessing the most robust forces and cohesive command system, the U.S. government should take the lead in a new round of voluntary actions by announcing that it will withdraw the U.S. warheads that most threaten Russia's nuclear deterrent (particularly those capable of hitting Russia's missile silos and underground command posts). The most menacing warheads are those deployed on the 50 MX silo-based missiles, which are armed with 10 warheads each, and the 400 high-yield W88 warheads fitted atop some of the missiles on Trident submarines. We also recommend immobilizing all of the land-based Minuteman IIIs (about 500 missiles), which are armed with three warheads each, halving the number of submarines deployed in peacetime and cutting the number of warheads on each submarine-borne missile from eight to four. The operation of ballistic-missile submarines should also be altered so that crews would require approximately one day to ready missiles for launching.

These measures would leave almost 600 U.S. warheads remaining invulnerable at sea, each capable of destroying the heart of a great city. With such a force, the U.S. would preserve ample capacity to deter any nuclear aggressor. Such a dramatic shift by the U.S. would fully establish its intention not to pose a first-strike threat to Russia. We believe this change in policy would persuade Russia to follow suit and take most of its missiles off hair-trigger alert. These changes would also help accelerate the implementation of agreements for disarmament already negotiated under START II and START III. We estimate that most of the job could be completed within a year or two.

Capabilities already exist to confirm that nuclear weapons have been taken off alert. For instance, the number of ballistic-missile submarines in port can be monitored using satellites, and most other measures could be checked during the random on-site inspections permitted by START I. Over the longer term, additional technical means could be engineered to provide more frequent checks that nuclear missiles posed no immediate threat. For example, electronic "seals" could be used to ensure that a component removed from a missile had not been replaced. The integrity of such seals could be verified remotely through satellite relay using encrypted communications.

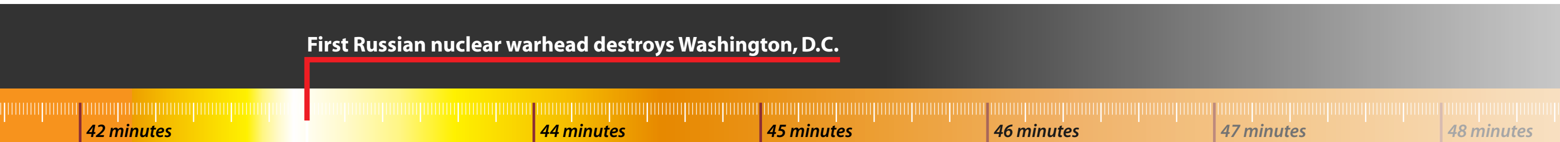
Global Zero Alert

This blueprint for taking U.S. and Russian nuclear forces off alert would substantially diminish the ability of either country to mount a first strike. Thus, it would eliminate both the ca-



CNES: LICENSED BY SPOT IMAGE CORPORATION *Photo Researchers*

ZONES OF DESTRUCTION, were a 500-kiloton nuclear warhead to explode over the Washington Monument, would cover hundreds of square kilometers around metropolitan Washington, D.C. The inner circle encompasses the area where most people would die from the immediate blast. The outer circle delimits the area where many more would perish from subsequent firestorms in built-up areas. The range of casualties would extend even farther.



A Prescription for Change

To reduce concerns that have driven Russia to maintain its missiles ready to launch on warning, the U.S. president should order the following:

1 Immediately remove to storage the warheads of the MX missiles (which will, in any event, be retired under START II).

2 Disable all Minuteman III missiles by having their safety switches pinned open (as was done for the Minuteman IIs in 1991). If Russia reciprocates, these missiles should be immobilized in a manner that would take much longer to reverse.

3 Remove to storage the warheads on the eight Trident submarines that are to be retired under START III and reduce the number of warheads on each remaining submarine missile from eight to four.

4 Take the W88 warheads off the Trident II missiles, place those warheads in storage and replace them with lower-yield weapons.

5 Allow Russia to verify these actions by using some of their annual inspections permitted by START I. Accept a greater number of inspections if Russia will also do so.

6 Put all U.S. ballistic-missile submarines at sea on a low level of alert, so that it would take at least 24 hours to prepare them to launch their missiles, and keep most submarines out of range of Russian targets. Consider ways to make these changes verifiable in the future and discuss possible reciprocal arrangements with Russian officials.

Even after these actions are taken, six submarines carrying up to 576 warheads would remain undetectable at sea, and the immobilized Minuteman IIIs could be destroyed only by a massive attack on about 500 silos.

In response to the U.S. initiative, the Russian president could order the following:

1 Remove the warheads from all 46 SS-24 rail- and silo-based missiles (which will, in any event, be retired under START II).

2 Immobilize all other silo-based missiles that are to be retired under START II.

3 Remove the warheads from the 15 ballistic-missile submarines most likely to be retired under the START agreements.

4 Place all ballistic-missile submarines (in port and at sea) in a condition such that their missiles could not be launched for at least 24 hours.

5 Disable the launchers of all truck-mobile ballistic missiles so that they cannot be activated for at least a few hours.

After these actions are taken, 128 to 400 warheads on two submarines will remain undetectable at sea, and nine to 18 SS-25 warheads on truck-mobile launchers will remain securely hidden in the field. In addition, about 2,700 warheads on silo-based ICBMs could be destroyed only by mounting successful attacks on some 340 missile silos. *—B.G.B., H.A.F. and F.N. von H.*

pacity and rationale for keeping missiles ready to fire on warning. Leaders would have to wait out any alarm of an attack before deciding how to respond, drastically reducing the risk of a mistaken or unauthorized launch.

We recognize that military leaders in the U.S. and Russia might insist on maintaining small portions of their current arsenals on high alert, perhaps hundreds of warheads each, until the other nuclear-weapon states—Britain, France and China—joined in adopting similar measures to reduce the readiness of their nuclear arsenals. But if the U.S. and Russia aspire to establish the highest possible standards of safety for their nuclear armaments, they should move as rapidly as possible to take all their missiles off alert and then follow with further steps to increase the time required to reactivate these weapons.

The ultimate goal would be to separate most, if not all, nuclear warheads from their missiles and then, eventually, to eliminate most of the stored warheads and missiles. To implement such an extensive program fully, the means for verification would have to be strengthened to ensure that every nuclear state would know whether another country was making nuclear missiles launch-ready.

Moving toward a global stand-down of nuclear arms will undoubtedly encounter strong resistance from those whose dominant fear remains a secretly prepared surprise attack. The design of procedures to take nuclear missiles off constant alert needs to take into account this already remote possibility. But these plans must urgently go forward to remove the much more immediate hazard—the mistaken or unauthorized launch of nuclear missiles. SA

The Authors

BRUCE G. BLAIR, HAROLD A. FEIVESON and FRANK N. VON HIPPEL have studied nuclear arms policy intensively. Blair served for four years in the U.S. Air Force Strategic Air Command before earning a Ph.D. in operations research in 1984 from Yale University. He is currently a defense analyst at the Brookings Institution in Washington, D.C. Feiveson received a master's degree in theoretical physics in 1959 from the University of California, Los Angeles. He worked in the U.S. Arms Control and Disarmament Agency for four years before moving to Princeton University to study public and international affairs. Feiveson received his Ph.D. in 1972 and joined the Princeton faculty in 1974. Von Hippel, who received a doctorate in theoretical physics from the University of Oxford in 1962, served in the office of the president's science adviser in 1993 and 1994 as assistant director for national security. He is currently a professor of public and international affairs at Princeton.

Further Reading

The Logic of Accidental Nuclear War. Bruce G. Blair. Brookings Institution, 1993.

Global Zero Alert for Nuclear Forces. Bruce G. Blair. Brookings Institution, 1995.

Caging the Nuclear Genie: An American Challenge for Global Security. Stansfield Turner. Westview Press, Boulder, Colo., 1997.

The Future of U.S. Nuclear Weapons Policy. National Academy of Sciences. National Academy Press, 1997.

The Parasitic Wasp's Secret Weapon

Parasitic wasps must develop inside living caterpillars. They survive this hostile environment by smuggling in a virus that suppresses their host's immune system

by Nancy E. Beckage

WASP AND CATERPILLAR are the David and Goliath of the insect world. The huge caterpillar's immune system threatens the wasp's eggs, which must mature inside a living host. But the tiny wasp prevails using a deadly weapon: a virus.

This caterpillar will never become a moth. It lurks deep in the foliage of a tasty tomato plant, hidden from predators, but its enemy has found it anyway. In search of a nanny for her offspring, the parasitic wasp has homed in on the distinctive scent of her lepidopteran victim and its lunch. Now the tiny wasp injects a clutch of eggs through the caterpillar's tough cuticle and into its body cavity, where her larvae will thrive by feeding on their living nursery. At a critical moment in development, the wasp larvae will burst through the flanks of the caterpillar to spin their cocoons on its surface. These wasps eventually depart as adults, metamorphosis complete, but their host is now destined to die as a caterpillar.

If this were a one-on-one interspecies scuffle, the caterpillar might stand a chance—it has an immune system capable of engulfing and killing invading wasp eggs before they can do permanent harm. The wasp, however, does not come to this encounter alone. In addition to her eggs, she injects hordes of virus particles. These viral warriors rapidly defeat the caterpillar's immune response, tipping the balance of power in favor of the wasp progeny. The caterpillar, doubly parasitized, slowly ceases feeding, fails to pupate and dies a premature death.

Host-parasite relationships such as those involving the wasp, the virus and the doomed caterpillar are among the most complex in nature. The wasp is an endoparasite—it must develop inside its host. If the caterpillar dies before the wasp larvae are properly fed, the wasps will die as well. Yet the caterpillar cannot be allowed to gain the upper hand using its immune defenses. Much of the responsibility for maintaining this delicate balance falls to the wasp's viral accomplice. Like the wasp, many parasites of insect hosts have evolved associations with bacteria and viruses that help them perform their often deadly deeds.

Microbial Weapons

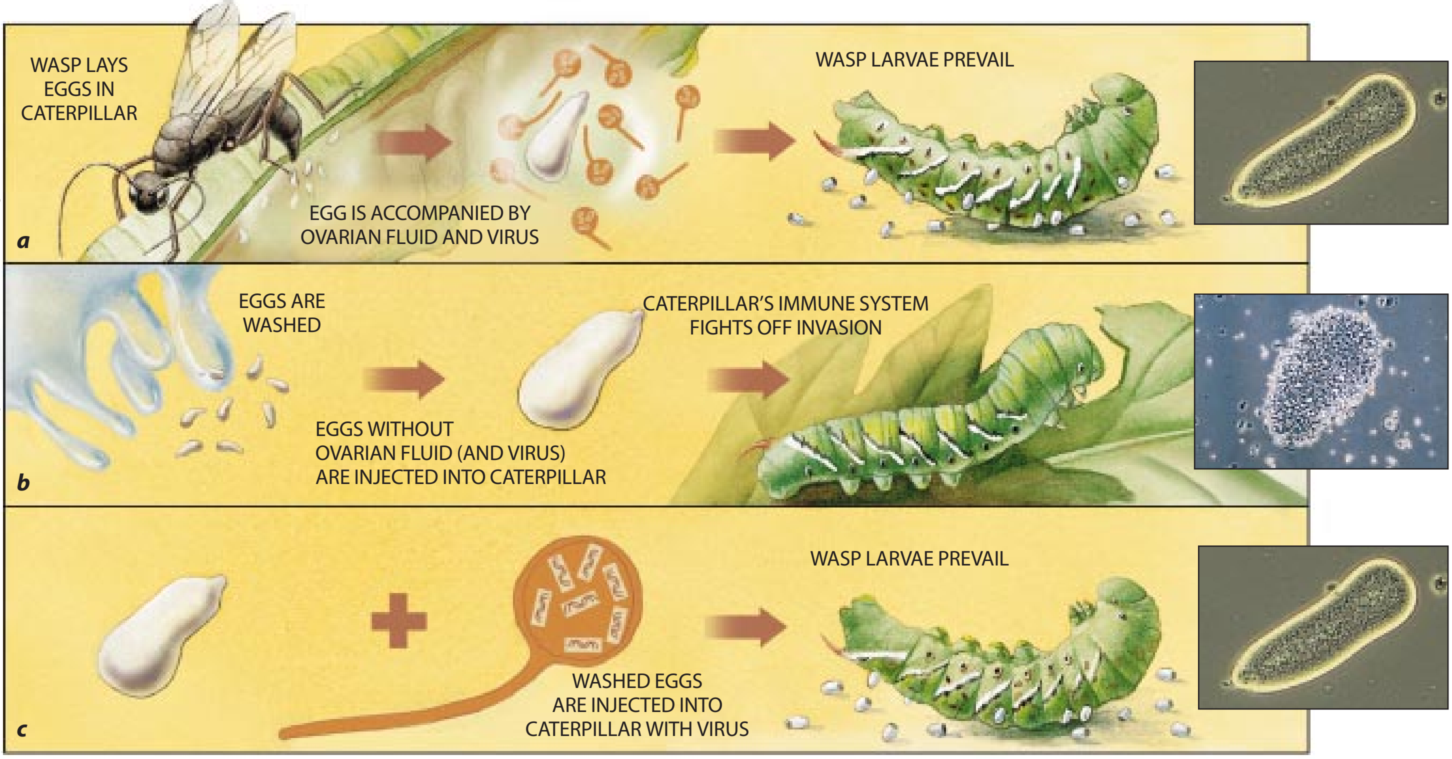
A simple example of such a partnership occurs in certain parasitic worms that carry a virulent bacterium in their digestive tracts. These worms regurgitate the bacteria into their insect hosts, killing the hosts within days of infection. The rapidly dividing bacteria are an immediate food source for the developing worms, and they provide further sustenance by secreting digestive enzymes that soon turn the host cadaver into a nutrient-rich soup. In return, the bacteria benefit by using the worms as vehicles for invasion of new hosts. These interacting partners are strictly independent organisms—they do not share genes.

In contrast, the interaction between endoparasitic wasps and the virus they exploit is more intimate. Not only are the fates of the partners intertwined, but their genetic material is also permanently mingled. And the relationship goes further—the wasp and the virus possess related genes. All this raises a thought-provoking question: Are the wasp and the virus two entities or one?

The first hint that endoparasitic wasps might have unusual weapons in their arsenal came in 1965. George Salt of the University of Cambridge suspected that female *Venturia* wasps inject substances required for successful development of their progeny into host larvae along with eggs during the process called oviposition. In particular, Salt noted that the ovary of the female wasp harbors substances that prevent destruction of wasp eggs by the caterpillar's immune cells.

Normally, injected wasp eggs float freely in the bloodlike fluid, called the hemolymph, that fills the body cavity of the caterpillar. When Salt washed the wasp eggs prior to injection, however, they provoked a rapid immune response. These eggs, stripped of an unidentified factor, were quickly attacked by the host's immune cells and ultimately killed. In 1973 electron micrographs taken by


ROBERTO OSTI



LAUREL ROGERS; STEVEN H. HARWOOD (*top and bottom micrographs*); MARK D. LAVINE (*center micrograph*)

BATTLE OF THE INSECTS rages between the caterpillar and the wasp. Wasp eggs escape attack by the caterpillar's immune system thanks to a virus injected into the caterpillar by the wasp along with her eggs (*a*). The virus disables the caterpillar's immune cells, allowing the free-floating eggs (*top micrograph*) to develop into normal wasp larvae. The fate of the caterpillar is not so fortunate. In the laboratory, however, wasp eggs that have been washed lack the protective virus (*b*) and are rapidly engulfed by the caterpillar's immune cells (*center micrograph*); no wasps survive this encounter. Injecting pure virus along with the washed eggs (*c*) allows the wasp eggs (*bottom micrograph*) to develop into larvae and emerge from the caterpillar.

Susan Rotheram, also at Cambridge, offered a clue to the identity of the protective substance. These images showed that the surface of a *Venturia* egg becomes impregnated with viruslike particles as it passes through the wasp's oviduct during oviposition.

Nearly a decade later Donald B. Stoltz of Dalhousie University conducted an extensive taxonomic survey of parasitic wasps in collaboration with S. Bradleigh Vinson of Texas A&M University. They showed that viruslike particles are invariably found in certain wasp species that develop as internal parasites of lepidopteran hosts. Moreover, they observed that these viruses replicate exclusively in the ovarian tissue of female wasps. During oviposition, the wasp injects thousands of virions into the caterpillar along with one or more eggs.

It seemed reasonable to suspect that these viruses are the ovary-derived substances that accompany wasp eggs into the host and squelch the host's immune response. The evidence was purely circumstantial until 1981, when Stolz, Vinson and their co-workers finally confirmed that this job can be performed by purified virus. But how exactly do the viruses—today called polydnaviruses (pronounced "puh-LID-nah-viruses")—disable the caterpillar immune system?

Immune Deficiency

To answer this question, my colleagues and I study the parasitic wasp *Cotesia congregata*, which can lay hundreds of eggs in each caterpillar. These eggs hatch into larvae that dine on the host's hemolymph fluid instead of consuming its tissue, thereby allowing the infected caterpillar to survive well past emergence of the wasp progeny. *Manduca sexta*, the tobacco hornworm, serves as our model host. Anyone who grows tomatoes has probably had a run-in with these giant leaf-green caterpillars, which forage on tomato, tobacco and jimsonweed and often grow to the size of a man's little finger. Tobacco hornworms are convenient to work with in the laboratory: obtaining blood samples from these enormous caterpillars is easier than obtaining samples from mice.

We have observed one immediate consequence of the parasitism by *Cotesia* wasps: certain cells, known as hemocytes, circulating in the caterpillar's blood undergo rapid physical transformation. Graduate student Mark D. Lavine has seen these effects within a few hours of oviposition. Affected hemocytes "round up," failing to adhere to substrates such as glass or parasite eggs. They also undergo extensive blebbing, or pinching off of bits of their membrane and cell contents. The damaged hemocytes clump together and are removed from circulation. Overall, this transformation bears a striking resemblance to the cell suicide, or apoptosis, that occurs in mammalian cells [see "Cell Suicide in Health and Disease," by Richard C. Duke, David M. Ojcius and John Ding-E Young; Scientific American, December 1996].

Granulocytes and plasmatocytes are among the caterpillar hemocytes most affected by parasitism; Michael R. Strand of the University of Wisconsin has shown that granulocytes in particular die by apoptosis. These are exactly the host immune cells that respond to foreign objects, including *Cotesia* eggs. In a normal immune response, granulocytes first release granules that coat the invading egg. Plasmatocytes then adhere to the egg surface in multiple layers, forming a thick capsule that eventually kills the egg inside. Selective removal of the granulocytes and the plasmatocytes from circulation disables the caterpillar's first line of defense against the endoparasite. Similar phenomena occur during human immunodeficiency virus (HIV) infection in humans. In that case, the virus targets lymphocytes, causing the clumping and apoptotic death of the cells. Opportunistic infectious agents are then free to ravage the victim, much like the wasp progeny that overtake their unfortunate caterpillar host.

When we inject unparasitized tobacco hornworm larvae with purified polydnavirus, caterpillar hemocytes undergo changes in appearance and behavior analogous to those we observe in normal parasitism. But if we chemically inactivate the virus prior to injection, the hemocytes remain unaltered. This result suggests that virus capable of directing manufacture of viral proteins is required for immune suppression.

Graduate student Steven H. Harwood has shown that polydnaviral proteins do indeed appear rapidly in the caterpillar host. We detect the first evidence that polydnaviral genes are active within 30 minutes of oviposition. By this time viral particles have spread throughout the host, entering cells, including hemocytes. In our tobacco hornworm system, we have also shown that at least one polydnavirus-encoded protein is produced inside host hemocytes following parasitism. This protein is called EP1, for "early protein 1."

We took great care to establish that EP1 is in fact a polydnaviral protein. EP1 production can be induced in tobacco hornworm larvae by injection of polydnavirus alone, suggesting that the *EP1* gene resides in the genome (the characteristic set of genes) of the virus; alternatively it could reside in the genome of the host and merely be activated by the virus. We deduced part of the sequence of the gene encoding the EP1 protein and then searched for this sequence in various organisms. We turned up no such gene in *Manduca* but instead found the *EP1* gene in the polydnavirus genome. Intriguingly, manufacture of this polydnaviral-encoded protein correlates temporally with the most dramatic effects of parasitism on host hemocytes.

We detected high levels of the EP1 protein inside hemocytes one day after oviposition, when these cells were quite disabled. We continued to find evidence of EP1 in the caterpillar for six days; hemocyte function returned to normal on the eighth day—but too late for the caterpillar to kill the wasp larvae.

Researchers led by Otto Schmidt of the University of Adelaide have found a similar correlation in a different host-parasite pair: hemocytes first become damaged during the brief period when viral protein is produced—then immune response rallies in two or three days. We speculate that hemocyte damage occurs as long as such viral proteins persist; once viral protein levels drop, damaged hemocytes would recover or be replaced, replenishing the functional supply.

One consequence of this timing is that the host immune response resumes in full force before developing wasps are ready to leave the caterpillar. Unlike vulnerable eggs or young larvae, however, older wasp larvae seem able to withstand active immune cells on their own. Polydnavirus provides a long but temporary reprieve from immune attack, allowing the wasps to become mature enough to protect themselves.

Bruce A. Webb of the University of Kentucky, who works with parasitized tobacco budworms, has discovered how the wasps fill one final chink in their immune suppression armor. Although the cellular immune response by the caterpillar is essentially immediate, there is a lag before polydnaviral proteins are available to alter the behavior of host hemocytes. Webb has shown that immediate but short-term protection against immune cells is conferred by ovarian protein molecules injected directly into the host by the wasp. The job of long-term protection then falls to the polydnavirus through sustained viral protein production in caterpillar cells.

Arresting Development

Another important aspect of the *Cotesia-Manduca*-polydnavirus tripartite relationship (and the one that first piqued my interest in the field) is the way the parasite manipulates the development of the host. A growing internal parasite benefits by extending the interval over which its host remains a feeding larva. For this reason, many endoparasites develop strategies to delay host metamorphosis. I was particularly interested by the case of the tobacco hornworm parasitized by *Cotesia,* because this host remains developmentally stunted long after the wasp progeny leave the body cavity; the caterpillars often linger two weeks before dying.

Developmental arrest in lepidopteran hosts is mediated through the endocrine system. I studied endocrine disruption caused by parasitism as a graduate student in Lynn M. Riddiford's laboratory at the University of Washington. There I observed that the concentration of a key hormone regulating metamorphosis is disturbed after parasitism of tobacco hornworms by *C. congregata.* The level of juvenile hormone (JH) is

Enlisting Insects in the War on Weeds

Wasps and viruses are not the only organisms capable of hardball tactics: humans are past masters. And lately scientists have turned caterpillars and their parasites into lethal weapons against weeds. Their worthy opponent is kudzu, a fast-growing, pernicious climber that carpets seven million acres in the southern U.S.

Entomologist David Orr and his colleagues at North Carolina State University have recently deployed soybean looper caterpillars to combat kudzu. In field tests the loopers defoliate the weed, and Orr believes the plants' efforts to replace lost leaves will slowly exhaust their enormous root systems (a single plant can have roots that weigh as much as 300 pounds).

