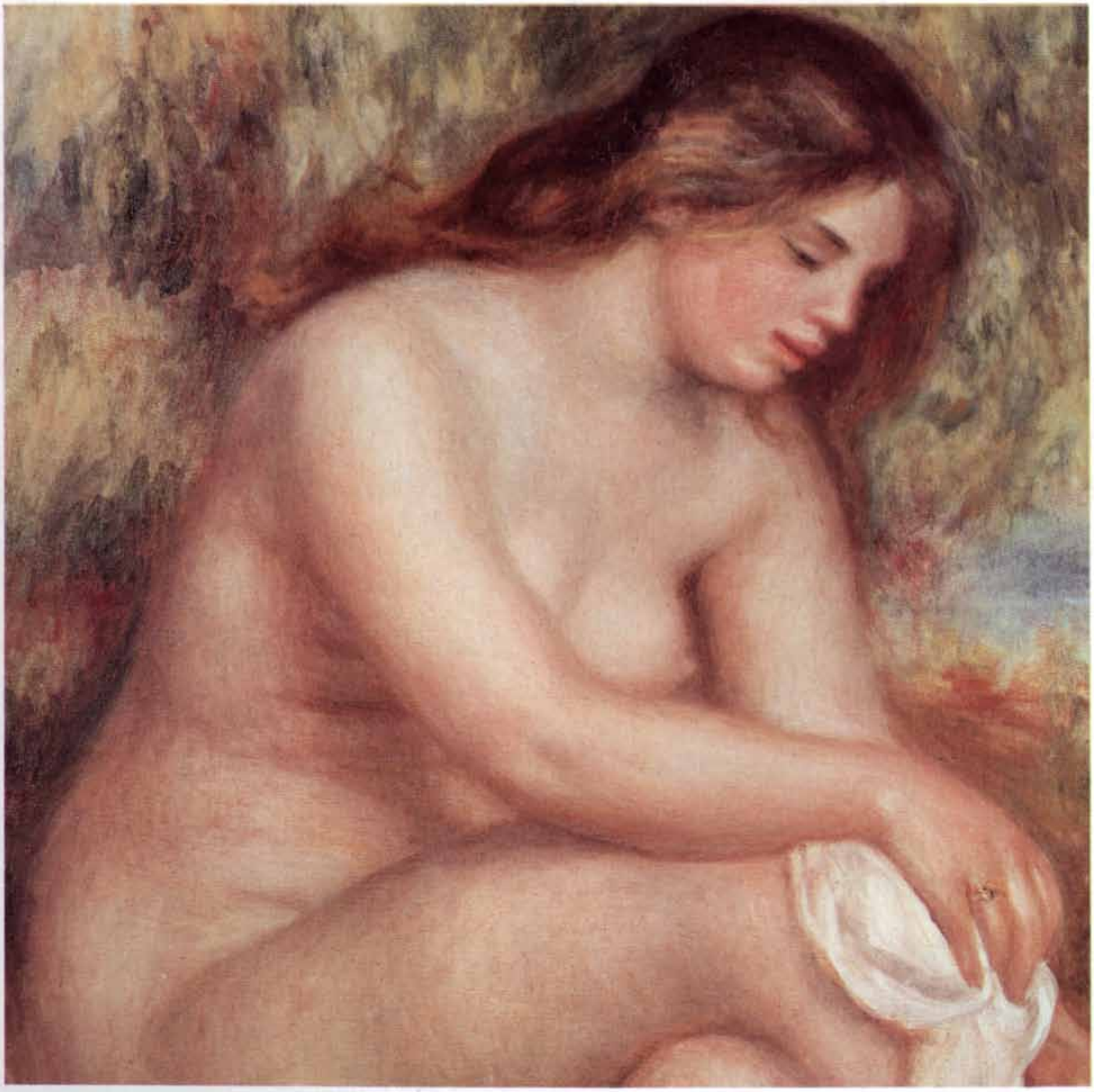
SCIENTIFIC AMERICAN

MARCH 1988
$3.50

Falling antimatter could overturn physics.
Can antibodies serve as made-to-order catalysts?
Beyond gigascale integration: the quantum semiconductor device?



Renoir's Seated Bather *would probably have been fertile: a critical ratio of fat to lean mass is important to fecundity.*

Don't drink and drive.

Join the American

Golf In America, there is a society of drivers who take joy in the knowledge that their cars can do what cars do on the German Autobahn.

Go quickly, confidently and so comfortably, the last kilometer of a journey ends up being as enjoyable as the first.

Their natural habitat is a Volkswagen. Like the Golf. Over the last three years, the best-selling nameplate in Europe.

The American Autobahn society likes it for its power, quick reflexes and substantial interior space.

The Golf comes with a lively 1.8-liter

Sales claim based on manufacturers' reported retail deliveries for 1985, 1986, and 1987 through August.

Autobahn society.

overhead cam engine. With 17% more power than before. Thanks to our new Digifant fuel-injection system.

For crisp handling

fort, the Golf has our highly refined 4-wheel independent suspension.

It seats five nicely. With the back seat down, it holds 39.6 cubic feet of cargo.

To join the American Autobahn society just stop by your local chapter. Your Volkswagen dealer.

German engineering.
The Volkswagen way.

SCIENTIFIC AMERICAN

March 1988 Volume 258 Number 3

DEPARTMENTS

THE COVER image, a detail of Auguste Renoir's *Seated Bather Drying Her Leg,* is in the tradition of art that links fatness with feminine beauty, a connection that turns out to make some biological sense (see "Fatness and Fertility," by Rose E. Frisch, page 88). Women must store a reasonable amount of body fat to be capable of reproducing. The fat apparently has a regulatory role and is important to the onset and the maintenance of normal menstrual cycles.

THE ILLUSTRATIONS

Cover painting from São Paulo Museum of Art (Giraudon/Art Resource)

Page	Source
42–45	Ian Worpole
46	Jet Propulsion Laboratory
47	Ian Worpole
48–49	La Specola Museum of Zoology, Florence (Scala/Art Resource)
50	Lawrence Berkeley Laboratory
51–56	George Retseck
59	Arthur J. Olson
60	Elizabeth D. Getzoff and John A. Tainer
65	Elizabeth D. Getzoff and John A. Tainer (*top*), Michael E. Pique, Elizabeth D. Getzoff and John A. Tainer (*bottom*)
66–67	Andrew Christie
68	Andrew Christie (*top*), Arthur J. Olson (*bottom*)
69	Andrew Christie
70	Michael E. Pique
73	Laurie Burnham
74–79	Hank Iken
81	Bruce Coleman Inc./ G. Harrison
82	Patricia J. Wynne
83	Animals Animals (*left*), Patricia J. Wynne (*right*)
84	James D. Murray (*top*), Patricia J. Wynne (*bottom*)
85	Avi Baron and Paul Munro (*top*), Patricia J. Wynne (*bottom*)
86	Bruce Coleman Inc./ Hans Reinhard (*top left*), Animals Animals (*top right*), Patricia J. Wynne (*bottom*)
87	Charles M. Vest and Youren Xu
89	Naturhistorisches Museum, Vienna
90–91	Bob Conrad
92–93	Carol Donner
95	Peter C. Ortner
97–100	George V. Kelvin
103	Anna Marguerite McCann
104–105	Tom Prentiss
106	Anna Marguerite McCann
107	Tom Prentiss (*top*); Barbara Bini, American Academy in Rome (*bottom*)
108	Tom Prentiss
109	John P. Oleson
110–113	Michael Goodman
114	James F. Blinn
115–116	Edward Bell
117	James E. Loyless

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Subscription rate: $24 per year in the U.S. and its possessions; $33 per year in all other countries. Postmaster: Send address changes to Scientific American, P.O. Box 953, Farmingdale, N.Y. 11737-0000.

SCIENTIFIC AMERICAN
Established 1845

We pay these people to fly TWA.

Introducing TWA's Quality Control Team.

Since many of our best ideas about improving service have come from our passengers, we thought, why not take this process a step further?

TWA's Quality Control Team is really a group of 30 "professional" passengers. Their job is to fly TWA, and report directly to the Chairman what they liked, and what they didn't like. They have a checklist of over 100 service items to go through, analyzing everything from the temperature of the entree to the cleanliness of our aircraft.

Introducing TWA's In-Flight Service Manager.

Recently, TWA has instituted an In-Flight Service Manager on every flight who is responsible for all aspects of TWA's in-flight service. You have a question about a connection? They'll answer it. You need special assistance? They'll take care of it. You've seen the movie three times already? They'll find you a great magazine.

You get the idea.

The Chairman of TWA Has Made Service Top Priority.

TWA has always recognized that good service makes all the difference in air travel; so we're determined to ensure that ours is the best in the business. That's why the Chairman of TWA has made service top priority. In fact, if you have any constructive criticism, major gripe, or just lavish praise, he'd like to know about it.

You can write him at TWA, 605 Third Avenue, New York, NY 10158.

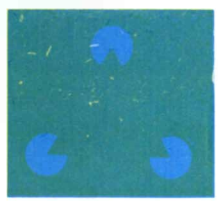
TODAY'S TWA.
FIND OUT HOW GOOD WE REALLY ARE.

TWA

LETTERS

To the Editors:

If I had known that the topic of January's "The Amateur Scientist" was to be subjective contours, I would have included the following observation in my article, "Art, Illusion and the Visual System," in the same issue. David Hubel and I have done some experiments indicating that the experience of illusory borders is not cognitive. We made Kanizsa figures (figures that produce the impression of borders where there are none) in two colors instead of in black and white. We found that when the two colors are equiluminant, or of the same brightness, the illusory borders vanish, even though the nonillusory parts of the figures are still perfectly clear. Hence if the three shapes below are blue on a green background, you do not see any triangle when the colors are equiluminant.



That is, at equiluminance you see quite distinctly three circles with wedges cut out, but you have no impression of a triangle lying in front of the circles. The level in the visual system where illusory borders are generated cannot be so high as to be considered "cognitive," because your conscious impression of the nonillusory parts of the figure is no different at equiluminance and nonequiluminance.