Because soybean loopers eat crops as well as kudzu, each caterpillar enters the field equipped with a safety mechanism to prevent its escape—parasitic wasps that execute the caterpillar as it spins its cocoon, ensuring that no moths will emerge to fly away and reproduce. An added benefit is that the parasitized loopers eat more kudzu: the wasps extend both the feeding interval of the caterpillars and their appetite.

It is not clear how the eggs of this wasp, *Copidosoma truncatellum*, escape attack by the immune system of the caterpillar; the wasp does not carry polydnaviruses. Several wasps that do carry polydnavirus have also been used in biological-control strategies, though, often as weapons against populations of pest insects, including destructive fruit flies, moths and aphids.

—Mia Schmiedeskamp

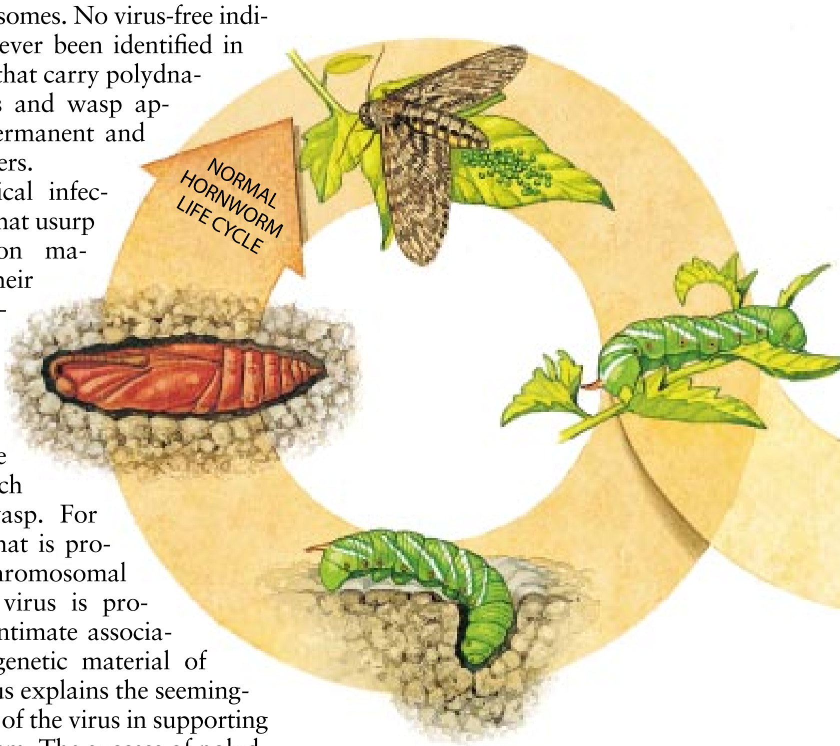
dramatically elevated in parasitized hosts, never descending to the low level needed for pupation. The high levels of JH are probably caused by a lack of sufficient juvenile hormone esterase, an enzyme that clears JH from the organism. Parasitism apparently leads to sustained low esterase levels, in a manner that prevents pupation even after departure of the wasps.

It turns out that developmental arrest is largely executed by polydnavirus, although the presence of the wasp itself is needed to obtain full arrest. When we inject unparasitized larvae with low doses of polydnavirus, they fail to pupate normally. The amount of virus required to retard development appears to be less than that required to suppress the immune response. In fact, postdoctoral fellow Mitch Dushay showed that eggs that have been washed prior to injection retain trace amounts of virus particles or viral proteins that do not suffice to prevent encapsulation by immune cells but may cause developmental failure of the host. Polydnavirus may also contribute to developmental effects even after departure of the wasps from the host. We speculate that the virus remains as a latent infection in the caterpillar—perhaps mediating lasting effects on development.

Polydnavirus is clearly responsible for manipulating a number of developmental and immune programs of the caterpillar host in a manner that is beneficial to the wasp. In many ways, the virus is essential to successful parasitism. The amazing strength of this relationship between wasp and polydnavirus becomes even clearer when one considers the genetics of the partners.

Permanent Partners

The size and complexity of polydnaviral genomes greatly exceeds that of other DNA viruses: each polydnavirus comprises up to 28 separate circles of double-stranded DNA (thus their name, from "polydisperse DNA viruses"). In 1986 Jo-Ann G. W. Fleming and Max D. Summers of Texas A&M discovered that the extremely complex polydnavirus genome is integrated into the genome of both male and female wasps. This viral DNA is thought to be scattered throughout the wasp chromosomes. Wasp inheritance of the virus appears to be strictly Mendelian—the viral sequences are copied and passed to successive generations as part of the wasp chromosomes. No virus-free individuals have ever been identified in wasp species that carry polydnaviruses. Virus and wasp appear to be permanent and integral partners.



Unlike typical infectious viruses that usurp the replication machinery of their hosts to reproduce wildly, polydnavirus reproductive success is affected by the survival of each and every wasp. For every wasp that is produced, a chromosomal copy of the virus is produced. This intimate association of the genetic material of wasp and virus explains the seemingly selfless role of the virus in supporting wasp parasitism. The success of polydnavirus depends on efficient reproduction of the wasp, which in turn depends on an essential host-parasite relationship. Any role the virus plays in ensuring the success of the parasite also ensures the success of viral transfer to the next generation.

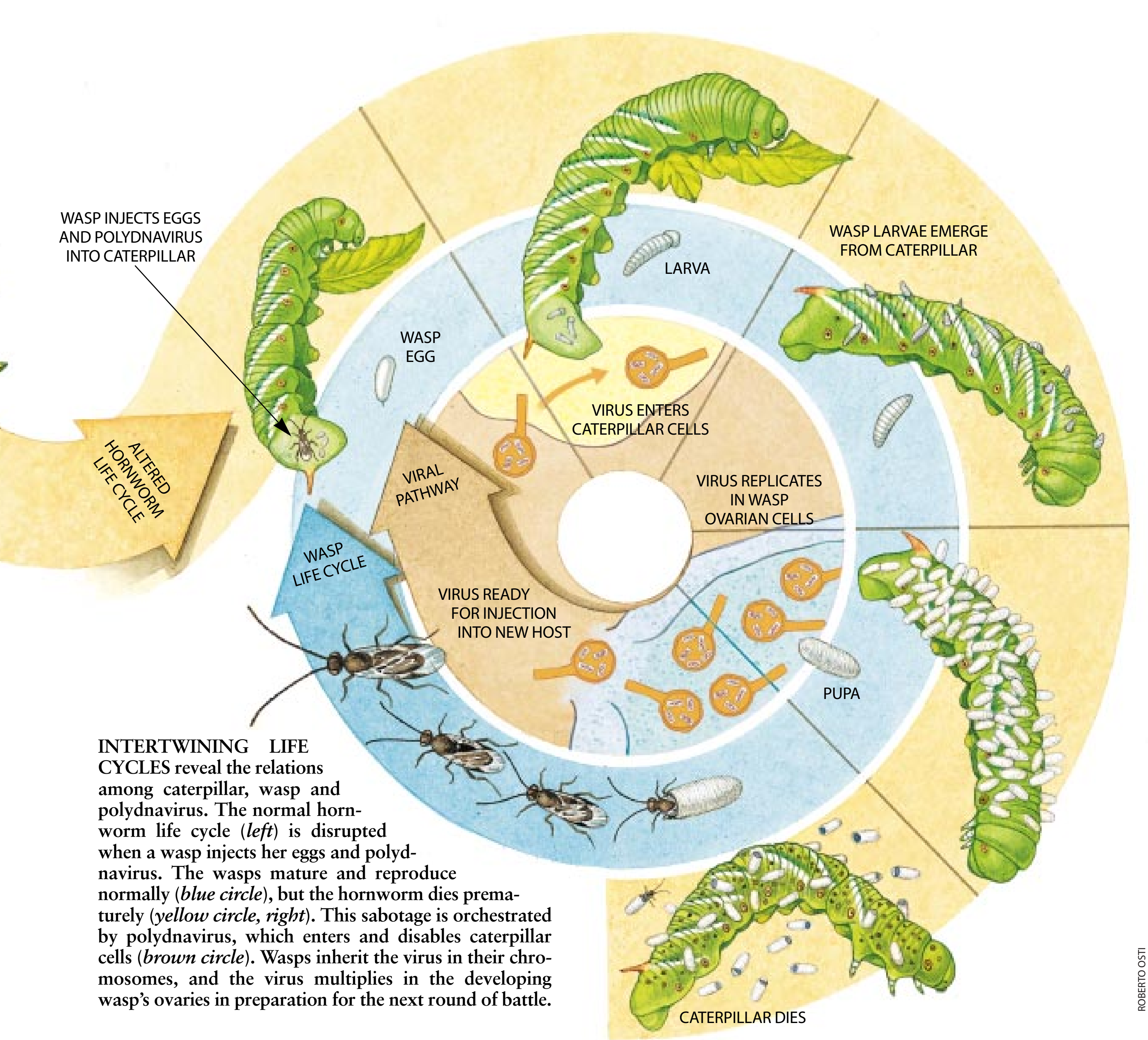
Because viral transmission from wasp to wasp takes place through inheritance of a virus integrated into the chromosome, there must be some other rationale for the mass production of virus in the wasp's ovaries. In fact, the packaged viruses produced at this step seem to be useless for the typical viral mission of infection with intent to replicate—but they are masters of host manipulation. Viruses are expert at spreading throughout a host and entering host cells. Parasitic wasps appear to have harnessed this talent to target useful viral protein production to caterpillar cells, allowing for insider manipulation of their host's biology.

The integration of polydnaviruses into wasp chromosomes prompts questions about the origin of the virus. A typical answer might be that the viruses originated as independent pathogens of the caterpillar hosts, or of the wasps themselves, and later combined with the wasp DNA. A much more intriguing possibility is suggested by the apparently permanent and exclusive association of viral DNA with wasp DNA. Perhaps there was never a separate viral entity. Instead wasps may have acquired the ability to copy and package a subset of useful genes selectively from their own genomes, for shipment into caterpillar cells. Work by Webb and Summers on wasp venom proteins may fit with this last hypothesis.

These researchers have found that some wasp venom genes are similar to polydnaviral genes. Moreover, antivenom antibodies raised in the laboratory also recognize viral proteins that are important for manipulation of the caterpillar. It seems, then, that wasp venom genes and polydnaviral genes may be evolutionarily related. This result is especially exciting because certain wasp venom proteins are known to play a supporting role in manipulating caterpillar physiology.

In one evolutionary scenario, initially independent polydnaviruses may have picked up useful venom genes from the wasp genome. In another scheme, the wasp may have found an incredibly efficient way to utilize its own venom proteins, by copying their genes, packaging them and routing them to caterpillar cells where they can maintain a sustained effect. Either scenario results in increased fitness of both the wasp and virus; either way the genetic boundaries between the wasp and virus are obscured.

Whatever the origins of the polydnaviruses, their associations with wasps and caterpillars offer rich opportunities for the study of evolutionary biology.



INTERTWINING LIFE CYCLES reveal the relations among caterpillar, wasp and polydnavirus. The normal hornworm life cycle (*left*) is disrupted when a wasp injects her eggs and polydnavirus. The wasps mature and reproduce normally (*blue circle*), but the hornworm dies prematurely (*yellow circle, right*). This sabotage is orchestrated by polydnavirus, which enters and disables caterpillar cells (*brown circle*). Wasps inherit the virus in their chromosomes, and the virus multiplies in the developing wasp's ovaries in preparation for the next round of battle.

The vicious tactics of the wasp and its viral accomplice against their hijacked caterpillar host argue against the tenuous hypothesis that the most highly evolved parasites exhibit only minimal virulence to their hosts [see "The Evolution of Virulence," by Paul W. Ewald; SCIENTIFIC AMERICAN, April 1993].

Endoparasitic wasps invariably kill their hosts. The event is exquisitely timed and coordinated, however, to ensure the success of the wasp. In beautiful contrast to these hardball tactics, the mutually advantageous relationship between wasp and virus is so intimate that it blurs interspecies genetic boundaries. The complex question of why the caterpillar does not become a moth should keep a raft of scientists from evolutionary biologists to endocrinologists busy for years to come. SA

The Author

NANCY E. BECKAGE is associate professor of entomology at the University of California, Riverside, where she has been on the faculty since 1990. Her interests include the strategies adopted by parasites and pathogens to disrupt development of host species and the coevolutionary relationships of host organisms with parasites. Previously, Beckage held positions in the U.S. Department of Agriculture Stored Products Insects Research Unit at the University of Wisconsin and at the Seattle Biomedical Research Institute. She received a Ph.D. from the University of Washington in 1980.

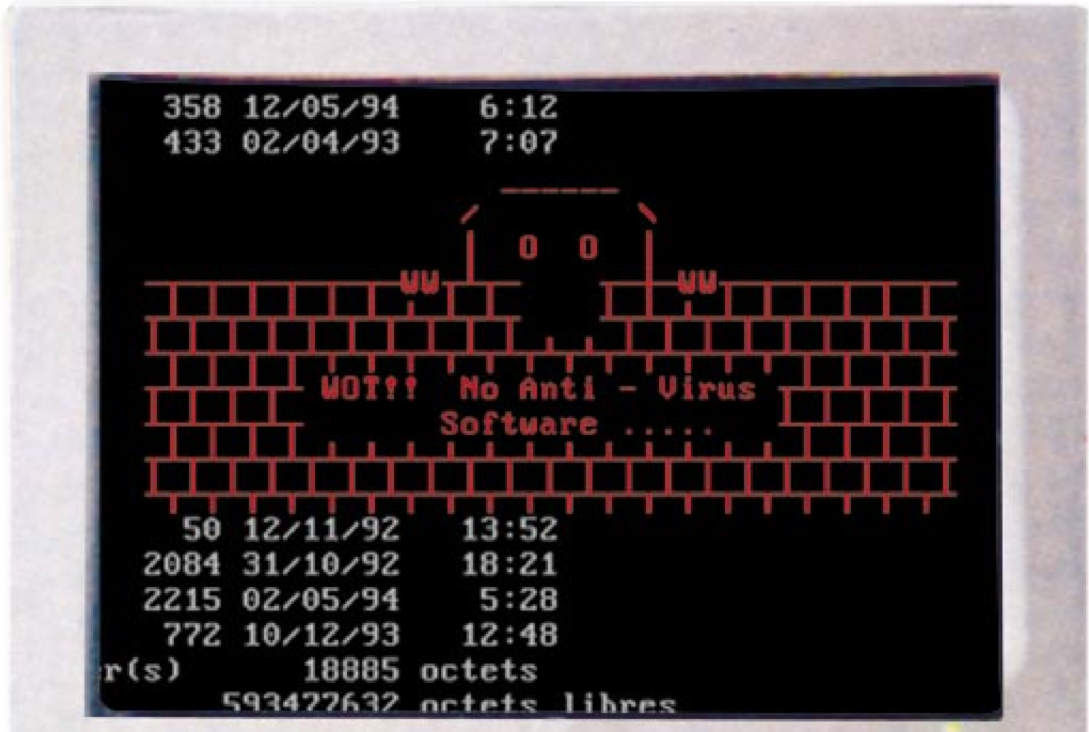
Further Reading

POLYDNAVIRUSES: MUTUALISTS AND PATHOGENS. Jo-Ann G. W. Fleming in *Annual Review of Entomology,* Vol. 37, pages 401–426; 1992.

HOW PARASITIC WASPS FIND THEIR HOSTS. James H. Tumlinson, W. Joe Lewis and Louise E. M. Vet in *Scientific American,* Vol. 268, No. 3, pages 100–106; March 1993.

POLYDNAVIRUSES: POTENT MEDIATORS OF HOST INSECT IMMUNE DYSFUNCTION. M. D. Lavine and N. E. Beckage in *Parasitology Today,* Vol. 11, No. 10, pages 368–378; 1995.

PARASITIC WASPS. Donald L. J. Quicke. Chapman & Hall, 1997.



JANA BRENNING

Fighting Computer Viruses

Biological metaphors offer insight into many aspects of computer viruses and can inspire defenses against them

by Jeffrey O. Kephart, Gregory B. Sorkin, David M. Chess and Steve R. White

Computer viruses have pervaded popular culture at least as successfully as they have the world's computer population. Capitalizing on the same fearful fascination with man-made life-forms that Mary Shelley tapped in *Frankenstein,* viruses have become the subject of widespread urban legends and hoaxes, popular television shows and movies. Yet they have not received much scientific scrutiny.

Much of their popular presence is attributable to an obvious but deep biological analogy: computer viruses replicate by attaching themselves to a host (a program or computer instead of a biological cell) and co-opting the host's resources to make copies of themselves. Symptoms can range from unpleasant to fatal. Computer viruses spread from program to program and computer to computer, much as biological viruses spread within individuals and among individual members of a society. There are other computer pathogens, such as the "worms" that occasionally afflict networks and the "Trojan horses" that put a deceptively friendly face on malicious programs, but viruses are the most common computer ill by far.

We and our colleagues at the IBM Thomas J. Watson Research Center have found the biological analogy to be helpful in understanding the propagation of computer viruses on a global scale and inspirational in our development of defenses against them. Building on decades of research by mathematical epidemiologists, we have obtained some understanding of the factors that govern how quickly viruses spread. Our efforts to find efficient methods of detecting viruses and the relations among them owe much to pattern-matching techniques developed by computational biologists. Furthermore, we have also drawn inspiration for defenses against pathological software from the vertebrate immune system and its astounding ability to repel or destroy pathogens.

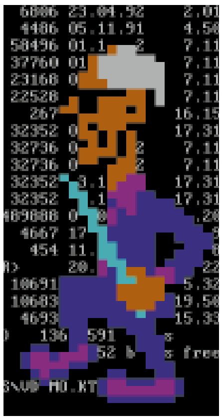
Computer viruses can trace their pedigree to John von Neumann's studies of self-replicating mathematical automata in the 1940s. Although the idea of programs that could infect computers dates to the 1970s, the first well-documented case of a computer virus spreading "in the wild" occurred in October 1987, when a code snippet known as the "Brain" virus appeared on several dozen diskettes at the University of Delaware. Today viruses afflict at least a million computers every year. Users spend several hundred million dollars annually on antivirus products and services, and this figure is growing rapidly.

Most viruses attack personal computers (PCs). More than 10,000 viruses have appeared so far, and unscrupulous programmers generate roughly another six every day. Fortunately, only a handful have been detected far afield.

There are three main classes of PC viruses (and the categories for other systems are analogous): file infectors, boot-sector viruses and macro viruses. Roughly 85 percent of all known viruses infect files containing applications

COMPUTER VIRUSES exist in more than 10,000 strains, a few hundred of which have succeeded in becoming widespread. The screen images shown on these pages are associated with the colorful but unsuccessful "Chad," "Walker" and "Rescue" viruses.

such as spreadsheet programs or games. When a user runs an infected application, the virus code executes first and installs itself independently in the computer's memory so that it can copy itself into subsequent applications that the user runs. Once in place, the virus returns control to the infected application; the user remains unaware of its existence. Eventually a tainted program will make its way to another computer via a shared diskette or network, and the infection cycle will begin anew.



Boot-sector viruses, which account for about 5 percent of known PC virus strains, reside in a special part of a diskette or hard disk that is read into memory and executed when a computer first starts. The boot sector normally contains the program code for loading the rest of a computer's operating system (hence the name, a reference to lifting oneself up by one's own bootstraps). Once loaded, a boot-sector virus can infect any diskette that is placed in the drive. It also infects the hard disk, so that the virus will be loaded into memory whenever the system is restarted. Boot viruses are highly effective: even though there are fewer strains, they were for a time much more prevalent than file infectors were.

The third category, macro viruses, are independent of operating systems and infect files that are usually regarded as data rather than as programs. Many spreadsheet, database and word-processing programs can execute scripts—prescribed sequences of actions—embedded in a document. Such scripts, or macros, are used to automate actions ranging from typing long words to carrying out complicated sequences of calculations. And virus writers have created scripts that insert copies of themselves in other documents. Macro viruses can spread much more rapidly than other kinds of viruses because many people share "data" files freely—consider several workers swapping drafts of a jointly written report. "Concept," the first macro virus observed in the wild, infected its first Microsoft Word document late in 1995 and is now the most prevalent virus in the world. Today more than 1,000 macro viruses are known.

As well as basic replication code, viruses can contain whatever other code the author chooses. Some virus payloads may simply print a message or display an image, but others will damage programs and data. Even those without malicious payloads can cause damage to systems whose configuration differs from what the virus designer expected. For instance, the "Form" virus, which usually produces only a slight clicking noise once a month, overwrites one disk directory sector in a way that is harmless to older PCs but lethal to newer ones that arrange disk information differently.

Antivirus Technology

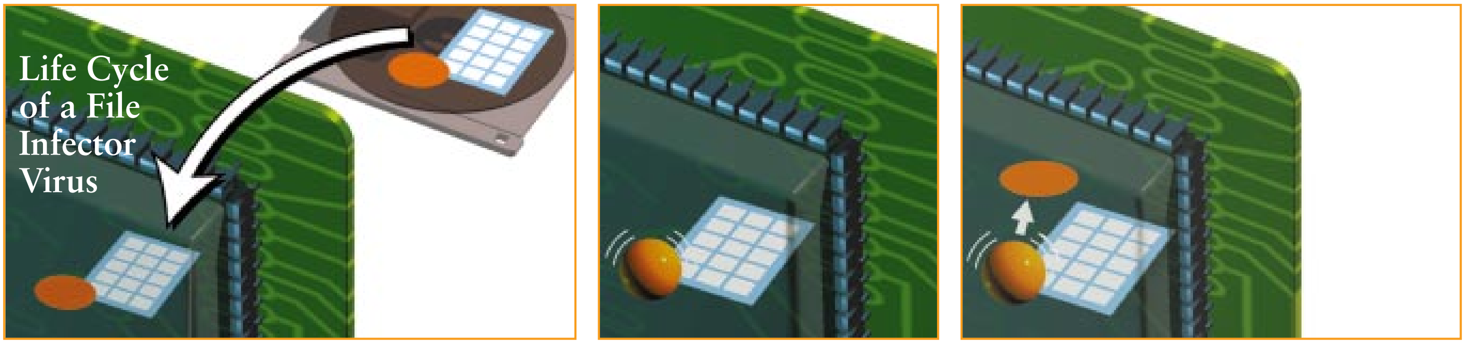
Antivirus software has existed since shortly after computer viruses first appeared. Generic virus-detection programs can monitor a computer system for viruslike behavior (such as modification of certain crucial files or parts of main memory), and they can periodically check programs for suspicious modifications. Such software can even detect hitherto unknown viruses, but it can also be prone to false alarms because some legitimate activities resemble viruses at work.

Scanning programs, in contrast, can search files, boot records and memory for specific patterns of bytes indicative of known viruses. To stay current, they must be updated when new viral strains arise, but they only rarely raise false alarms. The viral signatures these programs recognize are quite short: typically 16 to 30 bytes out of the several thousand that make up a complete virus. (Similarly, biological immune receptors bind to sequences of eight to 15 amino acids out of the thousands in a viral protein.) It is more efficient to recognize a small fragment than to verify the presence of an entire virus, and a single signature may be common to many different viruses. Most computer-virus scanners use pattern-matching algorithms that can scan for many different signatures at the same time: the best can check for 10,000 signatures in 10,000 programs in under 10 minutes.

Once a virus has been detected, it must be removed. One brutal but effective technique is simply to erase the infected program, much as certain types of immune cells destroy an infected cell. Body cells are generally easy to replace, but computer programs and documents are not so expendable. As a result, antivirus programs do their best to repair infected files rather than destroy them. (They are aided in this endeavor by the fact that computer viruses must preserve their host program essentially intact to remain undetected and multiply.)

If a virus-specific scanning program detects an infected file, it can usually follow a detailed prescription, supplied by its programmers, for deleting viral code and reassembling a working copy of the original. There are also generic disinfection techniques that work equally well for known and unknown viruses. One method we developed gathers a mathematical fingerprint for each program on the system. If a program subsequently becomes infected, our method can reconstitute a copy of the original.

Virus-specific detection and removal techniques require detailed analysis of each new virus as it is discovered. Experts must identify unusual sequences of instructions that appear in the viral code but not in conventional programs—a process that relies on carefully developed knowledge and intuition. They also must develop a prescription for verifying and removing the virus from any infected host. To keep up with the influx of half a dozen new viruses a day, antivirus technologists have developed automated tools and procedures to assist human virus experts or even replace them.



When a user runs an infected program, the computer starts by copying the program from the disk, where it is stored and inactive, into RAM, where it can be executed.

The viral code begins running first, while the infected program is still quiescent.

The virus copies itself in a part of RAM separate from the program so that it can continue its work even after the user starts running other software.

We have developed a brute-force statistical technique to extract high-quality signatures very quickly. We started by measuring the frequencies of short byte sequences in a large group of legitimate programs. When a new virus is sent to us, our software finds the sequence of viral bytes that is statistically least likely to appear in a legitimate program. This method is much faster than analysis by hand, and tests suggest that it produces signatures that are less prone to false alarms than those selected by expert humans. Our signature-extraction method is somewhat analogous to the outmoded "template" theory of the immune system, according to which antibodies mold themselves to a particular foreign invader—our signatures are made specifically for each new virus we encounter.

Stephanie Forrest of the University of New Mexico and her collaborators at Los Alamos National Laboratory have developed an alternative that is more faithful to the currently accepted "clonal selection" theory of the immune system, in which the body generates an enormous range of immune cells and then mass-produces the ones that turn out to recognize a pathogen. Their scheme generates code signatures randomly, without reference to any particular virus. Each signature is checked against existing code on the system; if it does not match anything, it is retained in a huge database. Finding one of these signatures in a program is a sure sign that the program has been modified, although further analysis is required to determine whether a virus is at fault.

In another twist on the biological metaphor, virus hunters have learned to exploit the fact that programmers often make new computer viruses from key parts of existing ones. These viral "genes" enable us to trace the evolutionary history of computer viruses, in the same way that biologists determine the family trees of related species. By processing large collections of viral code, we can automatically derive a set of family signatures that catches all the different members of a viral family, including previously unknown variants. This technique reduces signature storage requirements substantially: a single 20-byte family signature can recognize dozens of distinct viruses.

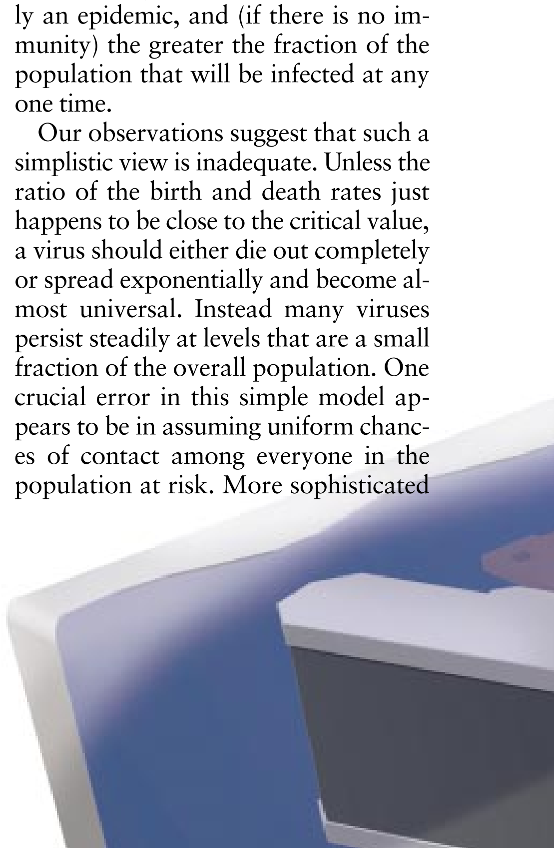
We have also developed a neural-network technique to recognize viruses by scanning for several, very short patterns, each only three to five bytes long. These tiny fragments represent computer instructions that carry out tasks specific to viral infection. Although conventional software might occasionally contain one of these fragments, the presence of many of them is an almost certain viral hallmark. Antiviral software can check for such short sequences very quickly; even more important, because these patterns of data are directly linked to the virus's function, we can now recognize a wide variety of viruses without ever having seen them before.

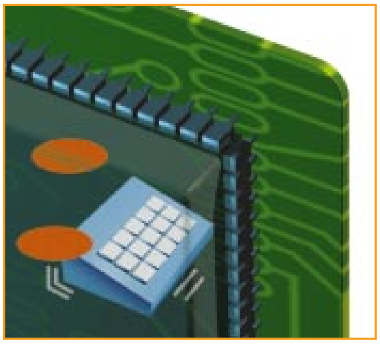
Hunting Viruses in the Wild

Since 1990 we have been collecting virus statistics from a population of several hundred thousand PCs among our corporate customers. We record the location and date of each incident along with the number of infected PCs and diskettes and the identity of the virus. These statistics have permitted us to infer a good deal about the behavior of viruses in the wild, including the fact that only a small fraction of viruses are genuinely problematic. Only about 5 percent of all known viruses have been observed within the population we have studied, many of them just once. The 10 most common viruses account for two thirds of all incidents. In addition, the prevalence of these successful viruses appears to follow a common pattern: a virus will spread over the course of a year or so, increasing its numbers in a roughly linear fashion until it reaches a plateau. After that, it will continue to appear in computers at a roughly constant level, although sometimes its numbers decline to near extinction.

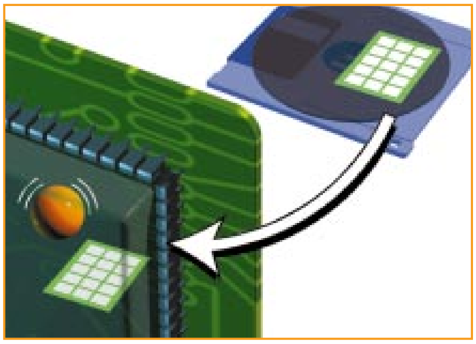
In an effort to understand these characteristics, we have borrowed from mathematical models of biological epidemics. The simplest models predict the behavior of a disease from a few parameters—most significantly, the "birth rate" at which sick individuals infect others and the "death rate" at which the sick either die or are cured. If the ratio between these two rates is less than a critical value, any infection will quickly die out. The larger the ratio, the more likely an epidemic, and (if there is no immunity) the greater the fraction of the population that will be infected at any one time.

Our observations suggest that such a simplistic view is inadequate. Unless the ratio of the birth and death rates just happens to be close to the critical value, a virus should either die out completely or spread exponentially and become almost universal. Instead many viruses persist steadily at levels that are a small fraction of the overall population. One crucial error in this simple model appears to be in assuming uniform chances of contact among everyone in the population at risk. More sophisticated

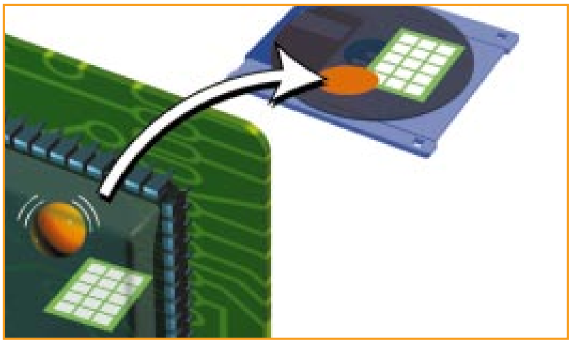




Its initial work done, the virus passes control back to the infected program.



When the user runs a different program, the dormant virus begins running again.



It inserts a copy of itself into the previously uninfected software so that the cycle of virulence can repeat.

JAMES GARY

models take into account the extraordinary cliquishness of typical patterns of software exchange. Each person shares software and data only with a few other people, on average, and most of the sharing takes place within groups. If Alice shares with Bob and Bob shares with Carol, then Alice and Carol are reasonably likely to share with each other.

Computer simulations have shown that locality of contact slows the initial growth in a way that is qualitatively consistent with our observations. Sparse sharing reduces the likelihood of an epidemic and lowers the plateau, but not by enough to explain the data.

Evolution in Action

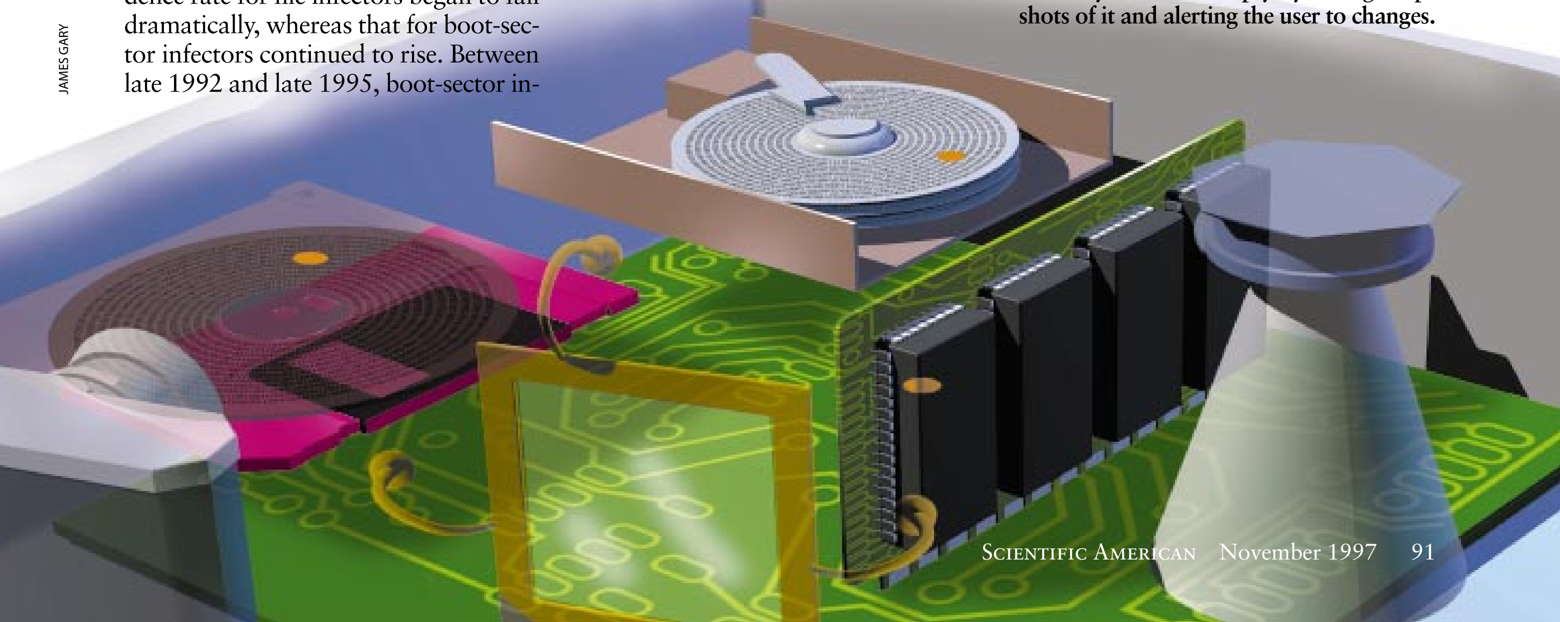
Just as external factors such as drought, sanitation and migration have a strong influence on biological epidemics, changes in the computing environment are responsible for the presence of several distinct epochs in viral infection. Until 1992, reported sightings of file-infecting viruses and boot viruses occurred at roughly equal (and steadily rising) rates. Then the incidence rate for file infectors began to fall dramatically, whereas that for boot-sector infectors continued to rise. Between late 1992 and late 1995, boot-sector infectors reigned supreme. Why did the file infectors essentially become extinct?

We believe the cause was the widespread acceptance of Windows 3.1, an enhancement to MS-DOS—the operating system used on most computers—that became popular around 1992. Windows crashes readily in the presence of typical file viruses, and so necessity will lead afflicted users somehow to eliminate the virus from their systems (perhaps by wiping out the hard disk and reinstalling all the software), regardless of whether they know that the symptoms are caused by a virus. Boot viruses, in contrast, tend to coexist peacefully with Windows 3.1; they do not kill their hosts before the infection has a chance to run riot.

The wide use of Windows 95, yet another new operating system, has now led to a precipitous decline in the prevalence of boot viruses. Windows 95 warns the user about most changes to boot sectors, including many of those caused by viruses, and most boot viruses cannot spread under Windows 95. We have already seen a handful of viruses specifically designed for Windows 95 and other 32-bit operating systems, although the ones we have seen are unlikely to become widespread.

We are now in the era of the macro virus. Because users tend to exchange documents and other data files capable of harboring macro viruses more frequently than they exchange programs, macro viruses enjoy a higher birth rate and thus spread faster than the traditional boot or file infectors. Sophisticated mail and file-transfer functions now permit users to share documents or programs more quickly and easily than before, exacerbating the problem.

Macro viruses are also the first viruses to exploit the growing trend for interoperability among computers. A



DIGITAL BATTLEGROUND holds both malicious programs and the specialized software that hunts them down and repairs the damage they cause. Computer viruses come in many varieties and can be recognized both by their actions and by unique patterns of program code within them. Antiviral programs can also protect ordinary software simply by taking snapshots of it and alerting the user to changes.

JAMES GARY

DOS file infector can never endanger a Macintosh, for instance, but a macro virus can infect any computer that supports a vulnerable application program. The fact that Microsoft Word runs on many different kinds of computers enables Concept and other macro viruses to move beyond traditional system boundaries.

A Digital Immune System

Today viruses mainly travel from one computer to another through intentional, manual exchange of programs, and human response time is generally sufficient to cope with them. A successful new virus typically takes months or even years to gain a foothold. In the densely connected world of the near future, viruses might be able to propagate much faster. As early as 1988, Robert Tappan Morris launched what came to be known as the "Internet Worm," a program that exploited security holes and invaded hundreds of computers around the world in less than a day.

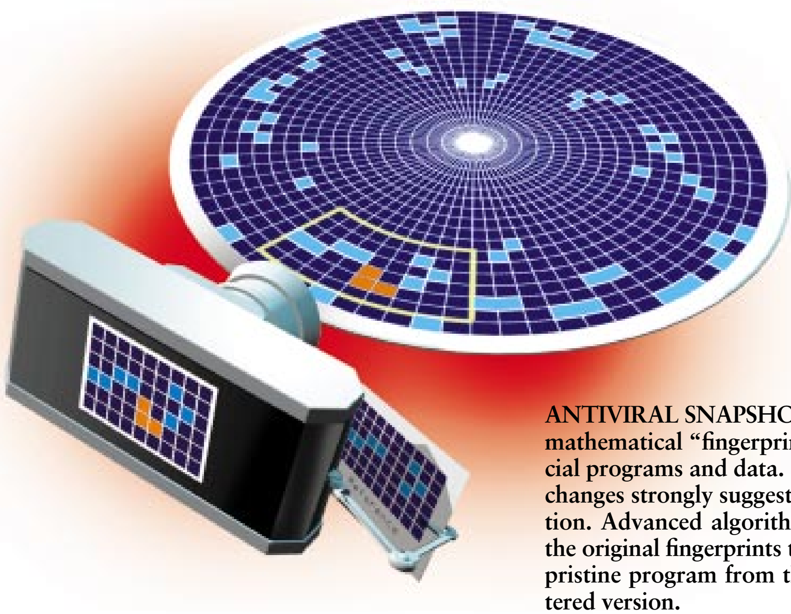
New technologies (such as Web browsers that use "ActiveX") for silently downloading software and data to a user's computer make the problem even more pressing. Already modern-day mail programs permit text documents or spreadsheets to be sent very simply as e-mail attachments. Opening the attachment can cause the appropriate application to start up automatically, and any macro viruses contained in the attachment may be executed. Soon software agents may be routinely authorized to send and open mail containing attachments. With humans no longer participating in the replication cycle, viruses could be free to spread orders of magnitude faster than they do now.

These changes in the digital ecosystem suggest that a more automatic response to computer viruses is needed, one that is not limited by human response times or by the rate at which humans can dissect novel viruses. IBM, Symantec Corporation and McAfee Associates are among the companies developing technology to help respond quickly and automatically to new viruses.

At IBM, we are creating what may be thought of as an immune system for cyberspace. Just as the vertebrate immune system creates immune cells capable of fighting new pathogens within a few days of exposure, a computer immune system derives prescriptions for recognizing and removing newly encountered

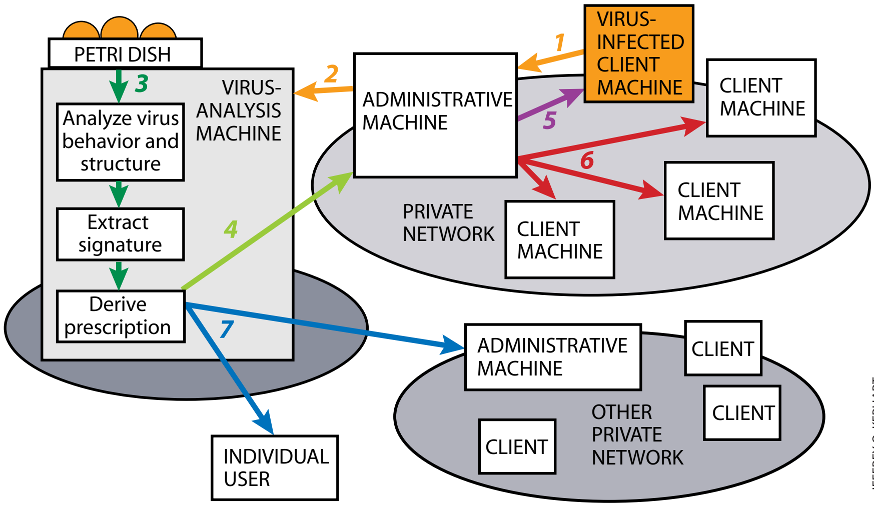
GENERIC ANTIVIRAL PROGRAM flags activities—such as the alteration of critical sites in RAM or particular files on disk—that are likely to arise from a virus in action. Preventing these illicit acts will not eliminate the virus but can stop it from infecting additional programs or interfering with the computer's normal operation.

SIGNATURE SCANNER searches a user's disks looking for fragments of program code that appear in known viruses.



JAMES GARY

ANTIVIRAL SNAPSHOTS capture mathematical "fingerprints" of crucial programs and data. Subsequent changes strongly suggest viral infection. Advanced algorithms can use the original fingerprints to recover a pristine program from the virus-altered version.



JEFFREY O. KEPHART

DIGITAL IMMUNE SYSTEM for cyberspace is planned to work as shown here. An unknown virus spurs a client machine to forward a sample to an administrative machine (*1*), which in turn sends an encrypted sample to a central virus-analysis machine (*2*). This machine replicates the virus in a "petri dish" device and analyzes its behavior and structure (*3*). The resulting prescription is sent back to the administrative machine (*4*), which forwards it first to the infected client (*5*) and then to other machines on the local network (*6*). Subscribers around the world receive regular antivirus updates that protect them from the new virus (*7*).

computer viruses within minutes. In a current prototype, PCs running IBM AntiVirus are connected by a network to a central computer that analyzes viruses. A monitoring program on each PC uses a variety of heuristics based on system behavior, suspicious changes to programs, or family signatures to infer that a virus may be present. The monitoring program makes a copy of any program thought to be infected and sends it over the network to the virus-analysis machine.

On receiving a putatively infected sample, the machine sends it to another computer that acts as a digital petri dish. Software on this test machine lures the virus into infecting specially designed "decoy" programs by executing, writing to, copying and otherwise manipulating the decoys. To replicate successfully, a virus must infect programs that are used often, and so the decoy activity brings the viral code out of hiding. Other behavioral characteristics of the virus can be inferred during this phase as well.

Any decoys that have been infected can now be analyzed by other components of the immune system, which will extract viral signatures and produce prescriptions for verifying and removing the virus. Typically it takes the virus analyzer less than five minutes to produce such prescriptions from an infected sample. The analysis machine sends this information back to the infected client PC, which incorporates it into a permanent database of cures for known viruses. The PC is then directed to locate and remove all instances of the virus, and it is permanently protected from subsequent encounters.

If the PC is connected to other machines on a local-area network, it is quite possible that the virus has invaded some of them as well. In our prototype, the new prescription is sent automatically to neighboring machines on the network, and each machine checks itself immediately. Because computer viruses can exploit the network to multiply quickly, it seems fitting that the antidote should use a similar strategy to spread to machines that need it. By allowing the latest prescriptions to be propagated to subscribers at uninfected sites, it is possible in principle to immunize the entire PC world against an emerging virus very rapidly.

Regardless of how sophisticated antivirus technology may become, computer viruses will forever remain in an uneasy coexistence with us and our computers. Individual strains will wax and wane, but as a whole, computer viruses and antivirus technology will coevolve much as biological parasites and hosts do. Both will also evolve in response to such changes in the computing environment as itinerant software agents—which will have to be protected from corruption by the computer systems they traverse even as those systems guard themselves from agent malice. Perhaps computer viruses and computer immune systems are merely precursors of an eventual rich ecosystem of artificial lifeforms that will live, die, cooperate and prey on one another in cyberspace. SA

The Authors

JEFFREY O. KEPHART, GREGORY B. SORKIN, DAVID M. CHESS and STEVE R. WHITE, all currently at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y., come to the study of computer viruses from disparate backgrounds. Kephart studied electrical engineering and physics at Princeton and Stanford universities and then became interested in analogies between large-scale computer systems, ecosystems and economies while doing postdoctoral work at Xerox PARC. Sorkin, who earned an A.B. in mathematics from Harvard University and a Ph.D. in computer science from the University of California, Berkeley, specializes in combinatorics, mathematical modeling and optimization. Chess tackled many problems in computer-performance tuning, cooperative processing and security before turning to the implications of self-replicating code for distributed systems. He has an A.B. in philosophy from Princeton and an M.S. in computer science from Pace University. White received a Ph.D. in theoretical physics from the University of California, San Diego. He conceived IBM's antiviral products and continues to lead their research and development.