We have interpreted this result as meaning that of the two major subdivisions of the human visual system, called "magno" and "parvo" because of the difference in the size of their cells, only the color-blind magno system tends to link colinear borders. This linking ability ordinarily functions in determining which borders in a scene belong to the same object, even when the borders are interrupted; illusory borders are simply a manifestation of this function. At nonequiluminance—for example, when the shapes are black on a white background—both visual-system subdivisions "see" the incomplete circles, but it is only the magno system that links the wedges together and signals the illusory figure. At equiluminance only the color-sensitive parvo system can still see the shapes; the figure as a whole is invisible to the color-blind magno system, and so it does not generate illusory borders.

MARGARET S. LIVINGSTONE

Harvard Medical School
Boston, Mass.

To the Editors:

"Modeling Tidal Power," by David A. Greenberg [SCIENTIFIC AMERICAN, November, 1987], will revive interest, which for some of us has never died, in power from the Bay of Fundy. A major barrier on a tidal estuary would undoubtedly have ecological effects, some of them unfavorable. If these effects are compared with the effects of alternative, reasonably economical means of producing a large amount of energy, however, the ecology of tidal power shines brightly.

The environmental effect that appears most likely to raise opposition—and cost—is an increase of 15 centimeters in the maximum tidal level at Boston. This is a serious problem, but it would arise only at times when, if the tidal barrier did not exist, the tides would be within 15 centimeters of historical flood levels. Such tides might occur as seldom as two or three days a year. The question I should like to raise is whether the natural state could be restored on those days simply by leaving the sluices in the barrier open. This would sacrifice generation of power, but the loss might amount to no more than 1 percent of the annual output.

Could an additional benefit of a tidal-power barrage in fact be regulation of the tidal resonance in the Bay of Fundy in order to reduce the disasters that now occur in the Gulf of Maine when high tides combine with unfavorable winds? (One can well imagine the lawsuits that could result from real or fancied mismanagement of this mighty engine!)

Another question arises concerning the author's statement that "two-way generation...would yield more power, but the greater complexity of design would make the power more expensive." The question is whether it would really be more expensive than replacement energy from some other source, not whether it would be more expensive than energy from ebb generation. When we speak of the expense of alternative energy sources, we should include such factors as acid rain and long-term security of energy supply. (On the other side of the coin, of course, is a tidal barrier's potential for displacing fisheries and bird sanctuaries.) It may be that the extra energy available from two-way generation would be a bargain compared with viable alternatives for the 21st century.

ROBERT S. SPROULE

Montreal, Canada

To the Editors:

Mr. Sproule touches on some interesting aspects of the tidal-power investigation. Some consideration has already been given to his points.

David De Wolfe of Discovery Consultants Ltd. in Wolfeville, Nova Scotia, has used the same tidal model described in the article to see whether operation of the barriers for a short time, in different ways, could reduce the resultant rise in high-water level. Limited testing indicated that if the sluices were operated specifically to suppress the high water, the increase from a tidal dam in Minas Basin could be reduced to a few centimeters at Boston, but not eliminated. Given this result, it seems too much to hope that barrier operation could lessen the combined effects of storm surges and high tides.

Even though two-way generation could give more energy only in limited circumstances, the increased flexibility of the energy supply it would provide might render it attractive. With two-way generation it is possible to even out the energy supply and to some extent produce energy on demand. For example, in winter energy needs peak in the early morning owing to heating and lighting demands; in summer they peak in the afternoon owing to air-conditioning loads. In the 12-hour period preceding peak need the basin could be drained or filled (depending on the timing of the tidal cycle) so that energy production could be maximized for these few hours of peak demand. This energy at peak load would be worth much more than the base-load energy that would otherwise be produced. To the best of my knowledge this flexibility has not yet been fully worked into existing calculations, but it will be considered if feasibility studies go much further. The model does predict that the environmental effects of one- and two-way generation would differ only slightly.

DAVID A. GREENBERG

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1938: "Life's darkest, most fascinating secrets are hidden in the mysteries of enzymes—now known to be huge and intricate molecules, somehow able to dominate the uncountable chemical activities within living flesh. Only because of the play of enzymes can we digest our foods, or find chemical value in our very breath. And the use of food and oxygen in the multivaried events of growth, movement and reproduction, and even sensation and our thoughts—all is unlocked by these really miraculous keys."

"With an electronic beam for its brush and a fluorescent screen for its canvas, electricity can paint images of its own minutest variations, its lightning-fast changes, and its most complex moods. Devoid of the moving parts and inherent inertia of even the most delicate meter, the cathode-ray tube has become a new and startling tool in the hands of scientists, engineers, industrialists, and even radio servicemen."

"The single-use type of package is being developed for a number of purposes. Beer cans made a spectacular entrée into the container field, once a satisfactory lacquer was found to keep the liquid from the metal. Paper, pulp, and transparent-film containers for liquids have had less success."

"Carrots used to be short chubby roots, far less attractive than the long slim beauties seen in many markets today. By careful breeding, a deeper orange color has been developed and the core has been made more tender or practically eliminated. The modern carrot, if well grown, has as little in common with the carrot of the past as a modern streamlined car has with a pre-war gas buggy."

SCIENTIFIC AMERICAN

MARCH, 1888: "Electrical science may be said to have been established upon a new basis socially in New York City when, on Tuesday, January 31, the new headquarters of the Electric Club were formally opened. The building, at No. 17 East 22d Street, is lighted by electricity; for the principal apartments electroliers in brass and silver bronze, and thickly hung with strings of glass prisms and pendants, are used. In the committee room is the long-distance telephone, with which communication can be had with all connected cities. A machine for blacking boots, consisting of a motor that rotates a flexible shaft, to the end of which is attached a rotating brush, is kept ready for use in the basement. A safe with an electric lock is used to hold the valuables of the club, and an electric door opener for the main entrance is employed. The hope is that the club will have a true work to do in furthering the progress of electrical science and its rapidly increasing application to the commercial interest of the world."

"An exhibition of throwing the boomerang was recently given by a party of Australian natives at Munster, before some German scientific men, who are endeavoring to discover the cause of the boomerang's curious flight."

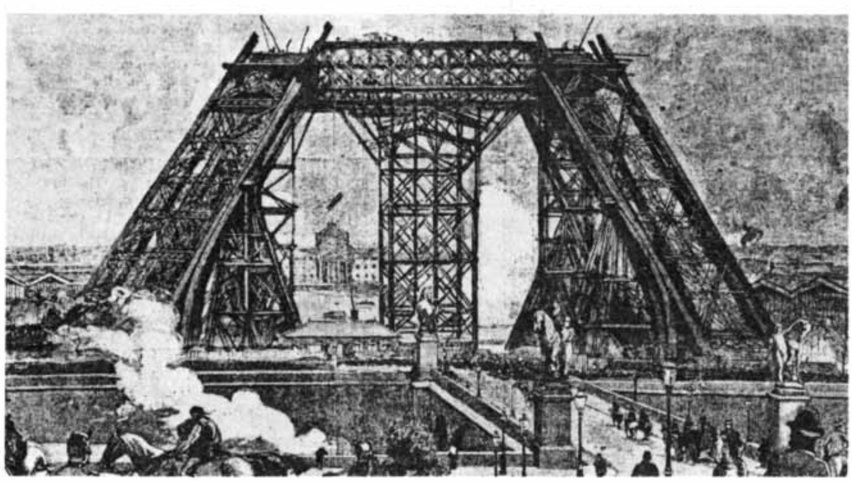
"The recent great snowstorm will not have been without some good results if it energizes the efforts of those seeking to introduce some hitherto obviously needed public improvements. Among such improvements is that of putting underground at least a portion of the telegraph and electric-light wires in all large cities, and burying some of the telegraph lines connecting the most important commercial centers. It would be ludicrous, were it not too serious a matter, to think of telegraphic messages being sent from Boston to New York via London, as was necessary on March 12 and 13."

"The transmission from the cow to man of scarlet fever and tuberculosis was the subject of an address at Marischal College, Aberdeen, Scotland. The subject deserves greater investigation, and certainly every effort should be made to prevent distribution of milk from tuberculous cows."