Further Reading

ROGUE PROGRAMS: VIRUSES, WORMS AND TROJAN HORSES. Edited by Lance J. Hoffman. Van Nostrand Reinhold, 1990.

COMPUTERS AND EPIDEMIOLOGY. J. O. Kephart, S. R. White and D. M. Chess in *IEEE Spectrum,* Vol. 30, No. 5, pages 20–26; May 1993.

A SHORT COURSE ON COMPUTER VIRUSES. Second edition. Frederick B. Cohen. John Wiley & Sons, 1994.

ROBERT SLADE'S GUIDE TO COMPUTER VIRUSES. Robert Slade. Springer-Verlag, 1994.

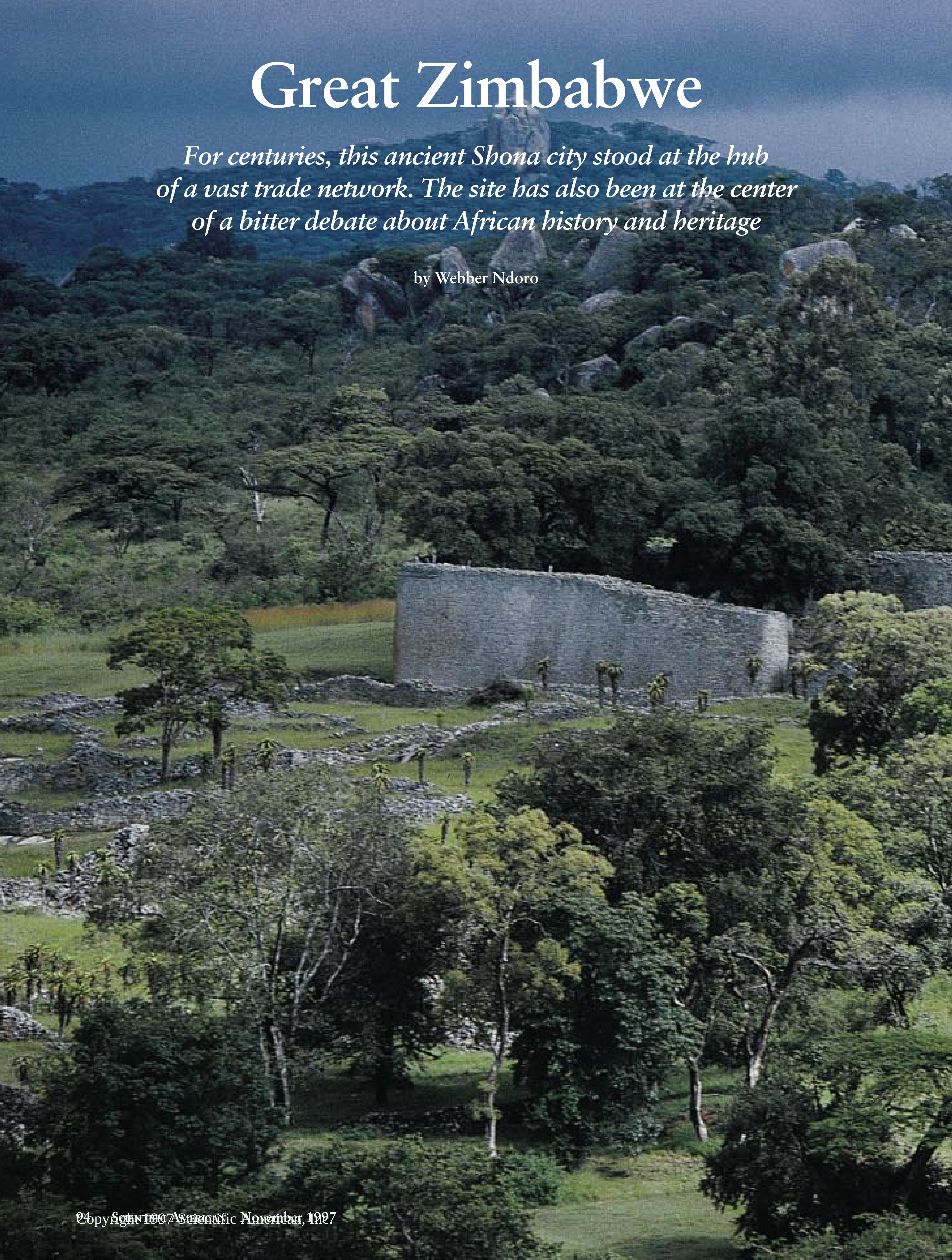
BIOLOGICALLY INSPIRED DEFENSES AGAINST COMPUTER VIRUSES. Jeffrey O. Kephart, Gregory B. Sorkin, William C. Arnold, David M. Chess, Gerald J. Tesauro and Steve R. White *in Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, August 20–25, 1995. Distributed by Morgan Kaufmann Publishers, Inc.

ANTIVIRUS ONLINE. IBM site on virus information is available at http://www.av.ibm.com on the World Wide Web.

Great Zimbabwe

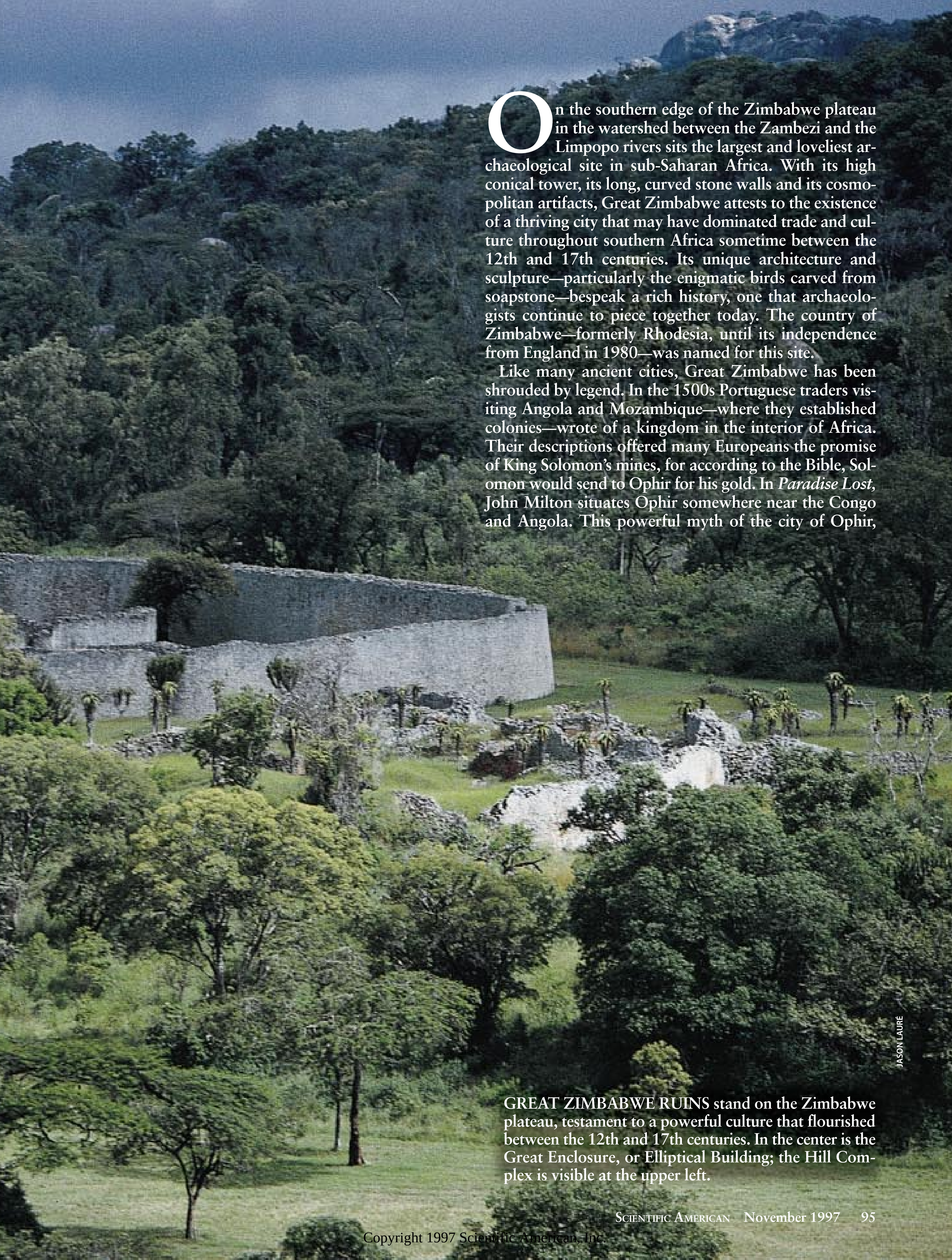
For centuries, this ancient Shona city stood at the hub of a vast trade network. The site has also been at the center of a bitter debate about African history and heritage

by Webber Ndoro



On the southern edge of the Zimbabwe plateau in the watershed between the Zambezi and the Limpopo rivers sits the largest and loveliest archaeological site in sub-Saharan Africa. With its high conical tower, its long, curved stone walls and its cosmopolitan artifacts, Great Zimbabwe attests to the existence of a thriving city that may have dominated trade and culture throughout southern Africa sometime between the 12th and 17th centuries. Its unique architecture and sculpture—particularly the enigmatic birds carved from soapstone—bespeak a rich history, one that archaeologists continue to piece together today. The country of Zimbabwe—formerly Rhodesia, until its independence from England in 1980—was named for this site.

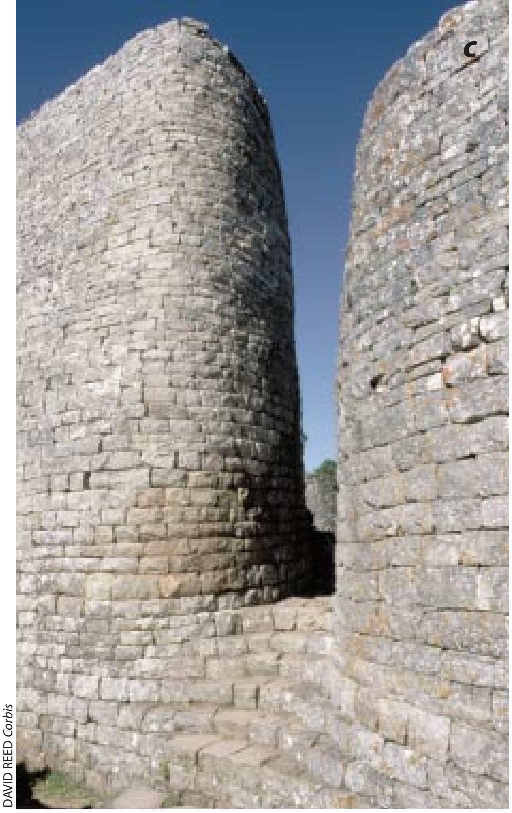
Like many ancient cities, Great Zimbabwe has been shrouded by legend. In the 1500s Portuguese traders visiting Angola and Mozambique—where they established colonies—wrote of a kingdom in the interior of Africa. Their descriptions offered many Europeans the promise of King Solomon's mines, for according to the Bible, Solomon would send to Ophir for his gold. In *Paradise Lost,* John Milton situates Ophir somewhere near the Congo and Angola. This powerful myth of the city of Ophir,



JASON LAURÉ

GREAT ZIMBABWE RUINS stand on the Zimbabwe plateau, testament to a powerful culture that flourished between the 12th and 17th centuries. In the center is the Great Enclosure, or Elliptical Building; the Hill Complex is visible at the upper left.

ROBERT HOLMES



DAVID REED *Corbis*

ROBERT HOLMES *Corbis*

CHARLES AND JOSETTE LENARS *Corbis*

SOPHISTICATED STONEWORK and curved walls are hallmarks of Great Zimbabwe, as seen in the conical tower (*a*), the elegant rounded steps (*b* and *c*), the chevron pattern on the walls of the Great Enclosure (*d*) and the inner passage of the Great Enclosure (*e*). Other features, such as the stone protrusions in this courtyard (*f*), remain entirely mysterious.

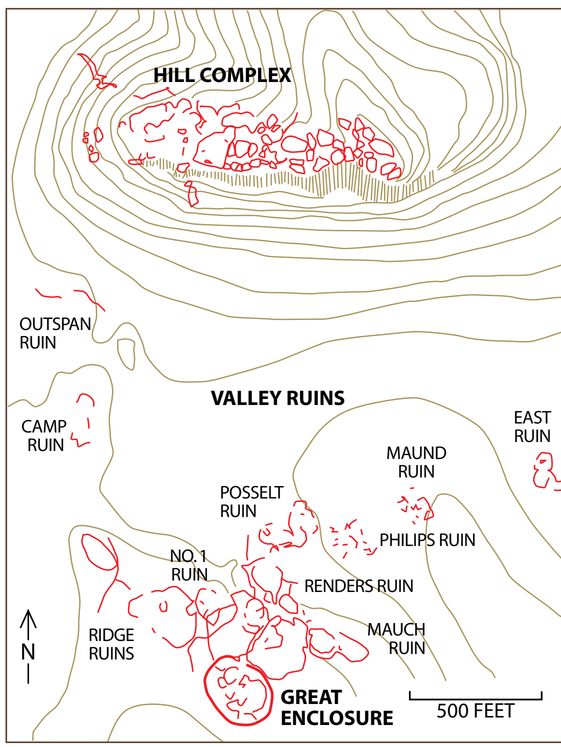
populated by Semitic people, shaped the later cultural and historical interpretations of Great Zimbabwe. The fable is, in large part, the reason so many archaeological mysteries remain about the site. Because whereas the story of Great Zimbabwe is ultimately that of early Shona culture and the African Iron Age, it is also a tale of colonialism and of often shoddy, politically motivated archaeology.

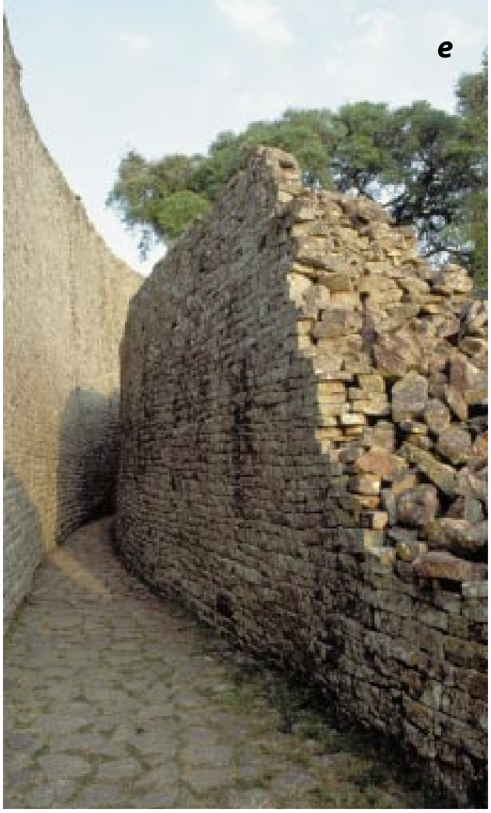
Masterful Stonework

Constructed between A.D. 1100 and 1600, Great Zimbabwe seems not to have been designed around a central plan but rather to have been altered to fit its changing role and population. Its scale is far larger than that of similar regional sites—including Danamombe, Khami, Naletale, Domboshava (in northern Botswana), Manikweni (in Mozambique) and Thulamela (in northern South Africa)—suggesting that Great Zimbabwe was the area's economic and political center. Because it is situated on the shortest route between the northern gold fields, where inland rivers were panned for the precious metal, and the Indian Ocean, the rulers of Great Zimbabwe most likely regulated the thriving medieval gold trade.

Great Zimbabwe covers 1,779 acres and comprises three main structures: the Hill Complex, the Great Enclosure and the smaller Valley Ruins. The Hill Complex, dubbed the Acropolis by Europeans, forms the oldest part of the site; evidence hints that farmers or hunters may have encamped there as early as the fifth century. From its position on the rocky, 262-foot-high hill, the oval enclosure—about 328 feet long and 148 feet wide—would have allowed its inhabitants to see potential invaders. The outer wall, which stands nearly 37 feet high, would also have afforded good protection. Inside the walls, as inside all the other enclosures, stand *daga* houses, curved, hutlike structures made of Africa's most common building material: dried earth, mud and gravel.

Below the Hill Complex sits the most stunning of Great Zimbabwe's structures, the Great Enclosure, or Elliptical Building. Called *Imbahuru,* meaning "the house of the great woman" or "the great house," by the Karanga-speaking people who lived there during the 19th century, the Great Enclosure was built at the height of Great Zimbabwe's power. (Karanga is the most common dialect of Shona and is spoken by the inhabitants of south-central Zimbabwe.) The enclosing wall is 800 feet long and stands 32 feet high at some places; an estimated one million blocks were used in its construction. An inner wall runs along part of the outer wall, creating





CORDAIY PHOTO LIBRARY LTD. *Corbis*

ROBERT HOLMES *Corbis*

a narrow, 180-foot-long passageway.

The function of the Great Enclosure is not known, although it is thought to have served as a royal palace. Because of the presence of grooves in the walls (perhaps representing the female anatomy) and of phallic structures, some historians have postulated that the compound was used for adolescent initiation rites or for other important ceremonies. It may have also housed the many wives of the ruler. The great conical tower, which stands 30 feet high and is 18 feet in diameter at the base, appears not to have been used for any particular purpose and may have served a merely symbolic function.

JOHNNY JOHNSON

ZIMBABWE, formerly Rhodesia, lies in southeastern Africa and has some 35,000 registered archaeological sites. The ruins of Great Zimbabwe are by far the largest of these, covering 1,779 acres.

In addition to the Hill Complex and the Great Enclosure, Great Zimbabwe is made up of the smaller Valley Ruins. This series of compounds stands in the valley between the two larger structures. The walls seem to be youngest here, suggesting that these structures were built as the population expanded and Great Zimbabwe needed more residential space.

Great Zimbabwe is unusual not only in its size but in its stonework. Many of the structures are made of rectangular blocks cut from nearby granite outcroppings. The city's name derives from the Shona term *dzimbabwe,* meaning "houses of stone." The blocks, set in layers without mortar, form stable freestanding, curved walls that are often about twice as high as they are wide. Although round, buttresslike structures rest along the base of many walls, they have no supportive role. Some archaeologists speculate that these curved extensions may have served to soften the approach to a doorway, or to have made passageways more complicated to navigate or perhaps even to have hidden rooms from direct view. They also may have served to control access to some areas, because people could have moved into the area in single file only.

The stonework is, in certain places, astonishingly sophisticated: rounded steps grace some of the entrances, and chevron designs decorate some of the walls. The walls are also punctuated by drains and occasionally by four-foot-wide doorways, some of which had wood lintels.

A Mysterious Culture

Although we know very little about the people of Great Zimbabwe, we can guess something about them from our knowledge of Mapungubwe sites, which appear to have been the center of Shona civilization around A.D. 1000. The largest Mapungubwe settlements, found in the Shashi-Limpopo area, are very similar to Great Zimbabwe. Wealth was apparently based on cattle production, ivory trade and gold. The Mapungubwe culture spread into western parts of Zimbabwe as the presence of Leopard's Kopje pottery (in Mapungubwe style) attests. With the rise of Great Zimbabwe, it appears that trade shifted and Mapungubwe declined as an important center, becoming abandoned just as Great Zimbabwe prospered.

Artifacts unearthed at Great Zimbabwe have not clarified the social and cultural organization of the settlement, but they have distinguished it from other Iron Age sites. In particular, a group of soapstone birds, many of them 14 inches high and sitting atop three-foot-tall columns, is unlike any sculpture found elsewhere. Each bird has a different pattern or marking; none is identifiable as a local creature. Because of the regard contemporary Shona people hold for their dead and because some Shona tribes use iron rods to mark tallies of their dead, some archaeologists have speculated that the avian icons indicate aggregates of ancestors used in rituals.

Other artifacts indicate that Great Zimbabwe was well established as a trading community by the 14th century. Objects from distant lands made their way to Great Zimbabwe: Syrian glass, Chinese celadon dishes (mostly from the Ming Dynasty, A.D. 1368 to 1644), Persian faience bowls, coral, bronze bells and an iron spoon—a utensil not used by the Shona. There is no blue-and-white Chinese porcelain, which became widespread during the mid-15th century; its absence suggests that Great Zimbabwe's economic importance was less by that time. Indeed, it does appear that the site was largely empty by 1700.

There are several reasons Great Zimbabwe may have been abandoned. By the late 1600s the northern rivers had been panned clean, and the gold trade began to move west. No longer centrally located, the city may not have been able to thrive when revenue and trade dried up. Another possibility is that the population became unsustainable. By some estimates, Great Zimbabwe had between 10,000 and 17,000 residents at its peak—a population equivalent to that of medieval London. (Other estimates are more conservative, placing the populace at a maximum of 2,000.) The area may have become devegetated as huge herds of cattle grazed it or as it was extensively farmed; recent environmental data suggest that a succession of severe droughts caused people to disperse. Or there may have been some other impetus, such as war, although there is no evidence besides minimal weaponry to support this argument. More archaeological clues, further digs at Great Zimbabwe and excavations at other Iron Age sites are needed to resolve the question of decline.

ROBERT HOLMES *Corbis*

SHONA BIRDS, often placed atop pillars, have been found only at Great Zimbabwe. These soapstone carvings do not resemble any local species; archaeologists and historians have been unable to determine their significance.

Plunder and Misappropriation

Empty for 200 years or so, Great Zimbabwe was probably used only irregularly for religious ceremonies—as it is today—until the late 1800s. It was then that Europeans arrived, lured by visions of gold from King Solomon's mines, and it was then that the archaeological record became so damaged as to become largely indecipherable.

A German explorer, Karl Mauch, was first to arrive, in 1871. He befriended another German, Adam Render, who was living in the tribe of Chief Pika, a Karanga leader, and who led him to Great Zimbabwe. (Had he known the outcome, Render, who was married to two tribeswomen and well integrated, might have steered Mauch into the Zambezi River.) On seeing the ruins, Mauch concluded very quickly that Great Zimbabwe, whether or not it was Ophir, was most certainly not the handiwork of Africans. The stonework was too sophisticated, the culture too advanced. It looked to Mauch to be the result of Phoenician or Israelite settlers. A sample of wood from a lintel bolstered Mauch's rapid assessment: it smelled like his pencil, therefore it was cedar and must have come from Lebanon.

Mauch's visit was followed by one from Willi Posselt, a looter, who lugged off a carved soapstone bird and hid others so he could return for them later. Posselt was followed by a series of visitors, some of whom worked for W. G. Neal of the Ancient Ruins Company, which had been created in 1895. Cecil Rhodes, founder of the British South Africa Company, gave Neal a commission to exploit all Rhodesian ruins. Neal and his rogues pillaged Great Zimbabwe, and other Iron Age sites, taking gold and everything of value, tearing down structures and throwing away whatever was not valuable to them (pottery shards, pots, clay figurines).

The first official archaeologist to visit the site, James Theodore Bent from Britain, had added to the confusion in 1891 by digging around the conical tower in the Great Enclosure—thereby completely destroying the stratigraphy and making it impossible for later archaeologists to make sense of its age. Bent also threw away clay and metal artifacts, including Persian and Arab trade beads, as insignificant. The archaeologist concluded that Great Zimbabwe had been built by a local "bastard" race—bastards because their fathers must have been white invaders from the north—since, as Rhodes and most European settlers maintained, native Africans could never have constructed Great Zimbabwe themselves.

A 1902 report written by Neal and a journalist named Richard N. Hall reiterated Bent's conclusions: the architecture was clearly Phoenician or Arabian. This attitude was pervasive in colonialist Africa: the continent had no history, no sophistication; its people and tribes were unchanging, unable to develop, culturally barren.

Archaeologists who suggested otherwise were not well received. In 1905 David Randall-MacIver, an Egyptologist who had studied under the famous William Matthew Flinders Petrie, excavated at the site and uncovered artifacts very similar to the ones being used by Shona, or Karanga, people living in the vicinity. By turning to indigenous people for cultural clues and interpretation rather than just for labor, Randall-MacIver was indeed doing something unprecedented. Had any other investigators of the time drawn on the lore or knowledge of the local people, many of the questions about Great Zimbabwe might well have been answered.

PAUL ALMASY Corbis

GREAT ENCLOSURE, shown here from the opposite side of the photograph on pages 94 and 95, perhaps served as the royal quarters of Great Zimbabwe. The structure was built during the city's heyday; its surrounding wall is 800 feet long and comprises an estimated one million stone blocks.

The continuity of artifacts suggested to Randall-MacIver that the site had been built by people whose culture was similar. He also demonstrated that the Arab and Persian beads were no older than 14th or 15th century and thus did not date back to biblical times and King Solomon. And he argued that the stonework was not at all Arabic, because it was curved and not arranged in geometric or symmetrical patterns. Randall-MacIver concluded that native Africans had built Great Zimbabwe.

Two subsequent researchers held the same opinion. In 1926 J. F. Schofield reiterated Randall-MacIver's conclusions, and in 1929 Gertrude Caton-Thompson did the same. Her excavations of the undisturbed Maund Ruin—which lies at the opposite end of the valley from the Great Enclosure—again supported the theory of indigenous construction. Caton-Thompson's detailed drawings and careful stratigraphy have been crucial in piecing together what little is known about Great Zimbabwe.

Despite the mounting evidence and archaeological testimony, most European settlers in Rhodesia rejected the record. From 1965 until independence in 1980, the Rhodesian Front censored all books and other materials available on Great Zimbabwe. This party, established by then prime minister Ian Smith to prevent Africans from gaining power, was based on a system of apartheid. Archaeologists, such as the noted Peter S. Garlake, who were vocal about the native origin of Great Zimbabwe were imprisoned and eventually deported. Africans who took the same view lost their jobs. Displays at the site itself were censored as well, although it hardly mattered because they were in English, and locals were not allowed to use the premises for any ceremonies.

Reclaiming the Past

Today Great Zimbabwe is a symbol of African cultural development. Popular books have made the monument somewhat more accessible to the people of Zimbabwe. Yet, at the same time, Great Zimbabwe remains largely inaccessible. Because of past archaeological mistakes, much of the history of the site is elusive. Given the condition of contemporary archaeology in southern Africa, there is little chance this will change soon.

The two archaeologists currently stationed at the site are responsible not only for the preservation of the decaying monument but for dealing with visitors and maintenance—and the 5,000 other sites under their jurisdiction as well (out of a total of 35,000 recorded sites in Zimbabwe). Although the ruins are protected by the National Museums and Monuments of Zimbabwe and were designated a World Heritage Site by UNESCO, only two conservators and fewer than 10 archaeologists are available in Zimbabwe to study and look after all the archaeological sites, including Great Zimbabwe.

The situation in other sub-Saharan countries is no better. According to Pierre de Maret of the Free University of Brussels, less than $150,000 is spent annually on archaeology in 10 sub-Saharan countries—and there are a mere 20 professional archaeologists among them. The sale of African objects abroad however, reaches into the millions of dollars every year.

It is clear that cultural legacies are being lost as monuments decay and artifacts are taken out of the various countries. If contemporary cultures, fragmented and ruptured by centuries of colonialism, are going to be able to piece together and to reconnect with their severed past, archaeology needs to assume a more important place in African society. Great Zimbabwe is so important not simply because of its masterful masonry but because it is a cultural clue that survived and that has been reclaimed. Now it needs to be fully interpreted and placed within the larger context of sub-Saharan history, a context that still lies hidden. SA

The Author

WEBBER NDORO is a lecturer in the history department at University of Zimbabwe, where he teaches cultural resource management and museum practice. Ndoro, who holds degrees in archaeology from the University of Cambridge and in architectural conservation from New York University, was the conservator for Great Zimbabwe National Monuments and coordinator of the Monuments Program from 1988 to 1994.

Further Reading

ZIMBABWE: A RHODESIAN MYSTERY. Roger Summers. Thomas Nelson and Sons, 1963.
GREAT ZIMBABWE. Peter S. Garlake. Stein and Day, 1973.
THE SHONA AND ZIMBABWE, 900–1850. D. N. Beach. Africana, 1980.
DZIMBAHWE: LIFE AND POLITICS IN THE GOLDEN AGE, 1100–1500 A.D. Ken Mufuka. Harare Publishing House, 1983.
CULTURE AND DEVELOPMENT IN AFRICA. Edited by Ismail Serageldin and June Taboroff. International Bank for Reconstruction and Development/World Bank, 1994.
UNCOVERING THE PAST: A HISTORY OF ARCHAEOLOGY. William H. Stiebing, Jr. Oxford University Press, 1994.

Making Rice Disease-Resistant

For the first time, scientists have used genetic engineering to protect this essential crop from disease

by Pamela C. Ronald

Rice is arguably the world's most important food. Almost two billion people—one third of the world's population—depend primarily on rice for basic nourishment. Rice fields cover more than 360 million acres of land around the globe and yield 560 million tons of grain every year. But farmers plant much more rice than they harvest, because insects, bacteria, viruses and fungi often claim a substantial portion of each crop. One of the most devastating of these pestilences is blight, caused by bacteria common throughout Asia and Africa.

These bacteria—*Xanthomonas oryzae* pv. *oryzae* (known as *Xoo*)—spread rapidly from rice plant to rice plant and from field to field in water droplets. Infected leaves develop lesions, yellow and wilt in a matter of days. In severely infected fields, bacterial blight can wipe out half of a farmer's rice crop.

And yet rice plants possess an amazing assortment of genes that offer protection from a host of diseases, including bacterial blight. The farmers' predicament is that no single variety has every gene and that all plants are vulnerable to some diseases more than to others. Breeders have exploited disease-resistance genes in rice for nearly a century, redistributing this genetic wealth from hardy species to agriculturally useful varieties. But conventional breeding is painstaking and time-consuming; often a decade or more is needed to produce desired traits.

With the advent of genetic engineering, we are now able to introduce isolated disease-resistance genes directly into rice plants, trimming years from the time required to develop a useful variety. My colleagues and I recently cloned the first such disease-resistance

RICE PLANTS are subject to many destructive diseases, including bacterial blight, which causes devastating leaf damage and reduces yield (*left*). Water droplets carry the bacteria into leaf wounds; yellow lesions develop on infected leaves in days. If infected while very young, the entire plant may succumb. Certain plants, however, have a genetic resistance to blight (*right*).

gene from rice—a gene that protects against common forms of bacterial blight. We have used this gene to generate the world's first transgenic disease-resistant rice plants. These new varieties of rice have tremendous potential to aid farmers around the world.

Gene Hunt

The story of the first disease-resistance gene cloned from rice begins in the developing world. *Oryza longistaminata* is a wild species of rice native to Mali, useless as a crop plant because of poor taste and low yields but quite hardy when it comes to resisting bacterial blight. In 1977 researchers in India evaluated this plant for resistance to various strains of *Xoo* and found it could withstand every type they tested.

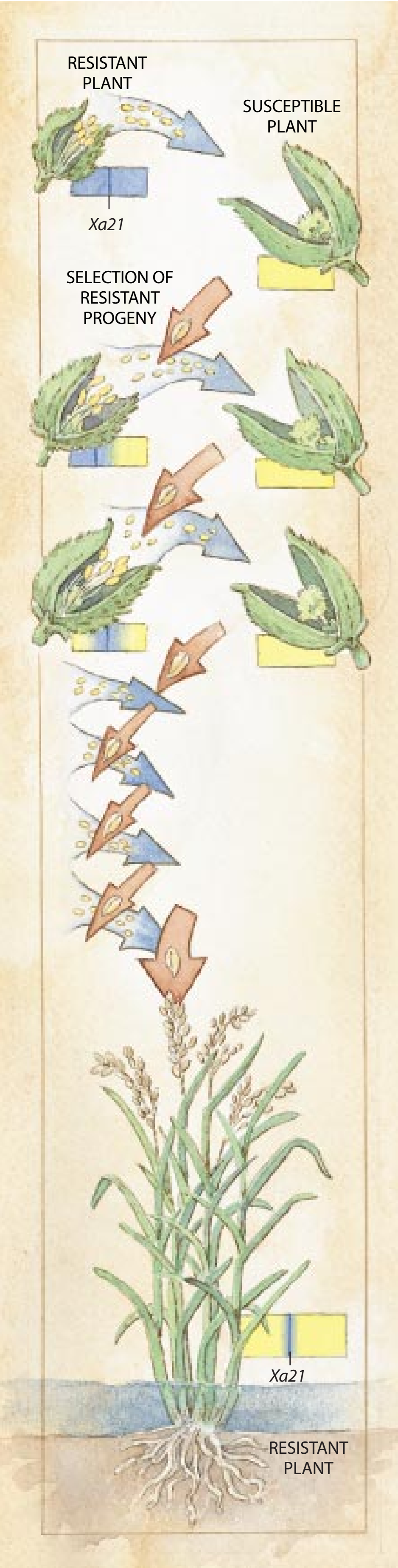
A year later Gurdev S. Khush and his co-workers at the International Rice Research Institute (IRRI) in the Philippines began work with *O. longistaminata.* Their goal was to transfer blight resistance from the wild rice to a cultivated variety, using conventional techniques. Twelve years of intensive breeding yielded one such resistant variety as well as the knowledge that the resistance was conferred by a small region of a single chromosome, perhaps even a single gene, which they named *Xa21.*

In 1990, just as they were reaping the intellectual rewards of this labor, I became a postdoctoral fellow at Cornell University. In an effort to learn about the mechanism of bacterial blight resistance, I decided to try to clone *Xa21* from the IRRI variety. A gene such as *Xa21* is a short stretch of genetic material (DNA) that encodes a protein and often, therefore, a trait (such as blight resistance). Although we frequently study traits in whole organisms, we cannot easily study single genes in the context of an organism's total genetic content, or genome. Nor can we accomplish genetic engineering without isolated genes. The solution is to generate accessible copies of individual genes in the process known as cloning.

To clone *Xa21,* I needed to identify the precise region of the rice genome that bears this gene, transfer this bit of DNA into bacteria where it could be easily copied, insert these copies into susceptible rice plants and then prove that the inserted DNA made these plants resistant to blight. When I began my effort, no one had yet succeeded in cloning a disease-resistance gene from any plant, although many such genes were known and used in traditional breeding. At that time, the barriers to genetic engineering of disease resistance fell into two categories: problems in finding the genes and problems in moving them around.

Finding a gene in a genome is a lot like the proverbial search in the haystack, and the genomes of most plants are particularly huge haystacks. As a rule of thumb, it is harder to locate specific genes in large genomes than in smaller ones; large genomes are also tough to manipulate. A simple yardstick is the tiny genome of the bacterium *Escherichia coli,* from which we can isolate genes quite easily. By this measure the rice genome is large—almost 100 times the size of the *E. coli* genome. A gene hunt in rice is challenging: my own search for *Xa21* would eventually require sever-

RICHARD JONES



RICHARD JONES

al years and many sophisticated techniques. Still, I was fortunate; researchers seeking to isolate genes from certain other grains face even bigger barriers. The genome of wheat, for example, is almost 3,500 times the size of the *E. coli* genome and fully five times the size of the human genome. Cloning a gene from grains such as rice and wheat is extremely difficult without some prior knowledge of the gene's location or sequence (think of trying to find a friend's house in New York City or Tokyo without an address or description).

In 1990 I felt the time was right for cloning genes from rice, because pioneering work led by Steven D. Tanksley and Susan R. McCouch, also at Cornell, had just produced a key development: a map to guide my exploration of the vast rice genome. The type of cloning I used is known as map-based cloning, and as the name implies, it requires some knowledge of the location of various landmarks, or markers, in the DNA. The genetic map constructed by the Cornell group showed the locations of hundreds of useful markers on the 12 rice chromosomes.

Over a period of a few years, first at Cornell and later at the University of California at Davis, my colleagues and I used this map to track down *Xa21.* During our search, we examined more than 1,000 rice plants to see how often the known DNA markers showed up in conjunction with resistance to bacterial blight. This strategy takes advantage of a certain amount of chromosomal swapping and rearranging that goes on during sexual reproduction: the closer two sites on a chromosome are, the less likely they are to be separated from each other during this process of recombination. In our case, the more often we saw resistance passed to progeny along with a given marker, the closer the resistance gene must lie to that marker.

TRADITIONAL BREEDING has been used for years to produce disease-resistant rice. Pollen from a resistant plant fertilizes a susceptible plant that also has desirable characteristics—it produces high yields of grain, for instance, or tastes good. Progeny inherit a random mixture of genetic material from both parents (*colored bars*). Resistant progeny are crossed again with the susceptible plant, endowing offspring with more of that parent's valuable traits. Selection of resistant progeny at each cross ensures the continuing presence of the resistance gene—in this case, *Xa21*.

By sheer luck, the first chromosomal landmark that my group and I identified as lying very close to *Xa21* turned out to be incredibly useful. One weekend in May 1994, two years after I set up my own laboratory at Davis, I discovered that the sequence of the marker DNA was similar to that of several disease-resistance genes recently cloned from tobacco, tomato, flax and a mustard plant. Alone in the lab that Sunday morning, I called my longtime friend and colleague John Salmeron of the University of California at Berkeley and asked him to compare my sequence more carefully with his tomato disease-resistance gene. We were thrilled to find very strong similarities among genes from such different plants. I felt confident that I was searching in the right neighborhood.

My group and I spent the next year cloning candidate *Xa21* genes and preparing to insert them into other rice plants. We knew that the crucial test would come when we transferred our isolated rice DNA into a plant normally susceptible to *Xoo.* If we had cloned the right gene, the resulting transgenic plants would be resistant to bacterial blight. We were anxious to begin these experiments, but we faced an undeniable obstacle: we had no experience introducing genes into rice cells. And at that time, only a very few labs in the world were able to carry out this process, called transformation, in rice.

Under the Gun

This problem of transferring genes into plant cells is the second great hurdle in engineering disease resistance. Many types of plant cells, including rice, are refractory to taking up extraneous DNA. The breakthrough came in 1987, when John C. Sanford of Cornell developed a gun that shoots microscopic particles into intact cells [see "Transgenic Crops," by Charles S. Gasser and Robert T. Fraley; Scientific American, June 1992]. Sanford's early versions were propelled by a gunpowder charge; later models are helium-driven and fire pellets made of gold. These pellets, which are less than a hundredth of a millimeter in diameter, can be coated with DNA that they then carry directly into cells.

Researchers did not use this technique in rice until 1991; when we were ready to test our *Xa21* clone, the International Laboratory for Tropical Agricultural Biotechnology (ILTAB) was one of the facilities doing so routinely. It is conveniently located in California and, to our delight, agreed to help us.

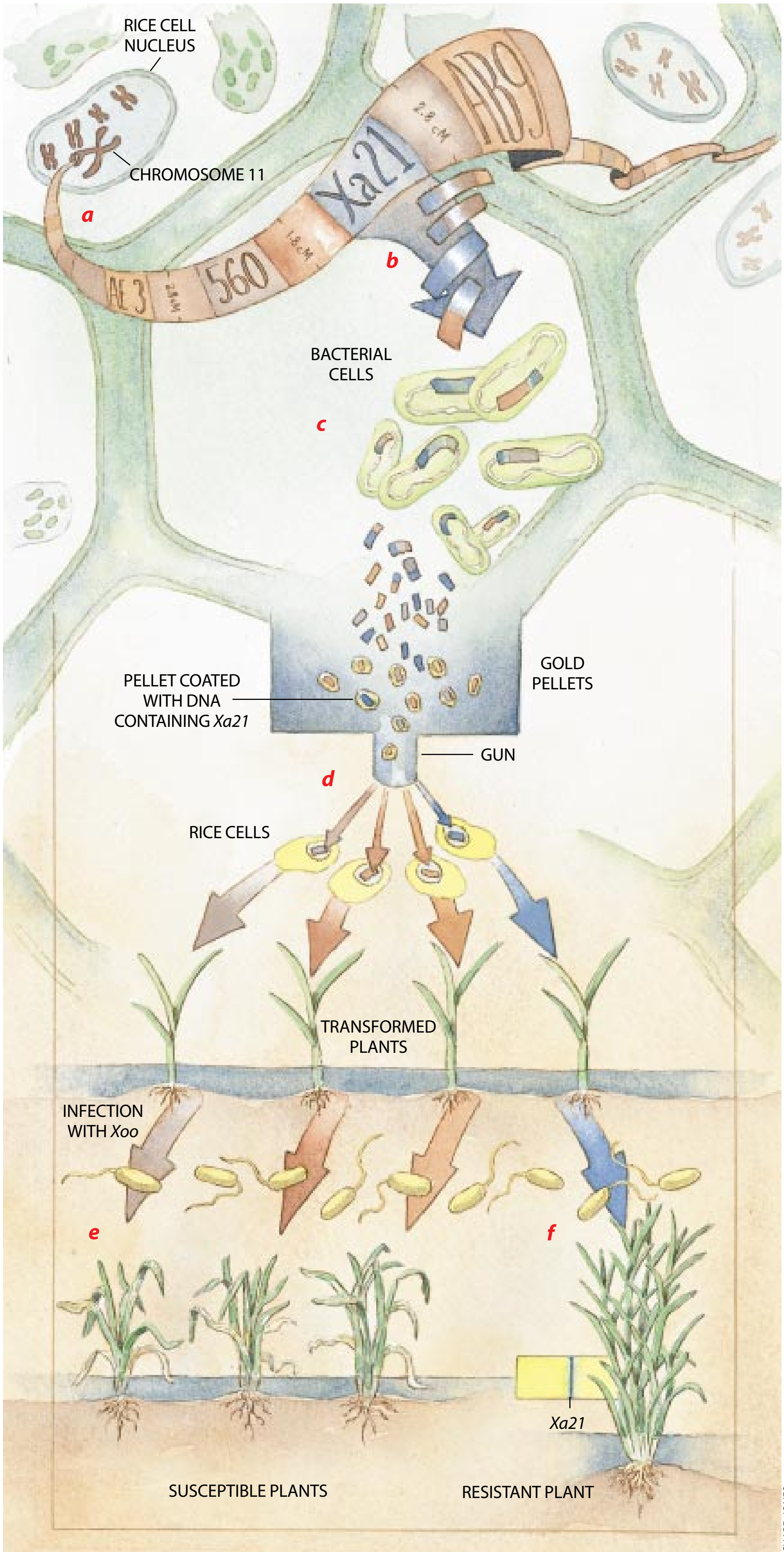
Researchers at ILTAB used the gun to transform our cloned DNA into rice cells of the variety Taipei 309. This is an old variety that is no longer grown, but we chose it because it is easily transformed and susceptible to *Xoo*. We grew 1,500 plants from the transformed cells; each plant had a bit of cloned DNA in every cell. When our plants were six weeks old, it was finally time to test for resistance to bacterial blight.

We exposed each of our transgenic plants to *Xoo* by trimming their leaves with scissors dipped in a bacterial suspension. Ten days later we examined the plants for lesions caused by the bacteria. We found that of the original 1,500 transgenic plants, 50 plants were highly resistant to infection with *Xoo:* each had lesions between 75 and 90 percent shorter than those in the original susceptible plants. In these 50 plants, the transformed piece of DNA contained an intact blight-resistance gene.

Pièce de Résistance

We had succeeded in cloning *Xa21*. Subsequently, we showed that *Xa21* was passed on to the next generation through self-fertilization, giving rise to seedlings that were also resistant to bacterial blight. We challenged our transgenic plants with 31 different *Xoo* strains from eight countries spanning Asia and as far flung as Colombia. The plants resisted infection by 29 of these strains, exactly replicating the disease-resistance profile of their wild African predecessor. For the first time, we were able to engineer rice for resistance to bacterial blight.

Our current goal is to insert *Xa21* into varieties that, unlike Taipei 309, are agriculturally important. In collaboration with ILTAB, we have successfully



GENETIC ENGINEERING of disease resistance requires isolation of individual genes that confer resistance. To isolate *Xa21,* the author determined its approximate location on rice chromosome 11 (*a*) and then generated pieces of DNA that spanned the region (*b*). Next, the DNA was copied in bacteria (*c*), and the replicas were placed on gold pellets that were shot into cells of disease-susceptible rice plants (*d*). Some cells took up the foreign DNA but not the gene for resistance (*e*); a few cells, however, did receive intact *Xa21* and grew into resistant plants (*f*).

introduced *Xa21* into two popular varieties—IR64 and IR72—which are grown on about 22 million acres in Asia and Africa. Our ongoing studies show that the transgenic plants are blight-resistant. And recently we have also engineered resistance in Ming Hui 63, a variety of rice widely grown in China.

With these exciting results in hand, I have sent *Xa21* to scores of scientists throughout Europe, Africa, Asia and the U.S., with the objective of introducing bacterial blight resistance into locally important rice varieties. Because growing conditions vary greatly from place to place, farmers often prefer to plant a variety of rice that is well adapted to their particular region. These varieties possess valuable traits such as drought resistance, cold tolerance, short stature (for wind resistance) or resistance to indigenous pests and diseases. The genetically engineered versions will be identical to the original plants except for the addition of the single cloned gene conferring resistance to bacterial blight.

Once we have generated these new varieties, we need to field-test the plants for yield, taste and hardiness to establish that the useful traits of the original varieties remain unchanged. In the next few years, researchers in California, Asia and Africa will field-test transgenic rice containing *Xa21*. If these lines perform as well as locally adapted varieties, national breeding programs will distribute seed to farmers in developing countries. Because the disease-resistance transgene is passed on to progeny, farmers can grow their own seed for the next season.

Crops of the Future

Compared with conventional breeding, genetic engineering is quick and flexible: we can shuttle individual cloned genes between plants in a matter of months. Donor and recipient need not be compatible for breeding; we can share genes among disparate species, even among different crops.