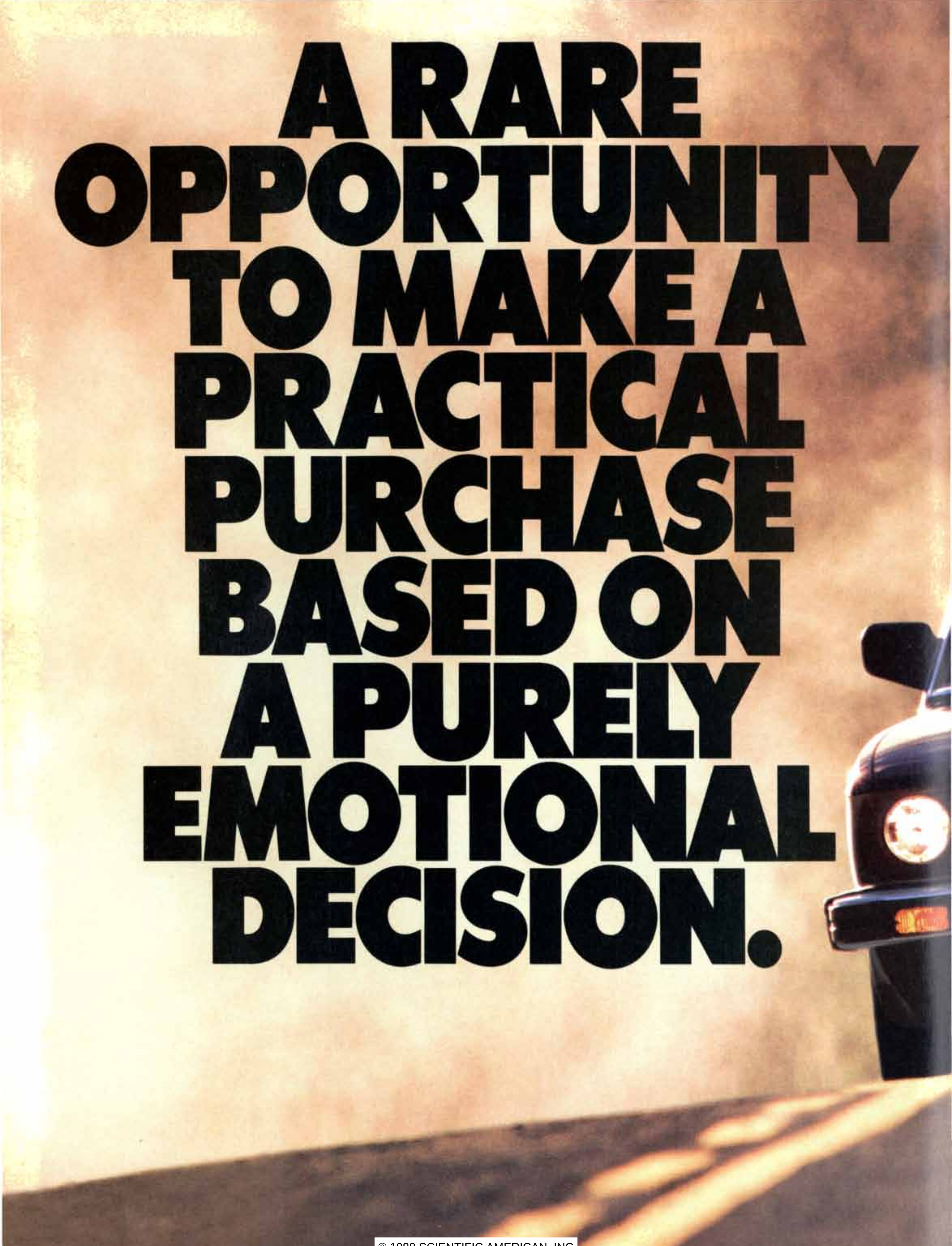
"The free use of onions for the table has always been considered by most people to be healthy and desirable, and but for their odor, which is objectionable to many, they would be found more generally on our dining tables. For a cold on the chest there is no better specific, for most persons, than well-boiled or roasted onions."

"Fire Marshal Whitcomb, of Boston, has been recently experimenting with rats and matches, shut up together in a cage, in order to ascertain whether they were likely to cause fires or not. The question may now be considered as settled. The very first night that Marshal Whitcomb's rats were left alone with the matches, four fires were caused. The rats were well fed, but they seemed to find something in the phosphorus that they liked."

"The greatest activity now prevails in Paris in respect to the preparation of the buildings for the International Exhibition of 1889, which promises to be by far the most wonderful and interesting industrial assemblage ever brought together."



Eiffel's great tower rises for the upcoming International Exhibition


A RARE
OPPORTUNITY
TO MAKE A
PRACTICAL
PURCHASE
BASED ON
A PURELY
EMOTIONAL
DECISION.

THE AUTHORS

TOM STEFANICK ("The Nonacoustic Detection of Submarines") is an American Association for the Advancement of Science fellow in Arms Control and National Security attached to the Armed Services Committee of the House of Representatives. He obtained a master's degree in civil engineering from the Massachusetts Institute of Technology in 1979, worked at the Institute for Defense and Disarmament Studies in Brookline, Mass., and was a research associate for naval policy at the Federation of American Scientists before accepting the fellowship. In addition to the detection of submarines, Stefanick has studied U.S. and Soviet naval strategy and the budget and force structure of the U.S. Navy.

TERRY GOLDMAN, RICHARD J. HUGHES and MICHAEL MARTIN NIETO ("Gravity and Antimatter") are theoretical physicists at the Los Alamos National Laboratory. Goldman has a 1973 doctorate from Harvard University and did postdoctoral work at the Stanford Linear Accelerator Center and at Los Alamos. He was a senior research fellow at the California Institute of Technology from 1978 to 1980, when he joined the staff of Los Alamos. He is currently working on the description of color confinement in quantum chromodynamics and its application to nuclear physics. Hughes, a native of England, got his bachelor's degree in mathematical physics in 1975 and his Ph.D. in theoretical elementary-particle physics in 1978 from the University of Liverpool. Since then he has held research positions at the University of Oxford, Caltech, the European Organization for Nuclear Research and the University of Edinburgh. He went to Los Alamos in 1985. Nieto, who entered graduate school at Cornell University intending to become an experimental high-energy physicist, became a theorist instead. On obtaining his doctorate he embarked on a "theorist's world tour," spending several years engaged in postdoctoral work at the State University of New York at Stony Brook, the University of Copenhagen, the University of California at Santa Barbara, the University of Kyoto and Purdue University. Nieto has been at Los Alamos since 1972 and is currently exploring the possible existence of unsuspected gravitational forces.

RICHARD A. LERNER and ALFONSO TRAMONTANO ("Catalytic Antibodies") are at the Research Institute of Scripps Clinic: Lerner is director of the institute and Tramontano is an associate member in the department of molecular biology. Lerner received his M.D. in 1964 at the Stanford University School of Medicine and has been at Scripps in various capacities since 1965, except for two years at the Wistar Institute of Anatomy and Biology in Philadelphia. This is his third article on immunochemistry for SCIENTIFIC AMERICAN. Tramontano received a bachelor's degree in 1976 at Columbia University and a Ph.D. in inorganic chemistry from the University of California at Riverside in 1980. He went to Scripps in 1983 after he completed postdoctoral studies at Harvard University.

LAWRENCE K. SHAPIRO and HOWARD I. SHAPIRO ("Construction Cranes") are son and father and also colleagues at the consulting engineering firm Charles M. Shapiro & Sons, which was founded by the elder Shapiro's father. Lawrence Shapiro, who earned a bachelor's degree in civil engineering from the University of Virginia in 1978, has been the project engineer for cranes since 1983. Before that he held several positions at Zurn Industries, based in Tampa, Fla. Howard Shapiro is principal engineer for cranes. He was graduated from the Polytechnic Institute of Brooklyn in 1953 with a degree in civil engineering and joined the firm in 1957 after serving in the U.S. Army Corps of Engineers. He is the author of a book on cranes and derricks.

JAMES D. MURRAY ("How the Leopard Gets Its Spots") is professor of mathematical biology and director of the Centre for Mathematical Biology at the University of Oxford. He holds a 1956 doctorate in applied mathematics from the University of St. Andrew's and a 1968 doctorate in mathematics from Oxford. From 1961 to 1963 he was a fellow and tutor in mathematics at Oxford. From there he went to the University of Michigan as a professor of engineering mechanics, and in 1967 he moved to New York University as a mathematics professor. He has worked at Oxford since 1970. Murray has been a visiting professor at many places around the world, including the National Tsing Hua University in Taiwan and the University of Florence.

ROSE E. FRISCH ("Fatness and Fertility") is associate professor of population sciences at the Harvard School of Public Health. She received her bachelor's degree at Smith College in 1939, a master's in zoology at Columbia University in 1940 and a Ph.D. in genetics from the University of Wisconsin in 1943. She began concentrating on the study of reproductive biology after returning to work and joining the Harvard Center for Population Studies in 1965. Frisch's current research includes the effects of exercise on the long-term reproductive and general health of women.

ROBERT T. BATE ("The Quantum-Effect Device: Tomorrow's Transistor?") is manager of the advanced concepts branch at Texas Instruments, Incorporated, where he has held several positions since 1964. The group he manages is studying nanometer semiconductor structures and revolutionary architectures with the goal of exploiting quantum-size effects to develop a new generation of ultradense integrated circuits. He has a 1955 bachelor's degree in engineering physics from the University of Colorado and a 1957 master's degree in physics from Ohio State University. Before going to Texas Instruments, he did research at the Battelle Memorial Institute. Bate holds 15 U.S. patents.

ANNA MARGUERITE McCANN ("The Roman Port of Cosa"), the first woman underwater archaeologist, is director of excavations at the port of Cosa. She was graduated from Wellesley College in 1954 and went on to get a master's degree at New York University's Institute of Fine Arts and a Ph.D. from the University of Indiana. She has taught at Swarthmore College, the University of Missouri, the University of California at Berkeley and New York University. Her work at Cosa began when she was a classical fellow at the American Academy in Rome. The Cosa project was supported by the National Endowment for the Humanities. Its results were summarized in *The Roman Port and Fishery of Cosa* (Princeton University Press, 1987). McCann, who is chairman of a new national committee for underwater archaeology of the Archaeological Institute of America, lives in New York with her husband, Robert Taggart.

There are no bruises.
And no broken bones.
She seems the picture of the perfect child.
But if you look closely you can see how rejection, fear and constant humiliation have left scars that have tragically affected her childhood.

So now only a shattered spirit remains.
And the light of laughter has gone out.
Remember that words hit as hard as a fist. So watch what you say.
You don't have to lift a hand to hurt your child.

Take time out. Don't take it out on your kid.