Thus, scientists should be able to harness cloned resistance genes to control disease in many crops besides rice. Species of *Xanthomonas* that cause blight, for example, infect virtually all crop plants. In Florida, 99 percent of the citrus crop is susceptible, and growers must closely monitor bacterial infections to prevent epidemics. In the mid-1980s more than 20 million orange trees were burned to thwart a suspected outbreak of this disease. State and federal governments spent more than $40 million on eradication alone, and hundreds of growers lost even more in unrealized produce. Scientists may someday be able to protect citrus and other lucrative crops by manipulating the rice bacterial blight-resistance genes and transferring them into susceptible species.

Genetic engineering may also help us cope with the problems any disease-resistant plant faces once it is in the field. In particular, pathogens may mutate and overcome the protection a given resistance gene confers. Breeders must therefore continually identify and introduce useful genes in order to minimize susceptibility to disease, whether through conventional methods or genetic engineering. Fortunately, many resistance genes are known and are ripe for cloning. Combinations of these genes may further enhance disease resistance, much the same way that combinations of antibiotics or antiviral drugs combat microbes such as tuberculosis bacteria and human immunodeficiency virus.

We also hope to incorporate resistance to more than one pathogen in a single transgenic line. In some instances, farmers cannot use rice varieties bred for resistance to bacterial blight, because they lack resistance to other pathogens and pests. The most serious of these pests is the brown plant hopper, an insect that causes severe damage to rice plants as it feeds. It also transmits damaging pathogens such as the grassy-stunt and ragged-stunt viruses. In an early effort at engineering resistance to multiple threats, we are collaborating with colleagues in China and England to incorporate resistance to both bacterial blight and the brown plant hopper into several important varieties of rice, using cloned genes, including *Xa21*. As more and more resistance genes are cloned, the number of available combinations will increase exponentially.

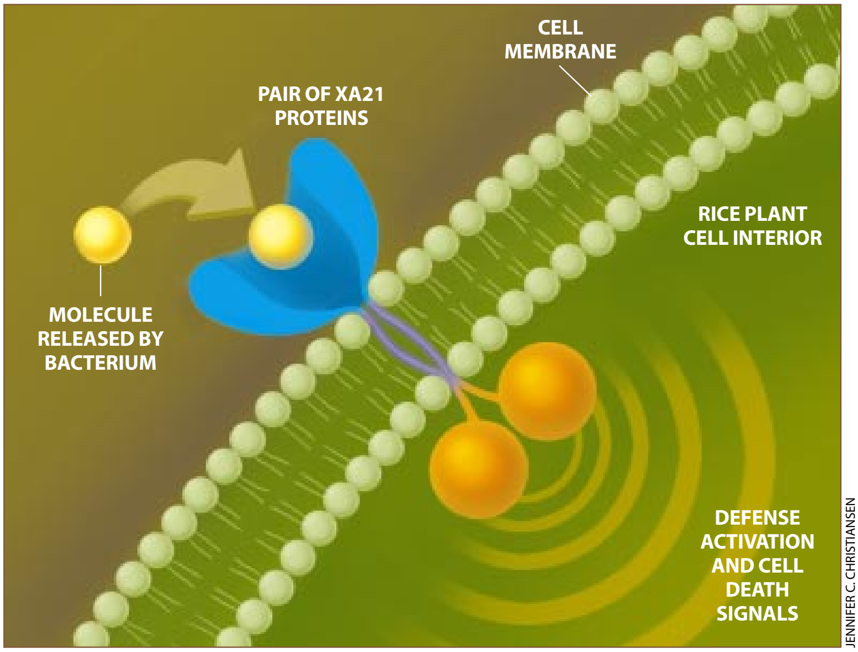
Transgenic disease-resistant plants hold great commercial promise. Although no farmers are actually growing such plants yet, U.S. companies are leading the commercialization of other transgenic crops. The Flavr Savr tomato, developed by Calgene for increased shelf life, was the first commercially available genetically engineered food. Soybeans resistant to the herbicide Roundup came on the market in 1996;

RICE LEAVES exposed to *Xoo* resist infection when the gene *Xa21* is present, both in our transgenic Taipei variety (*pair on left*) and in traditionally bred plants (*pair on right*). Infected leaves from susceptible plants develop extensive, yellow lesions (*center pairs*).

Sounding the Alarm

How do our blight-resistant rice plants sense bacterial intruders? We think the protein encoded by *Xa21* acts as a kind of receiver; pairs of the proteins most likely span the cell membrane, picking up external signals sent by the bacteria and relaying them inside the cell. A protective response ensues: the alerted cell signals its neighbors to mount a defense and then dies; groups of dead cells prevent further spread of the invader. The antennalike part of the protein outside the cell resembles proteins from animals that recognize and bind to other molecules. We have not yet found the bacterial molecule that tips off the rice cells to their enemy's presence, but my group is looking for it.

The portion of protein inside the rice cell is also familiar: it appears to be a kinase, a common type of enzyme responsible for prodding cells to action. We think this enzyme switches on in response to the bacterial signal, trumpeting a message throughout the cell to activate defenses. —*P.C.R.*



XA21 PROTEINS appear to consist of three parts: a section that detects signals from bacteria (*blue*), a section that spans the membrane of the rice cell (*purple*) and a section that generates a message inside the rice cell (*orange*).

maize engineered for resistance to herbicide was recently approved for sale in the U.S. and Canada.

Industrial nations will probably benefit most from the currently available transgenic products. In the developing world, for instance, farmers often cannot afford technologies that require input of expensive herbicides. In contrast, both developing and industrial countries are likely to find disease-resistant transgenic grains useful. These grains may also enjoy more acceptance than certain other new transgenics (such as controversial insecticide-producing plants that some fear will lead quickly to insecticide-resistant insects). Transgenic disease-resistant plants may eventually have a significant effect on the economics of crop growth, promoting more efficient land use, better global food supply and environmentally safer methods of disease and pest control.

With the commercial promise of transgenic disease-resistant crops comes a social responsibility. In 1996 Davis established the Germplasm Resource Recognition Fund to acknowledge the contributions of developing nations to the success of the university's programs, including *Xa21,* for which the university has filed a patent application. Financed by income from commercialization of genetic materials obtained in the Third World, the fund will provide fellowship assistance to researchers from developing nations. Farmers in these poor regions will also be able to obtain seeds of our transgenic lines at the same cost as the traditional parent lines. The fund provides a means for University of California scientists to patent their inventions and make them into commercially viable products while recognizing and fostering contributions from the developing world.

The potential of genetic engineering in rice and other grains will not be exhausted with disease resistance. The future will undoubtedly bring the cloning of many more genes responsible for other valuable traits (cold tolerance, perhaps, or drought resistance). Ultimately, breeders and farmers will be able to choose from a whole toolboxful of cloned genes—genes that will let them reap more of what they sow. SA

The Author

PAMELA C. RONALD received a Ph.D. from the University of California, Berkeley, in 1990. Her graduate work in the laboratory of B. J. Staskawicz focused on the genetic basis of disease resistance in tomatoes and peppers. She began her work on rice disease resistance as a postdoctoral fellow in Steven D. Tanksley's laboratory at Cornell University. This work continues at U.C.-Davis, where Ronald is an associate professor. When not working on the problems of the world's crops, she enjoys the harvest from her husband's organic farm.

Further Reading

PLANT DISEASES: THEIR BIOLOGY AND SOCIAL IMPACT. G. Schumann. American Phytopathological Society, 1991.

RICE, THE ESSENTIAL HARVEST. Peter T. White. *National Geographic,* Vol. 185, No. 5, pages 48–79; May 1994.

MOLECULAR GENETICS OF PLANT DISEASE RESISTANCE. B. J. Staskawicz et al. in *Science,* Vol. 268, pages 661–667; May 5, 1995.

A RECEPTOR KINASE-LIKE PROTEIN ENCODED BY THE RICE DISEASE RESISTANCE GENE, *Xa21*. Wen-Yuan Song et al. in *Science,* Vol. 270, pages 1804–1806; December 15, 1995.

SIGNALING IN PLANT-MICROBE INTERACTIONS. B. Baker et al. in *Science,* Vol. 276, pages 726–733; May 2, 1997.

by Shawn Carlson

Caught in a Wind Tunnel

Controlled flight might be the most vexing experimental problem ever to be solved. Some of the greatest thinkers through the ages have attacked it, including Leonardo da Vinci, who penned more than 500 drawings and 35,000 words on the subject. Of the many intractable difficulties human flight posed, aerodynamic stability ultimately proved to be the hardest to master. An airplane can pitch up and down, roll left and right, and yaw side to side. Straight and level flight requires all three of these motions to be managed simultaneously.

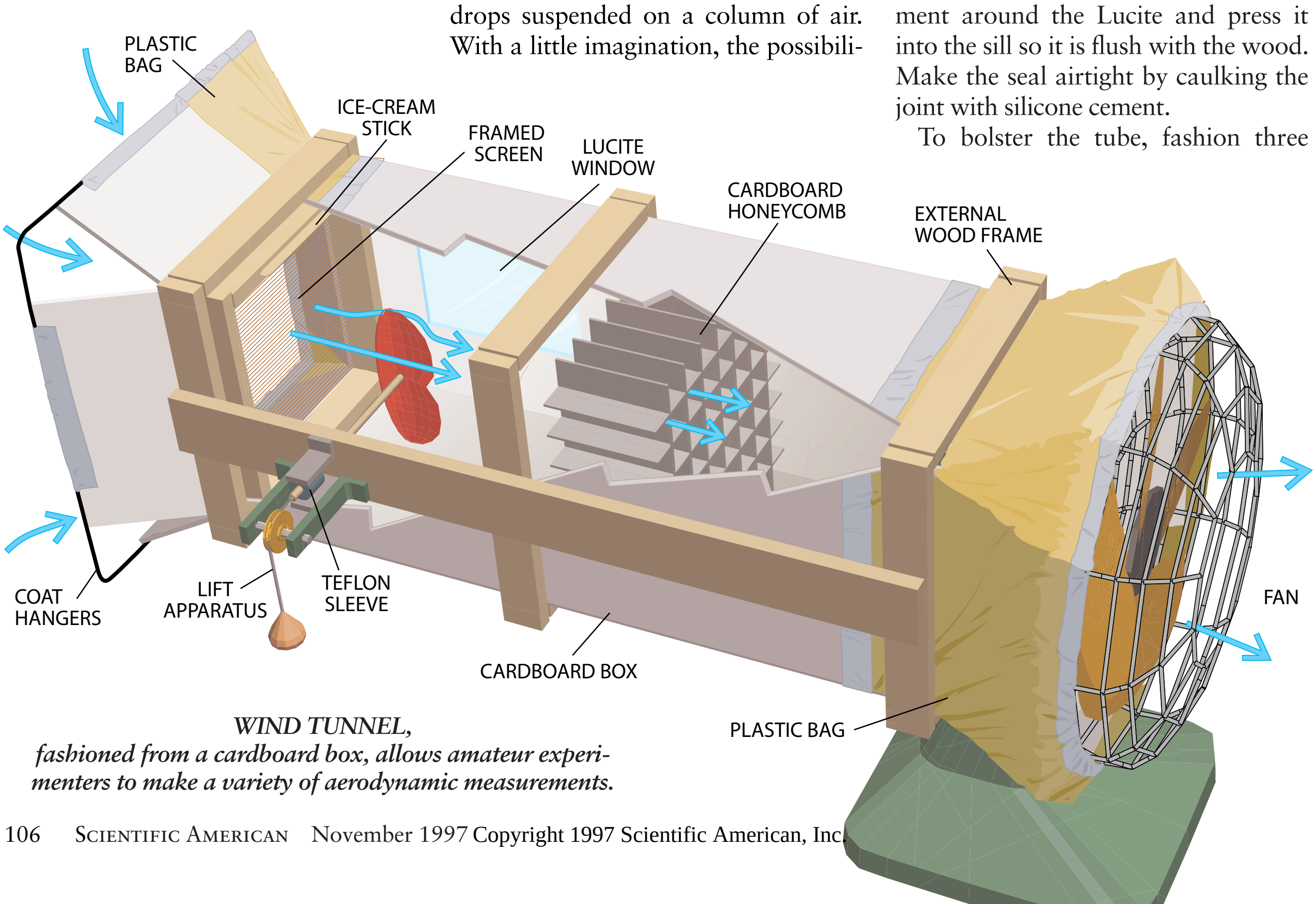
That challenge was finally answered by two of the most talented amateur researchers the U.S. has ever seen. Although most people think of them more as working-class do-it-yourselfers, Orville and Wilbur Wright would be better remembered as gifted scientists. One of their greatest contributions began in 1901, when they developed the wind tunnel into a precise research instrument. The Wright brothers used it to perform thousands of systematic measurements, the first truly reliable determinations of this kind in the emerging field of aeronautics. That effort made possible their success two years later at Kitty Hawk, N.C.

Wind tunnels still afford amateurs countless research opportunities in aerodynamics and beyond. Of course, a wind tunnel will allow you to tailor the design of kites, model airplanes, sailboats and racing cars for improved performance. In fact, a wind tunnel will let you study almost anything that is affected by moving air. For example, you can observe the interplay between waves and wind on a liquid surface. Environmentalists may want to examine evaporation through different types of soils or measure the wind's velocity profile above a tray of grass, soils or asphalt. With a wind tunnel, you can also investigate how insects cope with strong breezes or mount the chamber vertically to find the terminal speed of waterdrops suspended on a column of air. With a little imagination, the possibilities for exploration are probably endless.

For my own experiments, I recently put together an inexpensive wind tunnel that uses a household fan to draw air through the test region. A long cardboard box, available from Mail Boxes Etc. or other similar shipping outlets, forms the central tube. Mine measures 30 by 30 by 122 centimeters (12 by 12 by 48 inches), but the precise dimensions are not important. Just make sure that the length of the box is at least four times its width. Painting the interior of the box a glossy white will make your experiments easier to see. To keep slight air currents from passing directly through the cardboard itself, you should double-coat the exterior with latex-based enamel paint.

A sheet of Lucite forms the window. Have your local hardware store cut a rectangular pane 20 by 30 centimeters (eight by 12 inches) in size. Trace the windowsill around the pane [*see illustration below for placement*] and cut out the opening using a box cutter. Temporarily place waxed paper behind the hole and support it with a wide scrap of wood. Then run a bead of silicone cement around the Lucite and press it into the sill so it is flush with the wood. Make the seal airtight by caulking the joint with silicone cement.

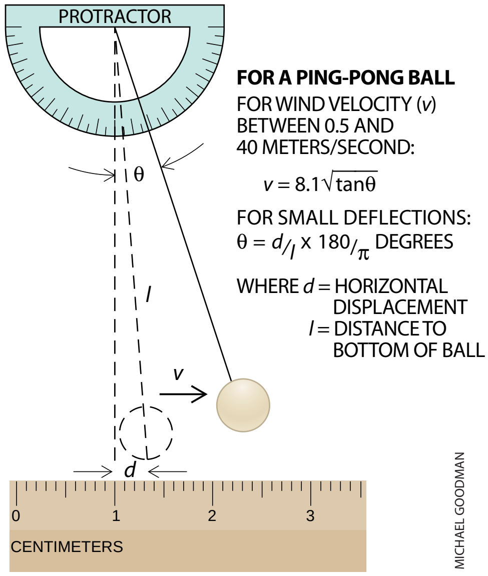
To bolster the tube, fashion three



WIND TUNNEL, fashioned from a cardboard box, allows amateur experimenters to make a variety of aerodynamic measurements.

tight-fitting rectangular frames from wood slats and glue them to the outside of the tunnel. Horizontal slats help to make the structure more rigid and support the measuring equipment.

Some household fans can drive wind as quickly as five meters per second. Adding a restriction, like the one shown in the illustration on the next page, increases the speed of the airflow (just as putting your thumb over the end of a garden hose shoots the water more rapidly). Decreasing the cross-sectional area in the test region by two thirds will, for example, nearly triple the airspeed you can attain. If you need still stiffer winds, check your local industrial liquidators for more powerful fans. Most commercial fans use a four-position switch to control their speed. For finer adjustment, replace this unit with a household light dimmer. Use duct tape to attach the flaps of the box to the fan. Plastic trash bags, slit at the bottom, make excellent airtight sleeves. Slip one over the fan and the box and tape it into place.



MICHAEL GOODMAN

PING-PONG BALL suspended from a string serves to measure wind speed.

Flair the entrance to the wind tunnel by bending the flaps on the opposite side outward and attaching them to a square fashioned from two wire coat hangers. Epoxy and then tape the wire square to the outside of the flaps as shown. Afterward, place another garbage-bag sleeve around this end of the tunnel to make it airtight as well.

The exhaust from the fan creates air currents in the room, and some of these swirling eddies will invariably drift back into the tunnel. A single layer of window screen at the mouth of the tunnel helps to smooth out these unwanted irregularities. Staple the screen to a wood frame that fits snugly inside the opening. Line the outer edge with felt to prevent air from leaking in around the sides. You will need access to the interior, so do not affix the frame permanently. Glue four ice-cream sticks onto the inner walls just deep enough inside from the mouth to keep the screen in place during tests. A "honeycomb" assembly (really, a square grid of cardboard baffles) just downstream of the test region helps to maintain a smooth flow of air for your experiments.

You can measure the airspeed inside the tunnel using a number of devices, including a hot-ball anemometer [see

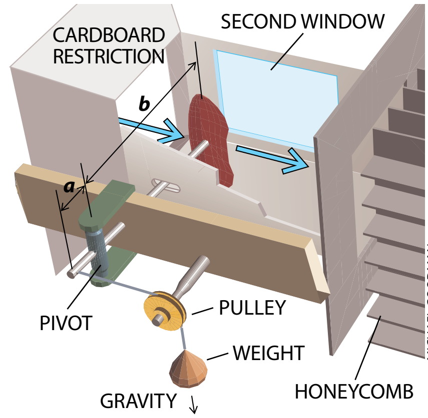
AVERAGE WEIGHTS OF U.S. COINS		
DENOMINATION	MASS (GRAMS)	WEIGHT (NEWTONS)
DIME	2.264	2.220×10^{-2}
PENNY	3.110	3.049×10^{-2}
NICKEL	4.999	4.902×10^{-2}
QUARTER	5.669	5.559×10^{-2}
HALF DOLLAR	11.50	1.128×10^{-1}
DOLLAR	26.73	2.621×10^{-1}

JOHNNY JOHNSON

The Amateur Scientist, November 1995, for construction details], a cup anemometer from an old weather station, a manometer or an ultrasonic anemometer. Or you can combine a Ping-Pong ball, a protractor and a length of white thread into a simple instrument [*see illustration on preceding page*]. The angle the string attains will depend on the relative strengths of the forces of gravity and aerodynamic drag on the ball. Use the equation given in the illustration. The value you obtain should be good to about 10 percent, but because of certain subtleties of fluid dynamics, it is valid only between 0.5 and 40 meters per second.

People with a passion for aeronautics will probably want to measure aerodynamic forces. Determining all possible forces and torques requires six simultaneous measurements. But experimenters are often interested in just one quantity—the lift on a wing or the drag on a surface, for example. The two setups shown will let you measure either lift or drag (but not both simultaneously). Mounting your models sideways takes gravity out of the equation.

To measure lift, increase the counterweight until the model remains in place when released. The weight then equals the aerodynamic force on the model. For drag, the counterweight applies a torque that balances the torque applied by the drag force. The drag force then equals the ratio of the moment arms (a/b) times the counterweight. If you don't have a set of calibrated weights, loose change will do [*see table above*]. If you have more sophisticated needs, the references listed on the World Wide Web site of the Society for Amateur Scientists describe more elaborate balance systems. As a final challenge, you may want to figure out how to record the forces electronically. SA



MICHAEL GOODMAN

***DRAG APPARATUS** uses a pivot and weights to gauge the aerodynamic force.*

For information about this project or other activities for amateur scientists, write the Society for Amateur Scientists, 4735 Clairemont Square, Suite 179, San Diego, CA 92117. You can also visit the society's World Wide Web site at www.thesphere.com/SAS/, call (619) 239-8807 or leave a message at (800) 873-8767.

by Ian Stewart

The Lore and Lure of Dice

STANDARD DICE resemble the white ones depicted: the faces with 1, 2 and 3 dots cycle counterclockwise around a common corner. The black dice have these faces arranged clockwise.

The die, more commonly known by its plural, "dice," is one of the earliest gambling aids. Herodotus claimed that dice were invented by the Lydians in the time of King Atys, but Sophocles disagreed, crediting the invention to a Greek called Palamedes, allegedly during the siege of Troy. Although it may seem plausible that dice were introduced to while away the time of bored besiegers, archaeologists have discovered cubical dice, to all intents and purposes just like today's, in Egyptian tombs dating from 2000 B.C. Dice have also been found with Chinese remains from about 600 B.C.

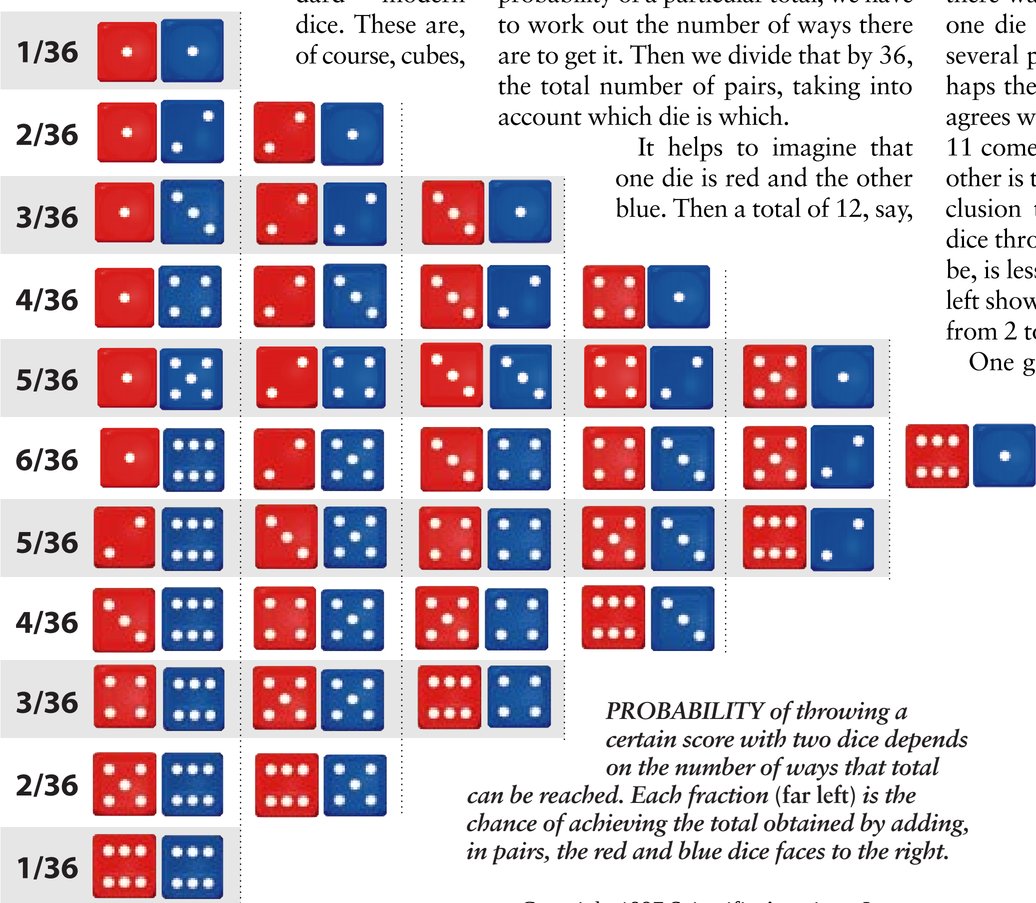
Dice with diverse shapes and strange markings, made of materials ranging from beaver teeth to porcelain, have been used by North American Indians, Aztecs and Mayans, Polynesians, Inuits and many African tribes. In this column I'm going to focus exclusively on standard modern dice. These are, of course, cubes, with a pattern of spots on each face, the numbers of spots being 1, 2, 3, 4, 5 and 6. Spots on opposite faces sum to 7, so the faces come in three pairs: 1 and 6, 2 and 5, 3 and 4. There are exactly two possible arrangements with this property, and one is the mirror image of the other. Nowadays virtually all dice of western manufacture resemble the white ones in the illustration at the right, in which the faces 1, 2 and 3 cycle around their common vertex in a counterclockwise direction. I am told that in Japan, dice with this handedness are used in all games except mah-jongg, where mirror-image dice, such as the black ones, are used instead. From now on I'll use western dice, unless otherwise stated.

Dice are often thrown in pairs, with the aim of achieving a desired total. Assume first that the dice are "fair," so that each face has a $1/6$ probability of coming up on top. To calculate the probability of a particular total, we have to work out the number of ways there are to get it. Then we divide that by 36, the total number of pairs, taking into account which die is which.

It helps to imagine that one die is red and the other blue. Then a total of 12, say, can occur in only one way: red die = 6, blue die = 6. The probability of a total of 12 is therefore $1/36$. A total of 11, on the other hand, can occur in two ways: red die = 6, blue die = 5, or red die = 5, blue die = 6. Its probability is therefore $2/36$, or $1/18$.

The great mathematician and philosopher Gottfried Leibniz thought that the probabilities of throwing 11 and 12 must be the same, because he thought there was only one way to throw 11—one die = 6, the other = 5. There are several problems with this theory. Perhaps the most significant is that it disagrees wildly with experiment, in which 11 comes up twice as often as 12. Another is that it leads to the unlikely conclusion that the probability that two dice throw some total, whatever it may be, is less than 1. The illustration at the left shows the probabilities for all totals from 2 to 12.

One game in which an intuitive feel for these probabilities is crucial is craps, which dates from the 1840s. One player, the shooter, puts up a sum of money. The others "fade" it—that is, bet an amount of their own choice. If the total faded is less than the shooter's initial bet, he reduces the bet to match that total. The shooter then rolls the dice. With a score of 7 or 11 (a "natural") on the first roll, he wins outright; with a score of 2, 3 or 12



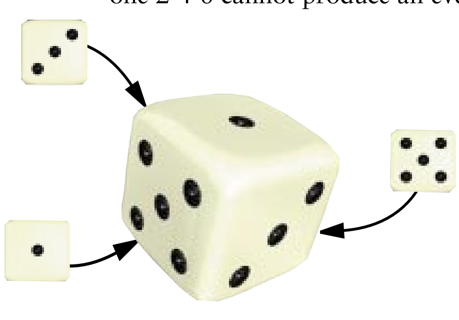
PROBABILITY of throwing a certain score with two dice depends on the number of ways that total can be reached. Each fraction (far left) *is the chance of achieving the total obtained by adding, in pairs, the red and blue dice faces to the right.*

ILLUSTRATIONS BY JENNIFER C. CHRISTIANSEN ("craps"), he loses. Otherwise the shooter's initial score—4, 5, 6, 8, 9 or 10—becomes his "point." He continues to roll, aiming to score the point again before he throws 7 ("craps out"). If he succeeds, he wins all the money; if he fails, he loses all.

From the probabilities and the rules of the game, it can be calculated that the shooter's chance of winning is $^{244}/_{495}$, roughly 49.3 percent. This is just less than even (50 percent). Professional gamblers can turn this slight disadvantage into an advantage by two methods. One is to accept or reject various "side bets" with other players. The other is to cheat, using sleight of hand to introduce rigged dice into the game.

Dice can be rigged in a variety of ways. Their faces may be subtly shaved so that their corners are not right angles, or they can be "loaded" with weights. Both techniques make some throws more probable than others. More drastically, the standard dice may be replaced by "tops" and "bottoms," dice bearing only three distinct numbers of spots. (Opposite faces have identical numbers.) The illustration below shows an example with the faces 1, 3 and 5 only. Because each player sees at most three faces of a die at any instant, and because no two adjacent faces have the same number of spots, nothing appears amiss at a cursory glance. It is not possible, however, to ensure that the arrangements at all vertices cycle in the standard order. Indeed, if the order is 1-3-5 counterclockwise around one vertex, then it must be 1-3-5 clockwise around an adjacent vertex.

Tops and bottoms can be used in craps for diverse purposes. A pair of 1-3-5 dice, for instance, can never throw 7, so with these a player can never crap out. A combination of one 1-3-5 and one 2-4-6 cannot produce an even



"TOPS," or fake dice having, say, just 1, 3 or 5 spots on any face, can never "crap out" in a game of craps.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Reprints: $4.00 each (minimum order, 10 copies) prepaid. Articles published within 3 months of current issue available. Write Reprint Dept., Scientific American, 415 Madison Ave., New York, NY 10017-1111.

Back issues: $8.95 each ($9.95 outside U.S.) prepaid. Most numbers available. Credit card (Mastercard/Visa) orders for two or more issues accepted. To order, fax (212) 355-0408.

Index of articles since 1948 available in electronic format. Write SciDex®, Scientific American Selections, P.O. Box 11314, Des Moines, IA 50340-1314, or call (800) 777-0444. E-mail: info@sciam.com

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of Scientific American for the fee of $3.50 per copy of each article plus $0.50 per page. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Specify CCC Reference Number ISSN 0036-8733 /96. $3.50 + 0.50.

Editorial correspondence should be addressed to The Editors, Scientific American, 415 Madison Ave., New York, NY 10017-1111. Unsolicited manuscripts are submitted at the authors' risk and will not be returned unless accompanied by a stamped, self-addressed envelope. E-mail: editors@sciam.com

Advertising correspondence should be addressed to Advertising Manager, Scientific American, 415 Madison Ave., New York, NY 10017-1111, or fax (212) 754-1138. E-mail: advertising@sciam.com

Subscription correspondence should be addressed to Subscription Manager, Scientific American, P.O. Box 3187, Harlan, IA 51537. The date of the last issue of your subscription appears on each month's mailing label. For change of address, please notify us at least four weeks in advance. Please send your old address (mailing label, if possible) and your new address. We occasionally make subscribers' names available to reputable companies whose products and services we think will interest you. If you would like your name excluded from these mailings, please send your request with your mailing label to us at the Harlan, IA address. E-mail: customerservice@sciam.com

Visit our Web site at http://www.sciam.com/

A	3	4	8
B			
1	A	A	A
5	B	B	A
9	B	B	B

B	1	5	9
C			
2	C	B	B
6	C	C	B
7	C	C	B

C	2	6	7
A			
3	A	C	C
4	A	C	C
8	A	A	A

NONTRANSITIVE DICE, labeled A, B and C, have the peculiar property that on average, B beats A, and C beats B, but A beats C. A has only 3, 4 or 8 spots on its faces; B has only 1, 5 and 9; whereas C has only 2, 6 and 7. The chart shows the possible outcomes when any of these strange dice competes with another.

total, so with these dice a player cannot throw 4, 6, 8 or 10. Tops must be used sparingly if their presence is to be undetected—even the most naive players will start to wonder when they keep throwing even totals.

Many conjuring or party tricks use dice. A lot of them are based on the rule that opposite faces sum to 7. One game is described by Martin Gardner in his book *Mathematical Magic Show* (Alfred A. Knopf, 1977). The magician turns her back and asks a member of the audience to roll three standard dice and add up the top faces. Then the victim is told to pick up any die and add its bottom number to the total. Finally, he rolls the same die again and adds its top number to the previous total. (The volunteer keeps all these totals to himself.) Now the magician turns back around and immediately states what the result was—even though she has no idea which die was chosen.

How does this work? Suppose that the top faces of the dice have numbers a, b and c and that die a is chosen. The initial total is $a + b + c$. To this is added $7 - a$, making $b + c + 7$. Then a is thrown again, giving d, and the final result is $d + b + c + 7$. The magician then looks at the three dice, which total $d + b + c$—so all she has to do is quickly add them up and add 7.

Henry Ernest Dudeney, the English puzzlist, includes a different trick in his book *Amusements in Mathematics* (Dover Publications, 1958). Again the magician asks for three dice to be thrown while her back is turned. This time the victim is asked to double the value of the first die and add 5; multiply the result by 5 and add the value of the second die; then multiply the result by 10; and add the value of the third die. On being told the result, the magician immediately says what the three dice values were. The result, of course, is now $10(5(2a + 5) + b) + c$, or $100a + 10b + c$

FEEDBACK

The mailbag about Juniper Green [March] was so enormous that Feedback can do even less justice to it than usual. First, apologies for several errors. The worst was my reference to "Richard Porteous" as the game's inventor; he is not Richard but Rob. Marc Loveday of Fruita, Colo., pointed out that step 18 of the first chart fails to observe that Bob can choose 18 instead of 2. Many readers, including Arlin Anderson of Madison, Ala., and William J. Shlaer of Vashon Island, Wash., corrected me on the status of JG-1, which is secondary because Alice cannot play 1 to open and hence cannot play at all. Anderson also corrected me on JG-9, which is secondary.

Porteous sent me an analysis of JG-100 carried out by Peter Conlon, Monique Barendse and Laurie Fischer, three of his current pupils. They show in particular that winning strategies exist if you start with 58 or 62. Michael D. Tibbetts of Clearwater, Fla., came to the same conclusion. He notes that prime numbers play a role in all aspects of the game and divides them into four kinds: large (53, 59, 61, 67, 71, 73, 79, 83, 89, 97), medium large (37, 41, 43, 47), medium (29, 31) and small (17, 19). The winning openings are twice the medium primes.

Paul J. Blatz of Van Nuys, Calif., recalled that the game was discussed in a number theory course given at Princeton University by Eugene P. Wigner in the late 1930s. A criterion for winning play in all cases was provided. The answer for JG-n depends on the oddness or evenness of the powers of primes that occur when $n!$ (or factorial n: $n! = n(n - 1)(n - 2)\ldots1$) is factorized. —*I.S.*

+ 250. So the magician subtracts 250 from the result, and the three digits of the answer are the numbers on the dice.

Other dice problems involve modified dice with nonstandard numbering. For example: Can you think of a way to label two dice, using only numbers 0, 1, 2, 3, 4, 5 or 6, to get a pair of dice such that all totals from 1 to 12 are equally likely? (See the end for the answer.) Perhaps the most counterintuitive dice phenomenon is that of "nontransitive dice." Make three dice, A, B and C, numbered like this:

A: 3 3 4 4 8 8
B: 1 1 5 5 9 9
C: 2 2 6 6 7 7

In the long run, B beats A. In fact, die B throws a higher number than A with probability $5/9$. Similarly, C beats B with probability $5/9$. So obviously C beats A, right? No, A beats C with probability $5/9$. The chart on the opposite page justifies these assertions. You can make a fortune with a set of such dice! Let your opponent choose one; then you choose whichever one beats it (in the long run, with probability greater than evens). Repeat. You will win on 55.55 percent of all plays. Yet your opponent has a free choice of the "best" die!

A word of warning, though: don't place too much reliance on probability theory without making the rules of the game very precise. In his marvelous little book *The Broken Dice* (University of Chicago Press, 1993), Ivar Ekeland tells the story of two Nordic kings who played dice to decide the fate of a disputed island. The Swedish king rolled two dice and scored a double 6. This, he boasted, was unbeatable, so King Olaf of Norway might as well give up. Olaf muttered something to the effect that he, too, might score a double 6 and cast his two dice. One turned up 6; the other split into two pieces, one showing a 1 and the other a 6. Total: 13! All of which goes to prove that what you think is possible depends on how you model the problem.

Mind you, there are a few cynics who think Olaf rigged the whole scam. SA

ANSWER: To make all totals from 1 to 12 equally likely, one die must have faces 1, 2, 3, 4, 5 and 6 and the other 0, 0, 0, 6, 6 and 6.


EVERETT COLLECTION

NUTTY PROFESSOR epitomizes one blatant scientific stereotype; more subtle attacks on science come from social critics.

BAD SCIENCE, BAD EDUCATION

Review by Douglas R. O. Morrison

Connected Knowledge: Science, Philosophy, and Education
BY ALAN CROMER
Oxford University Press, New York, 1997 ($25)

Yes, We Have No Neutrons: An Eye-Opening Tour through the Twists and Turns of Bad Science
BY A. K. DEWDNEY
John Wiley & Sons, New York, 1997 ($22.95)

I began to wonder some years ago why my children were learning science in such a crazy fashion. Teachers told them to do lab experiments but gave them no textbooks or notes to explain why they were doing those experiments or what they meant—evidently, the students were supposed to work it all out for themselves. At a P.T.A. meeting, I protested and was told that this was the new fashion in education. None of the other parents, I was informed, had made any complaint, except the ones who were scientists. This circumstance seemed to me to indicate a problem.

Most scientists have never heard of the "Science Wars"; they are too busy working to worry about how sociologists think their enterprise progresses. But it is becoming increasingly common knowledge that a harmful vision of science has been steadily taking over education in schools and universities. I only began to understand what was happening from an article by two deep-thinking physicists, Kurt Gottfried of Cornell University and Kenneth G. Wilson of Ohio State University, that was published this past spring in *Nature.* The two expressed concern that social scientists think scientific knowledge is merely a system of belief. This interpretation would imply that science is a subjective human construction, like art or music.

Many scientists do in fact underestimate how subjective their work is. In one briefly famous—and eventually disproved—instance, a group of researchers claimed to have discovered a heavy neutrino having a mass of 17 keV (17,000 electron volts). Such a particle would have profound implications for cosmology and particle physics. Several subsequent experiments claimed to confirm the initial observation; all purported to find a neutrino mass of 17 keV and not 16, 18 or any other number. Further investigation showed the heavy neutrino to be an illusion, however. The repeated appearance of the number 17 in the experiments proved only the power of expectation. But the fact that science's social critics have a point does not justify their broader denial of the vast body of reproducible experiments linked to well-tested theories.

It is difficult to speak uniformly about the so-called sociology of scientific knowledge because it embraces a number of schools of thought that adopt varying opinions about the external validity of such knowledge. Perhaps the most hard-line of these schools, according to Gottfried and Wilson, is based at the University of Edinburgh and, in their words, "contends that scientific knowledge is only a communal belief system with a dubious grasp on reality."

But the Edinburgh academics are not alone. Andrew Pickering, a sociologist at the University of Illinois, writes in his book *Constructing Quarks* that "there is no obligation upon anyone framing a view of the world, to take account of what 20th-century science has to say." These ideas may indeed seem incredible to a scientist.

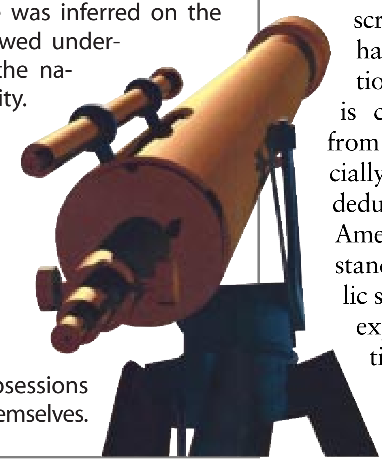
In *Connected Knowledge,* Alan Cromer, a professor of physics at Northeastern University, kindly guides the reader along through the social readings of science. He explains methodically the difference between science and a belief system and gives many examples, well chosen and carefully but interestingly explained. He also emphasizes that some parts of science can be considered complete: there are no new stable elements to be found or new continents to be discovered. Cromer points out that repeatability is taken as the essence of scientific knowledge, even though it is ignored or devalued by the social critics.

BRIEFLY NOTED

MARIHUANA, THE FORBIDDEN MEDICINE, *by Lester Grinspoon and James B. Bakalar. Yale University Press, New Haven, Conn., 1997 (paperbound, $16).*
Marijuana has been proved to lower ocular pressure in intractable glaucoma, relieve pain, reduce nausea and temper muscle spasms. It is also illegal, of course, an obstacle that has effectively barred doctors from a serious exploration of its therapeutic effects. Lester Grinspoon and James B. Bakalar marshal the remarkably voluminous evidence on marijuana's medical applications (and its widespread use) and on the policy decisions that keep it from being prescribed openly.

THE TRUTH OF SCIENCE: PHYSICAL THEORIES AND REALITY, *by Roger G. Newton. Harvard University Press, Cambridge, Mass., 1997 ($27).*
The quasireligious fervor of Roger G. Newton's arguments unintentionally appears to prove the point of the postmodernists who claim science as a human, subjective enterprise. The author's poorly examined, shifting premises and ad hominem attacks do little to advance his plea for the existence of an objective scientific "truth."

IN SEARCH OF PLANET VULCAN: THE GHOST IN NEWTON'S CLOCKWORK UNIVERSE, *by Richard Baum and William Sheehan. Plenum Press, New York, 1997 ($28.95).*
For nearly a century, mythology and mathematical calculations led some of the world's leading scientists to hunt for Vulcan, a hypothetical planet orbiting closer to the sun than Mercury. After years of fruitless investigations, Albert Einstein finally showed that Vulcan was an illusion: the planet's existence was inferred on the basis of a flawed understanding of the nature of gravity. This account efficiently reconstructs the events surrounding the search and exposes the ways in which misguided scientific obsessions can sustain themselves.



JULIAN BAUM

THE ILLUSTRATED PAGE

Alexander Graham Bell:
The Life and Times of the Man Who Invented the Telephone
BY EDWIN S. GROSVENOR AND MORGAN WESSON
Harry N. Abrams, New York, 1997 ($45)

Bell's trophy-winning June Bug

A. M. BENNER *Courtesy of the Grosvenor Collection*

Although his name will be forever linked with the telephone, Alexander Graham Bell was a wide-ranging inventor whose brain dreamed up an early metal detector, a respirator and various flying devices (including one, shown above, that competed for an aviation trophy offered by Scientific American in 1908). With its extensive quotations and archival photographs, this volume brings Bell's creativity boldly to life. *—Corey S. Powell*

He argues that science itself is the connection between theory and experiment.

Cromer gradually compares science and its methodology with the ideas of the "postmodernists," who question the objectivity of science and even the existence of objective reality. What I found particularly worrying in this section of the narrative was the author's description of how postmodernists have applied their ideas to education. In that arena, the movement is called constructivism, derived from the notion that all facts are socially constructed rather than being deduced from evidence. I often hear American scientists lament the low standard of education in their public schools. After reading Cromer's explanation of how constructivists have worked their ideas into science teaching programs and introduced their nonscientific ideas, I can well understand how these actions have exacerbated the problems.

Cromer also offers positive suggestions. He has been involved in school and adult education and in teaching in "Boston's most costly school, the Suffolk County House of Corrections"; in these capacities, Cromer has been able to put his thoughts into action. His work with Project SEED—an experimental, 13-day educational workshop—shows that a basic approach involving hands-on activity, steadily increasing complexity and a constant reference to quantification can be successful.

Cromer's wide-ranging book may be considered as a correction to excessive belief in the corrosive movements that he lists: "constructivism, postmodernism, multiculturalism, radical feminism, ecoradicalism and political correctness." Although such beliefs are welcome and even useful components of a free, dynamical society, Cromer's arguments drive home the ultimate importance of objective facts.

A. K. Dewdney's book is quite different, and the contrast between the two books is instructive. *Yes, We Have No Neutrons* shows how researchers can, and often do, go awry when they lose sight of science's guiding principles. The book's clever title refers to the embarrassing saga of cold fusion (researchers Martin Fleischmann and Stanley Pons, now with Toyota's research lab near Nice, France, experienced great difficulty in finding the neutrons that would be produced in abundance in a real fusion reaction). The subtitle, "An Eye-Opening Tour through the Twists and Turns of Bad Science," intriguingly describes the structure of the book. There are eight chapters, each telling a story of poor science; these are linked together by a clear introduction but otherwise are fairly independent.

Unfortunately, I was disappointed by the chapters that cover the two episodes of bad science that I have studied most seriously: N-rays and cold fusion. In both cases, Dewdney's accounts gave the impression that he had read a few review papers and distilled them without arriving at any deep understanding. In addition, the chapters contain factual errors. Dewdney has produced an easily readable account that gives the essentials in a racy style—although the minds and feelings of the people involved seem to have been interpreted rather freely with phrases such as "Blondlot sighed" and "Jones... boggled."

I was surprised to see that the sad and revealing story of neutrons claimed from cold fusion is barely described and its essence omitted. In fact, at a press conference in Utah on March 23, 1989, Fleischmann and Pons released a beautifully clear plot showing that their "cold fusion" experiment produced gamma rays peaking at 2.5 MeV (million electron volts). That energy is exactly the value one would expect for the fast neutrons produced by deuterium fusion when they interact with surrounding water molecules, producing gamma rays.

On March 28 Fleischmann presented the same plot at a scientific meeting at Harwell, England, but this time the audience objected. Fast neutrons, a detractor pointed out, hardly interact with water. Such interactions happen mainly when neutrons have slowed down almost to rest and are captured. The gamma rays emitted under these circumstances actually peak at 2.2 MeV. So two days later at another meeting in Lausanne, France, and again the following day at CERN near Geneva, Fleischmann showed a very similar plot—but the peak had suddenly moved to 2.2 MeV!

Also at Harwell, Fleischmann had been told that his gamma-ray peak was impossibly narrow; nature would not be so tidy. The new graph, released two days later, was twice as wide. For good measure, the peak number of events indicated in the plot had jumped from 2,000 to 20,000, even though the raw data I saw still showed just 2,000 events. Dewdney's account seems to end years ago; more recent information may be obtained on SCIENTIFIC AMERICAN's World Wide Web site in the physics area of "Ask the Experts" (www.sciam.com/askexpert/).

Other chapters in *Yes, We Have No Neutrons* stick more closely to Dewdney's specialties of mathematics and computers. These sections seem clearly

KENNETH CHAMPLIN

written and explained. In particular, his exposition on the unreliability of IQ tests and subsequent demolition of the book *The Bell Curve* are well done.