Write: National Committee for Prevention of Child Abuse, Box 2866E, Chicago, Illinois 60690

Great Products at Great Prices from a Great Company...
ELEK-TEK

DATA STORAGE

I·MEGA BERNOULLI BOX
For PC/XT/AT and most compatibles
10 meg (single) **$950**
20 meg (single) **1275**
20 meg (10 + 10) **1450**
40 meg (20 + 20) **1650**
Bernoulli Plus (80+20+20) **2999**
Interface priced separately
UNIVERSAL INTERFACE
For all 8″ Bernoulli Models **149**
Bernoulli Cartridges
IOMEGA 10 meg—3 pak **155**
IOMEGA 20 meg—single **84**
BB HEAD CLEANING KIT 59
Call for Bernoulli Box II pricing

HARD DRIVE CARDS

Western Digital 20 MB **$350**
Western Digital 30 MB **399**
PLUS DEVELOPMENT
20 MB Hard Card **550**
40 MB Hard Card $849

MULTI FUNCTION BOARDS

AST
Advantage 128 . . . . **$229**
I/O Mini II **79**
Six pak + with 384K **169**
Six pak premium with 512K **240**
Rampage/2 256 . . . . **229**
Rampage 286 . . . . **315**
INTEL ABOVEBOARD
All products . . . . **CALL**
ORCHID TECHNOLOGY
All products . . . . **CALL**
GENERIC COMPATIBLES OF MOST OF THE ABOVE . . . . CALL

HARD DRIVES

SEAGATE
Internal Hard Disk Subsystems
20 Megabyte $299
Each system includes:
- Seagate Drive
- Controller
- Documentation
- 2 Cables
- 1 yr limited warranty
- For use in IBM PC, XT, AT&T, most compatibles

30 MB RLL ½ ht./RLL 349
40 MB Drive full ht. for AT 599
40 MB ½ ht. for XT/AT 439
MINI SCRIBE
30 MB ½ height XT **349**
40 MB full height **675**

TAPE BACKUPS

Archive XL 40 MB **$329**
Mountain int. 40 MB XT or AT **379**
Mountain ext. 40 MB AT or XT w/power supply **575**

MODEMS

Hayes
Model 1200

HAYES Internal
(including software)
1200B 1200 Baud **$279**
2400B 2400 Baud **429**
HAYES External
(no software included)
1200 1200 Baud **289**
2400 2400 Baud **439**

OTHER MODEMS
Hayes compatible
Internal *(including software)*
Practical Peripherals 1200B **84**
Everex 1200 baud **94**
Everex 2400 baud **179**
Novation 1200 baud **125**
External *(no software included)*
U.S. Robotics 1200 baud **179**
U.S. Robotics 2400 baud **329**
Practical Peripherals 1200 baud . **109**
Practical Peripherals 2400 baud . **179**
Novation Parrot 1200 . . . . **92**
Novation 1200 baud **175**
Everex 2400 . . . . **225**

CHIPS - 9 packs

64K RAM (120, 150 or 200 NS) **CALL**
256K (120 or 150 NS) . . . . **CALL**
256K (100 NS) . . . . **CALL**
256K Static RAM (80 NS) . . . . **CALL**
128K RAM 150NS . . . . **2.20 ea.**
NEC V-20 . . . . **10 ea.**
Intel 8087 math chip (5MH) . . **110 ea.**
Intel 80287 math chip (6MH) . **175 ea.**

HEWLETT PACKARD
LaserJet Toner Cartridges
HP92285A for Original LaserJet
HP92295A for LaserJet Series II
Mfr. Sugg. Ret. $125 **$82**

PRINTERS

EPSON
FX 86e

Epson's FX-86e is the industry's standard DOT MATRIX PRINTER. Features include 8K buffer, NLQ, 40/200 cps.

OTHER EPSON PRINTERS
FX 86e (New)200 CPS . . . . **CALL**
FX 286e wide carriage (New) . . **CALL**
LX 800 180 CPS **199**
EX 1000 300 CPS . . . . **CALL**
LQ 500 180 CPS . . . . **CALL**
LQ 800 NLQ (Parallel) **299**
LQ 850 264 CPS . . . . **CALL**
LQ 1050 264 CPS . . . . **CALL**
LQ 2500 NLQ 324 CPS . . . . **CALL**

Diablo D80IF
Daisywheel Printer 80cps
Mfg. Sugg. Ret. $1995 **$499***
*price includes cut sheet feeder
6 month Elek-Tek warranty

PRINT BUFFERS

PRACTICAL PERIPHERALS
Microbuffer 32K **$69**
64K parallel or serial **129**
256K parallel or serial **159**
HANZON
64K universal **250**
64K memory expansion **56**
SMG
PB 36 256K parallel **$129**
PB 36 AB 2 pos. 256K parallel . . **175**

MONITORS

BEST BUY
Samsung 1252G 12″ Amber TTL . **99**
Samsung 13″ color **275**
Sakata 13″ EGA **325**
NEC
Multisync II color . . . . **CALL**
AMDEK
730 VGA color **350**
410A, 410G, 410W 150
722C Enhanced color **450**
PRINCETON GRAPHICS
All Models . . . . **CALL**
SONY MONITORS
All Models . . . . **CALL**

Huge Discounts on
TOSHIBA EPSON OKIDATA HEWLETT PACKARD XEROX / Diablo

DICONIX 8213
Portable Printer
Perfect for Laptop Computers
Elek-Tek Price $299

SPECIAL PURCHASE!
Diablo D-625 Daisywheel 25CPS
Elek-Tek Price $350*
*Price includes IBM parallel interface

TRIPPLITE & PERMA POWER POWER PROTECTION

Surge protector strips
PermaPower 6 outlet **$28**
Isobar4-6 4 outlet **42**
Isobar8-15 8 outlet **57**
Line conditioners/stabilizers
TL1200 4 out 1200W **135**
TL1800 6 out 1800W **199**
Command consoles
TLCCI 612 . . . . **79**
PermaRS562 Power Commander **65**
Standby power systems
TLBC450 450 watt system **360**
TLBC750 750 watt system **535**
TLBC1200 1200 watt system . . . **725**
TLBC2000 2000 watt system . . **1199**
SAFE400VA 400 volt system . . . **350**
SAFE1200VA 1200 volt system . . **899**

PUSH BUTTON SWITCH BOXES

448010 2 pos. 25-pin AB **$29**
448052 2 pos. 36-pin AB **32**

ROTARY SWITCH BOXES

448028 3 pos. 25-pin ABC **$29**
448060 3 pos. 36-pin ABC **33**

VIDEO BOARDS

Hercules Graphics PLUS **$179**
Generic Hercules compatible **80**
Hercules color **155**
Generic color compatible **80**
Hercules InColor Card **225**
Vega Deluxe **209**
Everex EGA Autoswitch **115**
Paradise EGA 480 . . . . **165**
Paradise EGA 350 . . . . **135**
Paradise VGA Plus **239**
Vega VGA **299**
Everex VGA **260**
ATI VGA improved performance . **299**

HUGE SAVINGS ON DISKETTES *Call for Quantity pricing for 10 boxes or more*

(*Packed in flip n' file) (**unformatted)	Dysan	maxell	3M	IBM	ELEK-TEK	SONY	
3½″ SSDD	14.00	13.00	12.00*	—	—	13.00	12.00
3½″ DSDD	17.00	18.00	17.00	22.00	—	17.00	16.00
3½″ DSDDHD	—	49.00	49.00	47.00	—	49.00	—
5¼″ SSDD	8.00	8.00	8.00	—	—	—	7.00
5¼″ DSDD	10.00	9.50	9.00	15.00	7.99	—	8.00
5¼″ DSDD96TPI	12.00	18.50	18.50	—	—	—	10.00
5¼″ DSDDHD (for IBM AT)	12.00	18.00	17.00	28.00	9.99	—	17.00
8″ SSDD**	19.00	20.00	20.00	—	—	—	—
8″ DSDD**	23.00	23.00	23.00	—	—	—	—

3M DATA CARTRIDGES

3M	1	10	25
DC100A	$13.50	$13.00	$12.00
DC300A	17.50	17.00	16.00
DC300XL/P	21.50	21.00	20.00
DC600A	23.00	22.50	21.50
DC600HC	25.50	25.00	24.00
DC615HC	19.00	18.50	17.50
DC1000	14.50	14.00	13.50
DC1000 alpha 20 MB*	16.50	16.00	15.00
DC1000 delta 10 MB*	16.50	16.00	15.00
DC2000	19.00	18.50	17.50
DC2000 gamma 40 MB*	22.50	22.00	21.00
DC2000 kappa 40 MB*	22.50	22.00	21.00

* formatted cartridges

ELEK-TEK
6557 North Lincoln Avenue, Chicago, Illinois 60645 312-677-7660
Call Toll Free 800-621-1269—Except IL, AK • Canadian Toll Free 800-458-9133

Corp. Accts. Invited. Min. Ord. $15.00. Visa or MasterCard by Mail or Phone. Mail Mashier's Check, Mon. Ord., Peronal Check Add $4.00 1st item. (AK, HI, P.R., Canada add $10.00 first item) $1.00 ea. add'l shpg. & handl. Shipments to Il address add 7% tax. Prices subj. to change. WRITE for free catalog. RETURN POLICY: All sales are final except defectives which will be replaced with identical merchandise only within various time frames as permitted by manufacturer's warranty policy. ALL ELEK-TEK MERCHANDISE IS BRAND NEW, FIRST QUALITY AND COMPLETE. Delivery subject to availability. **DUNS #09-718-0517**
Merchandise available at our Oakbrook/Willowbrook store.


GO FLY AN O-DAKO
The fighting kites of Asia fly in the same kind of environment you deal with every business day. A competitive one.
So, in the spirit of competition, Northwest has a way to keep you ahead of the rest of the field. Our schedule.
We fly daily to Asian business centers like Tokyo, Seoul, Taipei, Osaka and Hong Kong. On nothing but 747s.
In fact we fly to more of the Far East from more of the U.S. than any other airline.
So call your travel agent, or Northwest at 1-800-447-4747. When it comes to competition, we can make you a true champion.
NORTHWEST
©1988 Northwest Airlines, Inc.
LOOK TO US NORTHWEST AIRLINES

SCIENCE AND THE CITIZEN

Poisoning the Air

The U.S. resumes production of lethal nerve-gas weapons

One year ago the Soviet Union would not even acknowledge that it possesses chemical weapons. Then last April, General Secretary Mikhail S. Gorbachev announced that his country does indeed have such weapons and that to spur negotiations toward a ban he was halting production. Subsequently Soviet negotiators at the United Nations Conference on Disarmament in Geneva agreed to long-standing U.S. demands for on-site inspections, including so-called challenge inspections of any facility at any time, to verify a ban. In November, Soviet and American officials toured each other's chemical-weapons facilities. On December 10, during the summit, Gorbachev and President Reagan signed a joint statement proclaiming their mutual desire for a ban.

Six days later the U.S. began manufacturing chemical weapons for the first time in almost 20 years. Called binary weapons, they consist of two components, a canister and a weapon casing, each containing a nonlethal substance; after the canister is inserted into the casing the contents mix to produce deadly nerve gas. The Pentagon has sought funds for binaries for years; last year, after a bitter debate, Congress allocated $58 million to the start-up.

The Department of Defense already stockpiles, in the U.S. and Germany, tens of thousands of "unitary" weapons, including artillery shells, bombs, rocket warheads and land mines. The chemicals they contain include GB and VX, nerve gases that paralyze victims and can asphixiate them within seconds; mustard gas, which causes severe blistering of the skin and lungs and kills more slowly, and also nonlethal "incapacitants" such as BZ, an LSD-like hallucinogen.

Pentagon and Department of State officials maintain that most of the unitary weapons are obsolete and dangerous, and that binaries are needed to counter a buildup by the U.S.S.R. They also insist binaries are good for arms control and in fact have spurred the Soviets into making their recent concessions. "Showing our seriousness about modernization," Maj. Randal E. Morger says, speaking for the Pentagon, "has helped the negotiating process."

Proponents of binaries stress that even if the Soviet Union unilaterally disarmed, the U.S. would still need chemical weapons to deter attacks by such countries as Iran and Libya, which are rumored to have acquired chemical weapons. The U.S. should disarm, according to this reasoning, only after all nations agree to do so. Working out the verification procedures for such a treaty could take years, since inspections must involve not only Government facilities but also chemical companies, which are concerned about the loss of proprietary information; after a treaty is signed, destroying stockpiles may take another decade. In the meantime, proponents say, binaries will protect the U.S. and strengthen its hand in the Geneva talks.

Arms-control advocates vigorously dispute these views. Ivo J. Spalatin, a staff member of the House Foreign Affairs Committee who has fought the funding of binary weapons for years, says those who believe the binaries are still needed as a bargaining chip misinterpret the U.S.S.R.'s recent actions. "I think the Soviets realize that from a military point of view chemical weapons have little or no value," he explains.

Spalatin argues that the binary program may actually leave the U.S. with less of a deterrent. He says production of two out of three of the planned binary weapons, a bomb and a multiple-launch rocket, has been delayed indefinitely by technical problems; currently the Army can manufacture only 155-millimeter artillery shells, which have a range of about 15 miles. Congress has also encumbered the weapons with logistic constraints: the two components of each shell or bomb must be stored in different states and brought together only when and where they are to be

The Pentagon says building binary weapons will further negotiations toward a treaty

GAS MASK and a heavy, rubberized suit are the standard equipment issued to soldiers in the U.S. and the U.S.S.R. for protection against chemical weapons.

used. Spalatin thinks the U.S. should negotiate a bilateral agreement with the U.S.S.R. and then put pressure on other countries to join in the ban. "You will never get a multilateral agreement without getting a bilateral one first," he says.

Matthew S. Meselson of Harvard University, a biologist who has long championed biological and chemical arms control, agrees. He suggests that the U.S. decision to build more binaries, far from fostering disarmament, may encourage other countries to obtain chemical weapons. He also rejects the tit-for-tat concept of deterrence. "If we want to retaliate against Libya or Iran," he says, "we have many other more precise options than chemical weapons."

Meselson points out that because users of chemical weapons must don bulky protective gear, their mobility and hence their tactical advantage is reduced. This drawback, he notes, might explain why Iraq's recent use of mustard gas, in spite of killing and horribly injuring thousands of Iranian soldiers, has not led to a military victory. —*John Horgan*

Cousins or Brothers?

Chimps and human beings may be nearest relatives

Which are right, the genes or the bones? On the basis of anatomical evidence, including the fact that both walk on the knuckles of their hands, chimpanzees and gorillas would seem to be each other's closest relatives. Yet recent molecular findings have suggested that chimpanzees are actually closer to human beings than they are to gorillas. Morris Goodman of the Wayne State University School of Medicine has now provided the best molecular evidence so far that human beings and chimpanzees are indeed each other's nearest living relatives.

Goodman, together with his coworkers Michael M. Miyamoto and Jerry L. Slightom, determined the nucleotide sequence of the same stretch of DNA from an individual of each of the three species. The stretch they sequenced lies between genes that encode the protein portions of the hemoglobin molecule, but it does not itself code for a protein product. When the 7,000-nucleotide sequences from each species had been determined, the investigators inferred the existence of a common ancestral gene. Then they calculated the simplest branching pattern that would fully account for the observed divergence from the common ancestor.

It turns out that the simplest pattern is one in which the gorilla first splits off from the progenitor of chimpanzees and human beings; later the chimpanzee and human lineages diverge to yield their modern equivalents. Although Goodman's method cannot provide precise dates or time intervals for the events in the branching pattern, it does seem clear that the common chimp-human stem existed only briefly—perhaps for from half a million to a million years between five and 10 million years ago.

One of the intriguing consequences of Goodman's data is the implication that the ancestors of human beings were knucklewalkers. Since both the gorilla and the chimpanzee walk on their knuckles, either their common progenitor did the same thing or the complex trait evolved independently in the two lines. Since such independent evolution seems improbable, and since in the new branching pattern any common ancestor of chimps and gorillas was also the ancestor of human beings, the implication is that early human ancestors may have been knucklewalkers too.

That conclusion (along with much of the information from molecular-genetic studies) has aroused controversy among those whose evolutionary studies are based on morphology, or anatomical form. For example, little in modern human anatomy or in the hominid fossil record suggests knucklewalking. It will undoubtedly be some time before the apparent conflict between morphological and molecular data is resolved. Meanwhile Goodman proposes that chimpanzees, gorillas and human beings be put in the same subfamily in the overall scheme of classification of the species. That would be a radical move, because they are now not only in different subfamilies but also in entirely different families: Hominidae for human beings and Pongidae for the apes. —*John Benditt*

Did our ancestors walk on their knuckles? Molecular evidence suggests that the answer is yes

CHIMPANZEES AND GORILLAS both support their weight on the knuckles of their hands when they walk. Human anatomy shows no sign of knucklewalking, but new genetic information implies that early human ancestors may have used this form of locomotion.

Civil Defense

The military loses a fight for control of data—or does it?

At a quiet bill-signing ceremony in the White House in early January, a protracted battle by the Pentagon for control of civilian electronic information systems—particularly those containing scientific and technical data deemed valuable to the Soviet Union—ended in defeat.

Perhaps.

The battle officially began in 1984 when President Reagan issued a "national security decision directive" declaring that unclassified but "sensitive" information stored in computers and sent over telecommunications links, both Federal and privately owned, "can become targets for foreign exploitation." The directive created a committee headed by the National Security Agency, an arm of the Department of Defense, to devise stiffer security standards.

The NSA, which shields classified

Nuclear energy can help America find a way out of our dangerous dependence on foreign oil

Oil imports are increasing to dangerous levels. As the uncertainty in the Persian Gulf continues, the ability to rely on America's nuclear energy becomes more important than ever.

During the 1973 embargo, when we were importing 35% of our oil, prices skyrocketed as supply nose-dived. In the last 18 months, America's dependence on OPEC oil has increased dramatically. We're even more dependent now than we were in 1973. Oil imports have risen by over 25% while domestic oil production has fallen nearly 10%. Looking to the future, the situation is even worse.

In fact, if projections from the Department of Energy are correct, America may be importing as much as 50% of our oil by 1990. That would seriously jeopardize our national energy security.

Nuclear Energy Saves Oil

Electricity generated from America's 109 commercial nuclear electric plants saves us over 750,000 barrels of oil a day. Every day. Without nuclear energy's contribution, we would need to import even more foreign oil than we already do.

Nuclear Energy for the Future

America's use of electricity has been growing steadily to fuel our growing economy. At current growth rates, electricity demand will overtake supply in the early 1990s.

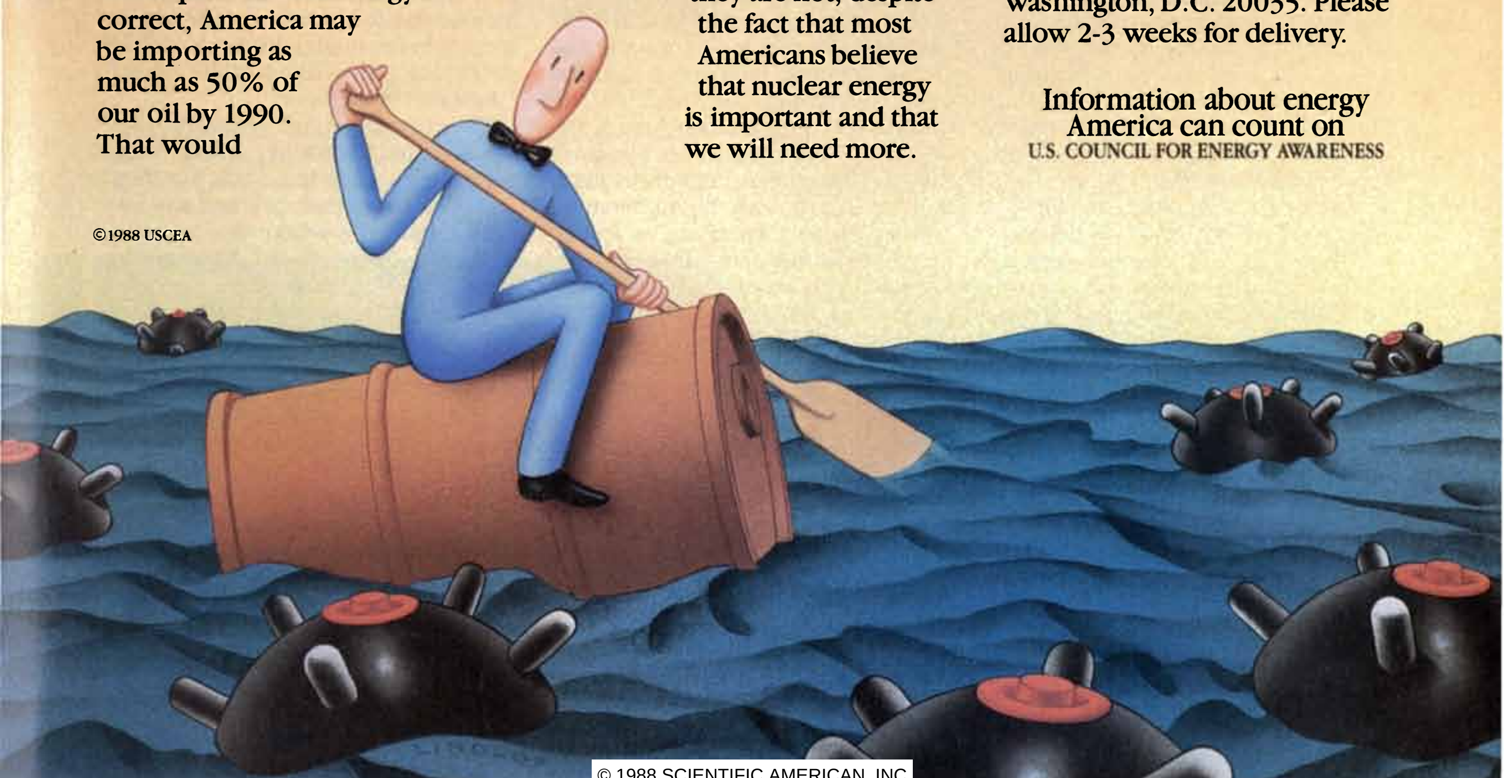
New nuclear electric plants should be in planning *now*. But they are not, despite the fact that most Americans believe that nuclear energy is important and that we will need more.