Dewdney's account of the scientific failings of Sigmund Freud, which I had not known before, horrified me. He writes that Freud had published only six case histories as a foundation for his enormous output of theoretical work. Worse yet, Dewdney claims that those accounts were not satisfactory justifications of Freud's conclusions. For example, in the Wolf Man case, Freud interpreted a dream to mean that as a child, the Wolf Man had interrupted his parents making love. Much later the Wolf Man himself said that Freud's interpretation was "all false"; he could not have witnessed the scene Freud posited, because in Russia "children slept in their nanny's bedroom, not their parents'."

Reading this chapter, one is strongly inclined to ask, "Was Freud a fraud?" Dewdney does not really provide enough information to answer this question, however. To make a real judgment of Freud's work, one would need to study all his writings. No matter the quality of his evidence, there can be no doubt that Freud created a profession and changed the tenor of medical practice, causing doctors to listen more carefully to their patients. His claim to fame is evident, although it is interesting to consider how little of it may be based on science.

As scientists like myself prepare our counterattack, the Science Wars are heating up. Anyone with an interest in the future of science should make haste to learn the order of battle. Cromer's *Connected Knowledge* should be required reading for people involved in teaching. Dewdney's book, in contrast, is a useful and easily read introduction to bad science and should sell well to the general public, who will enjoy the stories and will learn from them. For readers of *Scientific American,* some of the topics may be covered at too superficial a level, but the book should be considered as a helpful eye-opener to subjects with which they are unfamiliar. Beset by bad science or no science at all, scientists and their supporters must take a stand.

DOUGLAS R. O. MORRISON was a researcher at CERN, the European laboratory for particle physics near Geneva, for 38 years.

SINISTER SCIENCE

Review by Rebecca Zacks

The Death of Innocents: A True Story of Murder, Medicine, and High-Stakes Science
BY RICHARD FIRSTMAN AND JAMIE TALAN
Bantam Books, New York, 1997 ($24.95)

For reasons that remain inscrutable to doctors, babies sometimes die silently and unexpectedly in their cribs, victims of sudden infant death syndrome (SIDS). For reasons that seem unfathomable to most of us, parents sometimes murder their own infants and try to pass the deaths off as natural or accidental tragedies. In a small number of cases, the former may serve as a cover for the latter.

As journalists Richard Firstman and Jamie Talan reveal in this riveting account, that is exactly what happened in June 1970 and again a year later when Waneta Hoyt, a young mother in upstate New York, smothered her fourth and fifth children. Molly and Noah Hoyt, both less than three months old at the time of death, had spent most of their lives in a hospital clinic attached to machines that monitored their every breath. Waneta had already lost three children to mysterious causes, and so the babies were at risk for SIDS, her doctor believed. The last two deaths seemed only to confirm his suspicion, despite the odd circumstance that all Waneta's children died at home while alone with her. Autopsies on the two infants were inconclusive (no surprise: even under the pathologist's discerning gaze an intentionally smothered baby often looks no different from one who has stopped breathing for less sinister causes). The doctors attributed Molly's and Noah's deaths to SIDS, and their mother walked free for more than two decades thereafter.

The Hoyts' story is the thread that ties together a stunning examination of the interplay between criminal justice and medical research in the emotionally charged world of SIDS. In 1972 Alfred Steinschneider, the physician who had studied the Hoyt babies before their deaths, published a paper in the prestigious journal *Pediatrics.* Based on his observations of the infants, Steinschneider advanced what would become known as the apnea theory of SIDS. In essence, it suggested that during sleep some babies stop breathing for abnormally long periods (some short pauses are common) that occasionally prove fatal. Steinschneider further proposed that SIDS might run in families and that potential SIDS victims could be identified and protected if their breathing were carefully monitored.

Firstman and Talan piece together the genesis and impact of the apnea theory, drawing on thousands of pages of medical records, legal documents and scientific publications, along with interviews with more than 300 people. Their dissection of the 1972 paper describes Steinschneider's data as shaky in some places and blatantly inaccurate in others. The authors also examine the scientific and cultural conditions that favored the apnea theory. Despite its serious flaws, Steinschneider's paper shaped research agendas and popular beliefs and launched a lucrative business in home apnea monitors: by 1990 manufacturers were pulling in $40 million annually.

PAUL KENNEDY *Liaison International*

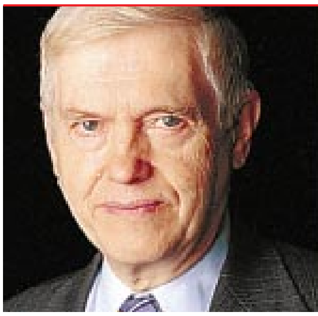
***SLEEPING NEWBORN BABIES** may have little to fear from apnea.*

Though more than two decades of subsequent research failed to support Steinschneider's assertions, Firstman and Talan write, those assertions continued to crop up in the courtroom. During those years, an attorney defending a parent accused of the serial murder of his or her children was likely to point to the paper as proof that SIDS can be familial. As former Dallas medical examiner Linda Norton explained in 1985 to a district attorney prosecuting yet another case of serial infanticide, "It's *the* defense in cases like this. It's apnea. It's SIDS. It runs in families."

Norton's testimony in 1986 helped the D.A. win a conviction. Moreover, her frustration and passion inspired that D.A. to begin digging into the Hoyt case. Science was being used to subvert justice; as Norton told the authors, "serial SIDS more than likely meant serial homicide." Some researchers estimate that 5 to 10 percent of the deaths attributed to SIDS may in fact be homicides, Firstman and Talan note.

A powerfully told detective story, *The Death of Innocents* ends with little sense of resolution. In 1995 Waneta Hoyt was finally sentenced to 75 years to life imprisonment for the murders of five of her children. Yet Steinschneider declined to write a correction to his 1972 paper. The same apnea theory that has been used as a cover for infanticide still occasionally finds its way into pamphlets, newspapers and—frighteningly—medical school classrooms.

REBECCA ZACKS is a science writer based in Boston.



WONDERS

by Owen Gingerich

In Praise of Fakes

"If it's a fake," Miriam declared, "it's worth the price as a piece of fine art."

Several years ago I became curious about a 17th-century Persian astrolabe maker called Abd al A'imma the Younger. My trail of research eventually led to the Boston Museum of Fine Arts, which had a brass instrument supposedly inscribed by him. The metalworking and calligraphy were fine enough, but it quickly became apparent to me that the astrolabe never could have functioned as an astronomical instrument. In a word, it was a fake.

The curator's response took me by surprise: "This will make a good display for us. The museum is planning an exhibition on forgeries. We have two astrolabes that look very much the same to the untutored eye, but one is genuine, the other fake. The comparison will help teach what is essential on an astrolabe and what is ornamental nonsense."

Unfortunately, the exhibition never came to be. The Boston trustees vetoed the idea because they didn't want to admit how often the museum had been caught out collecting forgeries. Yet later, in 1990, the British Museum staged a stunningly successful show of fakes—ranging from a medieval chastity belt to the infamous jaw of Piltdown Man. These artifacts, testimonials to deceit and gullibility, all raise the sensitivity of our perceptions. They teach us to see more critically as we (and fallible experts) gain greater insight into the often blurred line between the authentic and the counterfeit.

Yale's Vinland map, the Getty's bronze horse, Rembrandt's *Man in a Golden Helmet*—all of these challenged objects have attracted greater scrutiny and study than they ever would have had as complacently accepted antiques.

Recently my wife, Miriam, and I had our vision sharpened in an unanticipated arena. Ever since we had a glorious opportunity to sail in Melanesia during the 1986 appearance of Halley's comet, we have been intrigued by the astonishing variety and beauty of seashells. Over the years we have amassed a sizable collection of cowries, cones and conches. The invention of scuba diving—and the knowledge by fisherman that shells are eminently collectible—substantially democratized some once classic rarities. An example is the elegant Lister's conch, which fetched $1,000 at auction in 1970 but today is available for a few dollars. And the "matchless" cone *Conus cedonulli,* which brought six times the price of a Vermeer painting at a 1796 auction, can now be purchased for about $100. Still, many spectacular gastropods outstrip our modest budget.

We were therefore quite surprised to discover in a small shop several exquisite shells at bargain prices. Puzzled, I inquired about them. "We got them inexpensively from the fisherman, so they are good buys," the dealer explained.

Blissfully forgetting the adage that where money is to be made, forgeries abound, I succumbed to temptation, buying two beautiful cowries—a *valentia* and a smaller *sakurai,* the latter a shell at the top of the rarity scale. When I got back to my reference books, I found that the *sakurai* cowrie was a twin of the one illustrated in *The Shells of the Philippines.* That should have triggered an alarm, but it didn't.

I could not help gloating over our acquisitions, so I mentioned them via e-mail to Guido Poppe, one of the leading collector-dealers in Europe. Poppe promptly congratulated us on our *bonne chance,* then dropped his bombshell: "I hope you didn't buy painted specimens."

Shaken, we consulted *A Guide to Worldwide Cowries* to see if there were any similar but cheaper species that could be repainted into a rarity. This was an educational experience in itself, taking a close look at the shapes rather than the color patterns of the cowries: the bottom line was that both species had unique shapes.

Placing the specimens under magnification—10 to 30 power—revealed a wealth of detail, but it still left their status ambiguous. The single-colored *sakurai* pattern looked suspiciously like delicate penmanship. To get an expert opinion, I sent the shell to Gary Rosenberg of the Academy of Natural Sciences in Philadelphia. Although the academy's shell collection contains 12 million specimens, the *sakurai* is so rare that it is not represented in their holdings. Nevertheless, Dr. Rosenberg could give a definitive verdict.

"When I opened the box," he reported, "I thought, 'What a gorgeous shell!' But under the microscope, I saw that all the color lay sharply in the same plane. On a genuine pattern, the animal deposits the pigments all through the outer layer, producing a somewhat fuzzy appearance. Furthermore, the surface of your shell is too uniform, so a reflected light stays steady as the shell is rotated. On an unretouched shell, the growth lamellae give a jerky effect. Your shell is a genuine *sakurai,* but its splendid pattern is a forgery."

The fraudulent *sakurai* cast doubt on the larger *valentia.* The *valentia*'s pattern, however, was far more complex, with multiple layers and a palette of subtle colors. "If *it's* a fake," Miriam declared, "it's worth the price as a piece of fine art." An art historian made the next suggestion: "Take it to the physical conservation laboratory at the Fogg Art Museum. They can use infrared to see if anything has been repainted."

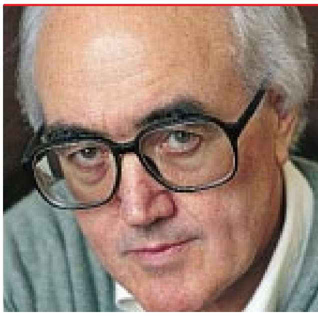
The conservators chuckled at that advice. Infrared is useful in some situations to detect overpainting on canvases but not appropriate here. In any case, the museum's experts gave the *valentia* a hard look under their microscope: "There are no signs of brushstrokes or edge bleeding that we would expect with painting on porcelain, for example. If this is a fake, it's a much better job than we can do in the lab." Still, they agreed that one minor blemish on the surface looked like a fingerprint in lacquer. Determined to get a definitive answer, I gave permission to dissolve the finish in some inconspicuous spot.

When I returned the next day, they were wreathed in smiles. "We've tried every solvent in the cabinet," they said. "Nothing touches it. Your shell has a natural surface, and that 'fingerprint' has got to be a natural defect."

As a final clincher, they suggested putting the shell under ultraviolet light. I knew the result already. Most shells, including the *valentia* and 180 other species of cowries that I have tested, have no ultraviolet features. The two exceptions are the *venusta* cowrie from southwest Australia and the relatively common *mappa,* found throughout almost the entire Indo-Pacific. Both species fluoresce with a magnificent orange. My discovery is well known to malacologists, although I'm not sure they know why these species glow that way.

In terms of sharpened perspectives, the money for the "enhanced" specimen was well spent. The happy ending of my story is that the dealer who sold the shells was as surprised as I when one turned out to be fraudulent, and she immediately exchanged the *sakurai* for a real one (which was not so easy, considering its rarity). And then she sold me the fake at a reduced price. SA

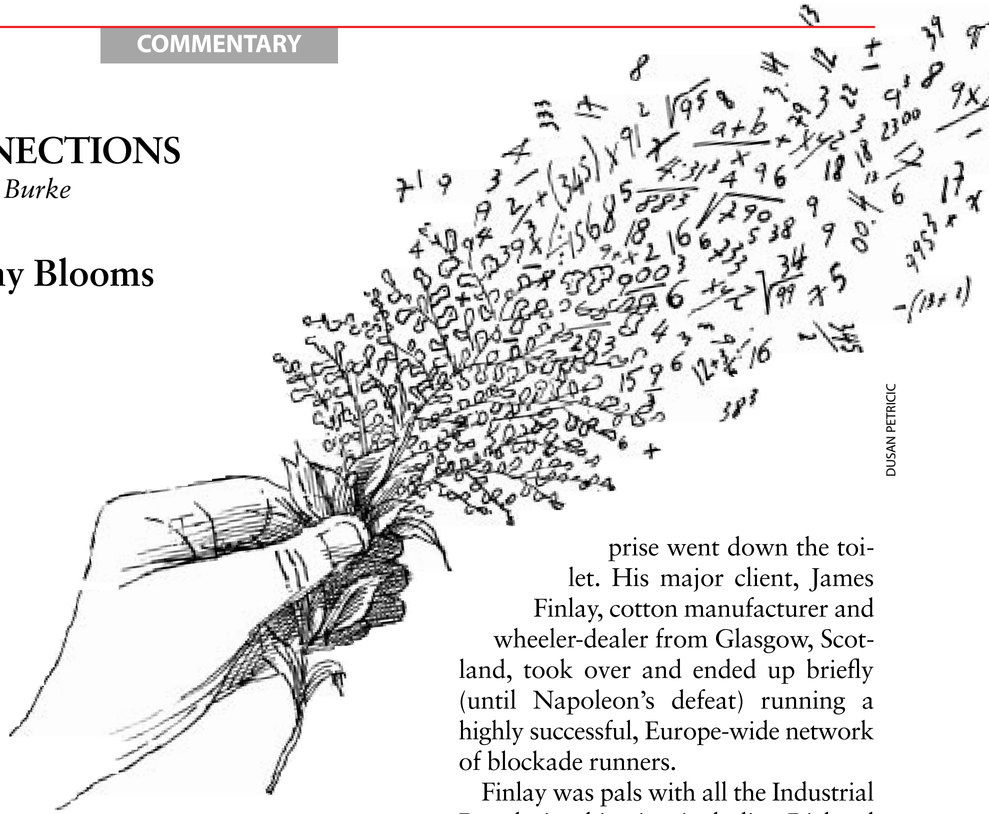
OWEN GINGERICH, professor of astronomy and history of science at the Harvard-Smithsonian Center for Astrophysics, has a "collecting gene" that gives him a passionate interest in rare books and elegant shells.



CONNECTIONS

by James Burke

Healthy Blooms



DUSAN PETRICIC

I'll risk a bet. You (like me) didn't know that the common lilac begins to flower when the sum of the squares of the mean daily temperatures (Celsius) since the end of the previous frost adds up to 4,264. This piece of mind-boggling botanical trivia sprang fully formed from the brain of Belgian astronomer and mathematician Adolphe Quetelet, whose obsession with numbers led him also to invent, in 1835, a concept that I'll bet you *have* heard of: the average joe.

Quetelet gathered data on this individual's propensity to commit crimes, get drunk, marry, die, be tall, commit suicide and so on. In the end, he uncovered so many regularities in the figures, he said, as to believe that there could be such a science as "social physics," which would put the analysis of behavior onto a mathematical basis.

Quetelet had taken some earlier thoughts on this matter to an 1833 meeting of the British Association for the Advancement of Science in Cambridge, where he persuaded other like-minded noodlers to set up the Statistical Society of London and so further the cause. At this time, British interest in social analysis of all kinds was at panic level, because living conditions in the overcrowded industrial cities had brought the laboring classes close to revolution. Statistics were soon avidly sought on such essential data as how many ragged families could sing a jolly song, how many starving mothers could knit, which filthy hovels sported morally improving prints on their walls, et cetera.

First prez of the London Statistical Society was one of the nerds Quetelet had met in Cambridge, a man with more ideas than time to do much about them. His notions included a speaking tube from London to Liverpool and an automated version of tic-tac-toe. Charles Babbage, the prop-head in question, was a great mathematician of his time, and I suppose that's why, among many pursuits, he *did* find time to invent shoes for walking on water. Most of his life, however, was taken up with trying to raise money to build two geared calculating machines of such complexity that he never built them. One used punched cards and had stored programs, and that's all I'll say about that. Ada, Countess Lovelace, said more than enough for all of us.

She was Babbage's aristocratic patron and promoter and introduced him to all the right people, in return (they say) for Babbage's providing her with a betting system. Like his machines, it never worked and caused a scandal. Rather like Ada's short-lived father, Lord Byron. Who spent much of his adulthood traveling around the eastern Mediterranean, where, in 1809, he met up with an odd cove called John Galt, who was trying to set up a grand international scam.

At the time, Napoleon's continental blockade was ruining the U.K. export industry, so Galt's idea was to sneak British manufactures through Istanbul and into Europe by the back door, over the Hungarian border. Almost as soon as it started, Galt's entire shaky enterprise went down the toilet. His major client, James Finlay, cotton manufacturer and wheeler-dealer from Glasgow, Scotland, took over and ended up briefly (until Napoleon's defeat) running a highly successful, Europe-wide network of blockade runners.

Finlay was pals with all the Industrial Revolution bigwigs, including Richard Arkwright. In 1771 Arkwright's water frame had turned the cotton industry from a cottage system for piecework into factory mass production. A single-power source (water) turned hundreds of rollers and spindles that pulled out the thread and then twisted and wound it, ready for use on looms. Five years later came the patent on a single-power source (steam) that would run Arkwright's machine and any other you could think of. James Watt's steam engine was so popular that he couldn't keep up with the paperwork, so he next invented a copying machine. Writing (or any design to be copied) was done on paper with a special ink whose ingredients included gum arabic. The completed original was then rolled against wet paper, on which the copy would appear (and last for 24 hours).

In 1823 Cyrus Dalkin of Concord, Mass., improved on the idea by coating one side of a sheet of paper with paraffin wax and carbon black. Pressing onto the sheet transferred a copy to the paper beneath. Dalkin called the product "carbon paper" and sold it to the Associated Press. In 1868 the AP covered a balloon ascent by Lebbeus Rogers (a biscuit maker). Rogers was in the AP office being interviewed when he saw Dalkin's paper at work. Instantly quitting biscuits and balloons, Rogers went into the carbon-paper business. In 1873 he attended a demonstration of the amazing new typewriter, where he persuaded the typist to try one of his carbon sheets. And the rest is history (unusually, repeating).

The typewriter Rogers saw had been manufactured by the gunmaker E. Remington, because the company had spare capacity and the kind of machine tools that would make the bits. There was little demand for the bits Remington had previously made, once the Civil War had ended, and there was a catastrophic drop in the demand for guns. The Remington had been one of the most successful guns ever made, rivaled only in sales volume by Sam Colt. Who made revolvers because his mines failed him. This may have been because in 1844, after he had successfully mined a ship on the Potomac River, at a distance of five miles, he wouldn't give the navy the secret, and so they wouldn't give him the money.

Immanuel Nobel was a great deal more open about *his* mining techniques when the Russians asked him to make mines for them. By the time the Crimean War started, in 1853, "Colonel Ogarev's and Mr. Nobel's Chartered Mechanical and Pig Iron Foundry" had been laying mines everywhere around Russia for 12 years. One place they'd sown up in this way was the harbor at Sevastopol. So the Allied Fleet supporting the troops in the Crimea was forced to anchor around the corner at Balaklava, where it was destroyed by the full force of the hurricane of November 13. A 7,000-ton cargo of medical supplies and clothing went down to the bottom, leaving British troops to suffer a terrible winter of pneumonia, starvation and dysentery.

One thing he did *find time to invent was shoes for walking on water.*

One week earlier an extraordinary woman named Florence Nightingale had arrived in the Crimea. She and the 38 other nurses accompanying her spent that dire winter discovering how bad British army medical services really were. She'd heard a few rumors: as the best means of warding off disease, British military doctors recommended smoking or else growing a mustache (to filter out germs). In one recovery area, 1,000 men suffering from diarrhea shared 20 chamber pots. In the hospitals the patients underwent operations on floors covered with blood. Wounds were often not dressed for five weeks. The hospital mortality rate in most cases reached close to 50 percent. By the end of the war, of the 18,058 British casualties, nine out of 10 had died from disease. When these facts hit the newspapers back home, everything hit the fan. Thanks to Nightingale's 1,000-page report, filled with horrifying detail, the Crimean War marked a turning point in military medicine.

Nightingale's obsession with statistics had started with a keen interest in botany. And it was while she was doing some botanical classification work that she had come across a statistical law that tickled her fancy and led to her striking up a lifelong friendship with its discoverer. It was Quetelet's law about flowering lilacs. Hope all this has planted a few useful thoughts. SA

WORKING KNOWLEDGE

ACTIVE-MATRIX LIQUID-CRYSTAL DISPLAYS

by Samuel Musa
Executive Director
Center for Display Technology
and Manufacturing
University of Michigan

Active-matrix liquid-crystal displays are standard on most new laptop computers. Two properties of the organic fluids called liquid crystals suit them for use as tiny switches turning picture elements (pixels) off and on. First, the crystals are transparent but can alter the orientation of polarized light passing through them. Second, the alignment of their molecules (and their polarization properties) can be changed by applying an electrical field.

In a color display the liquid crystals are held between two glass plates, the outsides of which have been coated with polarizing filters. Only light with a particular polarization can pass through these filters (*a*). Inside the plates are sheets of thin-film transparent electrodes and color filters, which form very small picture element regions called subpixels. A grouping of a red, a green and a blue subpixel forms a full-color picture element, or pixel. The combined activity of the subpixels defines the color that the pixel transmits.

Fluorescent backlighting illuminates a display from the rear. In pixels that are off, light passes through the rear polarizing filter, the liquid crystals (*b*) and the color filters, only to be blocked (absorbed) by the front polarizing filter. To the eye, these pixels appear dark. When a pixel is turned on, the liquid crystals reorient their position, and they in turn repolarize the light so that it can pass through the front polarizing filter (*c*).

The active matrix provides a superior method of electronically addressing (turning on) an array of pixels. For an image to appear on the screen, one row of pixels receives the appropriate voltage. At the same time, software in the computer dictates that voltage be applied to those columns holding active subpixels. Where an activated row and column intersect, a transistor turns on a subpixel electrode, generating an electrical field that controls the orientation of the liquid crystals. This process repeats sequentially for each of the 1,280 rows in an advanced display, which can take 16 to 33 milliseconds.

Liquid-crystal displays can provide increasingly better image resolution by raising the density of pixels and by refreshing the screen image at ever faster rates. The bright, sharp color images of the latest generation of flat-panel displays only serve to illuminate the remarkable properties of liquid crystals.