Too many financial, political, licensing, and regulatory uncertainties stand in the way of America's being able to fully utilize its nuclear energy resources. For example, it has taken some plants as long as 12 years to be completed. If nothing changes, that means that a plant begun now might not be operating before the year 2000.

As America's economy continues to grow, America must find ways to keep pace with its growing electricity needs. Nuclear energy can play a major role in meeting those needs as well as keeping us less dependent on foreign oil.

For a free booklet on energy independence, write to the U.S. Council for Energy Awareness, P.O. Box 66103, Dept. NS10, Washington, D.C. 20035. Please allow 2-3 weeks for delivery.

Information about energy America can count on
U.S. COUNCIL FOR ENERGY AWARENESS



U.S. data from foreign spies and also spies on other countries' transmissions, is extremely secretive. Nevertheless, in 1985 NSA officials emerged from their vast headquarters at Fort Meade, Md., to urge U.S. companies—including IBM, AT&T and others that deal in high technology—to buy equipment it had designed for scrambling, or encrypting, data. The NSA insisted, however, on managing the "keys" needed to unscramble messages and prohibited the use of its encryption equipment for international communications.

Because of such restrictions, even companies interested in encryption, such as banks, resisted the NSA's overtures, according to Robert H. Courtney, formerly a security analyst at IBM and now a consultant. The NSA, Courtney contends, is too "hopelessly introverted" to understand industry's security concerns, which involve not foreign espionage but fraud and embezzlement. Notwithstanding, according to John M. Richardson of the Institute of Electrical and Electronics Engineers (IEEE), a professional society that helps to set telecommunications standards, many companies worried that if they refused to comply with the NSA's "guidelines," they would lose Government contracts.

Then in 1986 agents of the Defense Department, the Central Intelligence Agency and the Federal Bureau of Investigation began visiting organizations that manage electronic data bases. The data bases, which transmit data over telephone lines to subscribers' personal computers, included *Chemical Abstracts,* run by the American Chemical Society, and Dialog, a service of the Lockheed Corporation. The Government agents reportedly asked to see lists of subscribers and inquired how access to non-U.S. citizens could be cut off.

In October, 1986, John M. Poindexter, then Reagan's national security adviser, aroused still more concern by issuing a memorandum stating what information the Government considered "sensitive" and therefore subject to regulation. Poindexter included any data related to national defense or foreign relations as well as "economic, human, financial, industrial, agricultural, technological, and law-enforcement information." "The memo included almost anything you could think of," says Charles K. Wilk of the Office of Technology Assessment, who has supervised a major study of data security for Congress.

Concerns over these actions were aired at hearings chaired by Representative Jack Brooks, head of the House Government Operations Committee, early last year. The American Civil Liberties Union, the American Association for the Advancement of Science, the American Physical Society and many other organizations argued that data controls hurt U.S. research and commerce.

The hearings, in the words of one observer, "turned the tide" against the military's program. Soon after the hearings ended Frank C. Carlucci III, who had replaced Poindexter as Reagan's national security adviser when the Iran-Contra affair erupted, rescinded the increasingly controversial "Poindexter memo." White House officials later voiced support for a bill, first proposed by Brooks, that makes the National Bureau of Standards, an arm of the Department of the Commerce, responsible for devising security guidelines for civilian information systems. After both the House of Representatives and the Senate had passed the bill, called the Computer Security Act, the president signed it into law on January 8.

Yet Pentagon officials, according to one congressional aide, "worked out a deal" with the Senate that may let the NSA keep a hand in the game. As part of the agreement, Senator Lawton M. Chiles declared for the *Congressional Record* that the Computer Security Act applies only to computers and "is not intended in any way to alter the assignment of responsibilities in the area of telecommunications security."

An aide to Chiles contends the statement was meant to allow the NSA to "save face" rather than to provide it with a semantic loophole. Other observers are disturbed both by the Chiles statement and by the ambiguity of the Computer Security Act itself. Robert L. Park of the American Physical Society maintains, however, that the act will have a positive effect, particularly in combination with the recent departure of two high-ranking officials in the Pentagon's policy office: Richard N. Perle and Fred C. Iklé. "These were the people behind the restrictive policies," Park says. *—J.H.*

BIOLOGICAL SCIENCES

Where's Rover?

Just when do children see that an object can be a symbol?

Sometime between the ages of two and a half and three, children make a leap in cognitive development that is as dramatic and clear-cut as the emergence of a first tooth in physical development. In that short time span, says Judy S. DeLoache of the University of Illinois at Urbana-Champaign, children learn that an object can be understood both as "a thing itself and as a symbol of something else." This capacity to symbolize has long been recognized as a hallmark of human cognition, but research had not established precisely when in the course of a child's development the capacity is acquired.

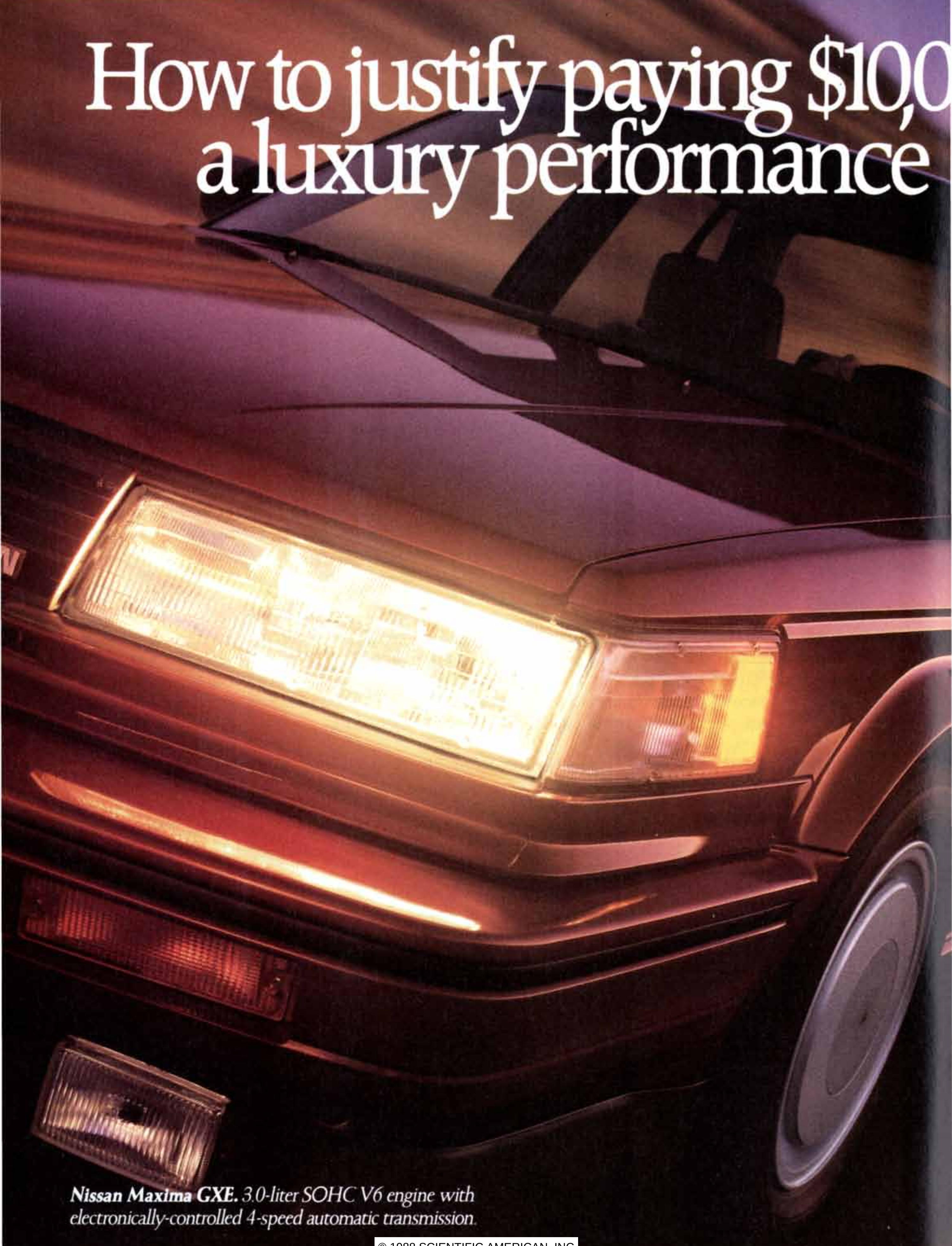
In *Science* DeLoache reports a successful effort to mark the event. She did an experiment with 32 children: 16 two-and-a-half-year-olds and 16 three-year-olds. Each child watched as a miniature toy was hidden in a scale model of a room (a tiny dog behind a couch, for example). The child was then taken to the full-size room and asked to find the real toy hidden in the corresponding place (a larger stuffed dog behind a full-size couch). Finally, as a memory check, the child was asked to retrieve the toy from the model.

"To succeed," DeLoache says, "the child had to realize that the model represented the room and that, by remembering the location of the object hidden in the model, he or she could determine the location of the object concealed in the room."

The two-and-a-half-year-olds, according to DeLoache, knew they were supposed to find a toy, but they were "unaware that they had any basis for knowing where the toy was without looking for it." As a group they searched with little success, even though they were adept at remembering where the miniature was hidden in the model.

In contrast, the approach of the older children highlights "the abrupt nature of the developmental change." Playing a memory game rather than a guessing game, they had "nearly universal" success in finding both objects. With a few months' difference in age, the three-year-olds were able to infer where the toy was hidden from the symbolic relation between the model and the room.

Why could the younger children


How to justify paying $10,0
a luxury performance
Nissan Maxima GXE. *3.0-liter SOHC V6 engine with electronically-controlled 4-speed automatic transmission.*

For some people, status means having the privilege of spending $27,000 for a European car.

Nissan has engineered a car for people who find status in something else: Just being smart.

The new Nissan® Maxima® GXE. A car *Motor Trend Magazine* called "The best performance sedan you can buy for under $25,000."* Which we take as an enormous compliment, since our car costs far less than that.

But as flattered as we are, we're not surprised. Because at the heart of the Maxima is the same basic fuel-injected V6 engine you'll find at the heart of our famous Z-Car.

And thanks to our revolutionary Sonar Suspension,™** the Maxima GXE not only monitors your driving behavior, it actually "reads" the road surface and automatically adjusts itself for improved comfort and handling.

Naturally, the Maxima also performs well in other areas. It has power windows, locks, mirrors, and antenna. As well as air-conditioning and a six-speaker electronically-tuned AM stereo/FM stereo radio and cassette deck.

Look at it this way. You could either buy a pricey, European luxury sedan. Or, for the same money, you could buy yourself a Nissan Maxima. And then go on a pricey European vacation.

NISSAN®

Built for the Human Race.™

*Based on manufacturer's suggested retail price. **Optional, but definitely worth it.

BELT YOURSELF

At three years of age children can recognize and exploit the symbolic meaning of things

CHILD LOOKS ON as an experimenter hides a tiny stuffed dog in a model of a room. Later the child will try to find a larger dog hidden in the same place in the actual room.

not see the correspondence? These children are accustomed to interpreting pictures as representations of something else, DeLoache hypothesized, but they are not yet capable of the "dual orientation" necessary to recognize that a real, three-dimensional object—such as a scale model—can also be a symbol. In a second experiment another group of two-and-a-half-year-olds were shown pictures rather than the scale model as a clue. They had no trouble finding the toy in the real room. With the scale model as a clue, however, the same subjects performed as poorly as the younger children had done in the first experiment. —*Elizabeth Collins*

Sex Switch

The gene determining maleness in human embryos is found

"Why can't a woman be more like a man?" asked Professor Higgins in Lerner and Loewe's *My Fair Lady*. The answer may have been found by a team of investigators led by David C. Page of the Whitehead Institute for Biomedical Research. The team, which also included workers from Canada and Finland, reports in *Cell* that it may have found the fundamental genetic switch that determines whether a human embryo will become a male or a female.

Male and female embryos develop identically until about five weeks after fertilization; it is only then that, in males, some genetic signal that was until now unknown sets in motion a cascade of biochemical events that leads to the development of male organs. In the absence of this signal the embryo develops into a female.

It has long been known that the mysterious signal must have something to do with the sex chromosomes. Normal males have an *X* and a *Y* chromosome, in addition to 22 pairs of ordinary chromosomes; normal females have two *X*'s. Since 1959 it has been thought that a "testis-determining factor" (TDF) must be encoded by the *Y* chromosome. Studies of people with abnormal combinations of sex chromosomes showed that male primary sexual characteristics do not develop in the absence of the *Y* chromosome.

Page's team picked up this trail. The workers studied the chromosomes of very rare individuals who although physically female have most of a *Y* chromosome. The investigators also studied individuals who were apparently male but who nonetheless seemed not to have a *Y* chromosome. Tests with DNA probes revealed that the apparently *Y*-less males did in fact have a tiny fragment of the *Y* chromosome; females with a *Y* chromosome lacked almost exactly the same fragment.

Page thinks the observations make a "very strong circumstantial case" that the fragment missing in females includes at least part of the gene that, when present, causes embryonic gonadal tissue to develop into testes. The sequence of bases (the chemical rungs of the DNA ladder) in the putative TDF gene suggests it would produce a protein similar to known proteins that bind to DNA and hence control the activity of genes; therefore Page supposes that the TDF gene may work in the same manner.

For direct proof that the TDF gene lurks in the missing fragment Page and his colleagues will inject female mouse embryos with the equivalent mouse gene; if the gene indeed includes TDF, it should make them develop as males.

Apart from its intrinsic interest, the discovery of the TDF gene is important because, Page says, it is the "first clear example of a gene controlling development of an entire organ system in mammals." The group has made a further "surprising finding," according to Page: there is a separate gene on the *X* chromosome very similar to the supposed TDF gene on the *Y* chromosome. Its role is not known, but Page thinks it too probably has a key role in sex determination.

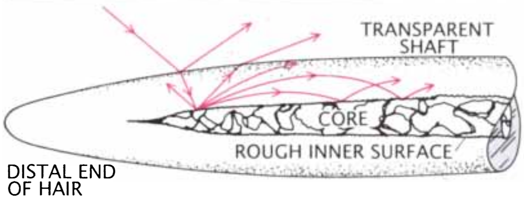
One reason to think the TDF-like gene on the *X* chromosome might be important is that it seems to have a long evolutionary pedigree: the team has discovered similar genes on the *X* and *Y* chromosomes in mammals ranging from rabbits to cattle. Even chickens appear to have similar sequences on their sex chromosomes, which is surprising because in birds sex is not determined by the *XX/YY* system. A role for a gene on the *X* chromosome would "completely rewrite our views on the chromosomal basis of sex determination," Page remarks. —*Tim Beardsley*

Solar Polar Bears

Polar bear fur sets a standard for solar-energy converters

On a clear day when the ambient temperature is zero degrees Celsius, even a very good solar-ener-

A polar bear's pelt converts sunlight into heat more than twice as efficiently as any solar heater



POLAR BEAR HAIR may be a natural fiber-optic cable. A cross section (right) *shows a solid shaft surrounding a reticulated core. The shaft apparently can trap ultraviolet light and aim it toward the skin* (above).

gy collector converts radiation into heat with an efficiency of only about 40 percent. There exists, however, a natural collector that converts part of the solar-radiation spectrum into heat with an efficiency exceeding 95 percent. The remarkable device is polar bear fur.

Clues to the capacity of polar bear fur for solar-energy conversion were found in a roundabout way some seven years ago when the Canadian government launched a population study of Arctic harp seals. The white mammals were difficult to track against the white background of the Arctic winter, but investigators assumed that infrared photographs made from airplanes would reveal the animals by detecting their body heat as it escaped into the frigid Arctic air. The seals defeated the scheme because their excellent natural insulation prevents almost all loss of heat.

Instead the answer to the tracking problem turned out to be ultraviolet photography. Whereas the environment reflected the ultraviolet region of the spectrum, the pelts of the seals absorbed it. Subsequent research showed that the pelts of other Arctic mammals absorb ultraviolet radiation; the polar bear proved to be the most efficient absorber.

The findings suggested to Richard E. Grojean of Northeastern University and the U.S. Army Natick Research and Development Command that a study of polar bear fur might further human attempts to harness the energy of solar radiation. Grojean found that individual polar bear hairs are actually colorless, not white. Under a microscope the transparent hair resembles a quartz fiber. The central core of each hair shaft scatters incoming radiation, giving the hairs their white appearance. (A similar scattering effect accounts for a snowflake's whiteness.) Grojean believes the hair shaft somehow conducts scattered radiation to the surface of the skin (which is actually black), where it is absorbed and converted into heat. This passively generated heat supplements metabolically produced heat.

Applications of these insights show some promise. According to Gregory Kowalski, a colleague of Grojean's at Northeastern, studies he has done demonstrate a 50 percent gain in efficiency when flat-plate solar collectors are filled with hairlike fibers. If solar panels could be modified to absorb more of the entire solar spectrum, Kowalski says, their efficiency could improve sharply. Much of the radiation, particularly ultraviolet wavelengths, pierces the cloud cover and so would reach a solar collector even on cloudy days. Such a system would also eliminate the need for elaborate tracking devices that turn collecting plates toward the sun: a collector filled with hairlike material should capture radiation no matter from what direction the radiation comes. *—Steven D. Mirsky*

PHYSICAL SCIENCES

Shedding Light

Supernova 1987A confirms and contradicts theories

Supernova 1987A has dimmed considerably, at least in the visible spectrum, since it first blazed forth in the Large Magellanic Cloud last year on February 24. But its contribution to astronomy is brightening rapidly.

Perhaps the most satisfying observations confirm for the first time the theory that heavy elements—from silicon on up—constituting much of the earth and its living inhabitants are forged in exploding stars. The first inklings that nucleosynthesis of heavy elements is occurring inside 1987A derived from its so-called light curve. For the first few weeks after its initial flash the supernova waned, and many theorists thought it would continue to do so.

Others, notably Stanford E. Woosley of the University of California at Santa Cruz, predicted otherwise. Woosley had constructed a supernova model in which the initial explosion generates large amounts of cobalt 56; this isotope, which has a half-life of 77 days, decays into stable iron, emitting gamma rays in the process. Woosley proposed that the supernova would brighten when its outer shell expanded far enough to allow the gamma rays, attenuated into visible light, to escape. Sure enough, the supernova brightened from mid-March through late May; since then it has waned at a rate that matches the decay of cobalt 56.

Woosley and other theorists also predicted that as the outer shell became still more transparent the decay of the cobalt would become more directly evident, first as X rays and finally as gamma rays. In October instruments on the U.S.S.R.'s *Mir* space station and on Japan's *Ginga* satellite detected the X rays. In the following weeks four detectors funded by the National Aeronautics and Space Administration—one aboard the *Solar Maximum Mission* satellite and the others on balloons lofted into the upper atmosphere above Australia—detected the gamma rays.

Further proof of the elemental richness of the supernova comes from NASA's Kuiper Airborne Observatory, an aircraft equipped with infrared sensors. In two flights in November the observatory detected infrared emissions typical of cobalt, nickel, silicon, sulfur and argon, according to Fred C. Gillett of NASA.

All these results, notes Edward L. Chupp of the University of New Hampshire, the principal investigator for the *Solar Maximum* project, represent "a real triumph of theory." He thinks the detection of the gamma rays, in particular, is comparable in importance to the detection of neutrinos when the supernova first

appeared, which confirmed theories about how such explosions begin.

Investigators hope future observations will help them to unravel the history of 1987A's precursor star: Sanduleak –69 202, a blue giant. Such stars were thought to be too hot and young to explode. Some investigators maintain, however, that Sanduleak was once a red giant and that thousands of years ago it shed its outer layers in the form of solar wind. According to Robert P. Kirshner of the Harvard-Smithsonian Center for Astrophysics, data from the *International Ultraviolet Explorer* satellite support this view: the data show the ultraviolet flash from the supernova heating up a slower-moving gas rich in nitrogen, thought to be a component of the outer shells of red giants.

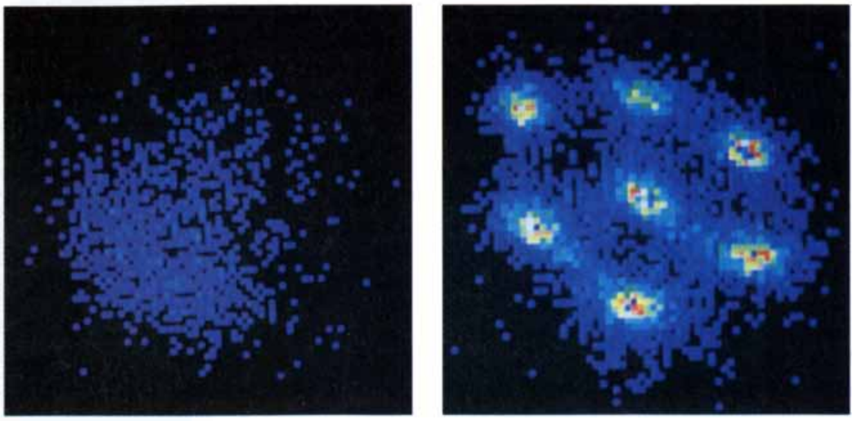
Workers are also seeking to determine whether the explosion of Sanduleak left a remnant. The burst of neutrinos that accompanied the explosion suggests that the core collapsed to form a neutron star. If the neutron star is rotating rapidly enough, it may be detectable as a pulsar—an emitter of highly periodic bursts of electromagnetic energy. Thomas A. Prince of the California Institute of Technology points out, however, that the fact that a pulsar has not yet been detected places an upper limit on its intensity.

Perhaps the most puzzling observations are those made by a team from Harvard working with a technique called speckle interferometry. Last June the team announced it had spotted a mysterious "companion" to the supernova (see "Science and the Citizen," July, 1987). The investigators have not detected the companion in later observations. More recently they reported that they had measured the diameter of the supernova and found it to be about five times wider than conventional theory would predict. Robert V. Stachnik of NASA, who is familiar with the group's work, maintains that the observations should be taken seriously. "They are over to the side somewhat now," he notes, "but they may turn out to be very important." —*J.H.*

Freezing Ions

Electromagnetic traps turn ion "gases" into "crystals"

An ionic "crystal" can be created if confining, inward forces balance the ions' mutual repulsion



IONS trapped by electromagnetic fields can form a diffuse cloud (left) *or a geometric array* (right), *depending on the power of an illuminating laser and the amplitude of the trap's fields. The images were made at the Max Planck Institute for Quantum Optics.*

For more than a decade investigators have been isolating ions (atoms that have lost or gained one or more electrons) for extended periods in electromagnetic traps: complex arrangements of electric or magnetic fields that confine slow-moving, charged particles in a small volume. Investigators do this, for example, to study the transitions of an electron from a higher to a lower energy state in individual ions. Now workers at the Max Planck Institute for Quantum Optics in West Germany and at the National Bureau of Standards in Boulder have observed trapped ions undergoing an entirely different kind of transition: from a cloudlike state in which the ions move randomly about to a crystalline state in which they are held fixed in a regular array.

Although the two groups used different metallic ions, their trapping techniques were essentially the same. They collected a small number of neutral atoms by vaporizing a bit of metal in a vacuum chamber. An electron beam traversing the chamber ionized the atoms, which were then immediately "cooled," or slowed down, so that their temperature (which reflects their kinetic energy) was only a few thousandths of a degree above absolute zero (about –273.16 degrees Celsius). This was achieved by directing a beam of laser light against the ions' direction of motion. The light in effect exerted an opposing pressure on the ions, causing them to come to a near standstill in the center of an electromagnetic trap. While the ions were confined in the trap they were kept bathed in laser light, since it induced them to fluoresce. The fluorescence (as a function of the laser's wavelength) could then be analyzed to gather information on the energy state of the ions. In addition, both groups of workers took advantage of the fluorescence in order to image the ions directly.

In general trapped ions are in constant (if bounded) motion, so that they appear as a diffuse "cloud" in such images. Yet it had been predicted that trapped ions could be held still if the inward force exerted on each one by the trapping electromagnetic field was balanced against the electrostatic repulsion among them (which arises from the fact that all the ions had like charges). In such a case the ions would naturally arrange themselves in an ordered, symmetrical structure.

As they report in the same issue of *Physical Review Letters,* both teams succeeded in stilling the ions by carefully adjusting the strength of the trapping field and the power of the cooling laser. As expected, the images distinctly show a small number of ions "frozen" in a geometric pattern, much as they are in ordinary crystals. The ionic crystals could then be easily "melted," so that the ions returned to a cloudlike state.

The two groups intend to analyze further the dynamics of the crystallized ions as well as their preferred geometric configurations. The U.S. group has already measured the vibrational frequency of one of a pair of frozen ions. Such experiments may yield insights on how physical systems change from ordered to disordered states. —*Gregory Greenwell*

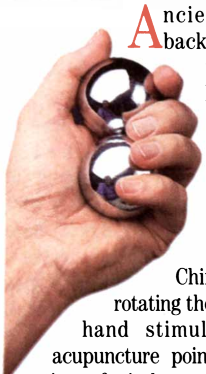
THE LIFESTYLE RESOURCE™

At The Lifestyle Resource we give you all the facts and details necessary to make an informed purchase. But your satisfaction is our primary concern. If your purchase doesn't meet your expectations, just return it within 30 days for a prompt refund.

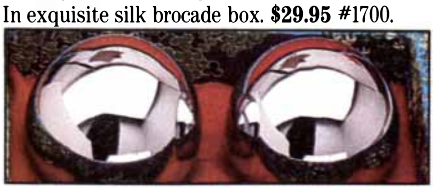
THE NEW INFRALUX

Pain emanating from inflamed joints and muscles can be excruciating. Just ask sufferers of bursitis, neuralgia, or arthritis. Infrared heat is the natural remedy preferred by doctors and therapists to dispel such nerve-racking pain. Recently redesigned, the Infralux infrared therapeutic heating unit for personal use now is contoured, angled for reaching back and neck. The heating surface is tripled and the replaceable 10-watt bulb gives 45% more heat. Hi/low settings. Plugs into household outlet. UL listed. Unit with case **$39.95** #1140.

TO YOUR HEALTH

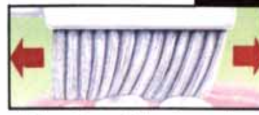


Ancient mandarins dating back 800 years believed that these Chinese Exercise Balls induced well-being of the body and serenity of spirit. These treasured gifts were given to President Reagan and his wife while visiting the Peoples Republic of China. The Chinese say that rotating the balls in the palm of each hand stimulates the fingers and acupuncture points, and improves circulation of vital energy throughout the body. Sports enthusiasts, musicians, computer users and health-conscious people everywhere consider them great muscle conditioners. Arthritis sufferers feel a decided benefit from this gentle but challenging excercise. Very effective for relaxation and meditation, Chinese Exercise Balls emit a distantly mysterious chime as you turn them. Beautifully hand crafted, 45mm. hollow polished chrome balls are perfectly weighted and fit comfortably into the average man's or woman's hand. In exquisite silk brocade box. **$29.95** #1700.



YOUR SMILE IS BEAUTIFUL

But you're only fooling yourself if you think it's proof of healthy teeth and gums. The unseen enemy is a natural mouth substance deposited on the teeth - plaque. Unchecked, plaque develops into rock-hard tartar behind which decay breaks down the tooth enamel and invades the soft, sensitive substructure. You have to remove the plaque before it hardens to prevent the process. Until recently there was no easy solution to this problem, but now there's Interplak, a precision-engineered, university-tested, three-speed plaque removal instrument. Its 10 independently driven bristle tufts rotate around their own axes: each tuft spins in the direction opposite to the spin of its neighbor at 4,200 rpm. The bristles are soft and rounded, adhering to dentistry recommendations, so that virtually every particle of plaque—*98% of it*—is removed gently and without irritation. Interplak even reaches between teeth and around orthodontic braces. Children think it's fun, so they're eager to brush. Interplak is cordless and easily stores in a free-standing or wall-mounted recharger/holder. Individual color-coded brush heads are available. The Interplak Home Plaque Removal Instrument with charger and two brush heads. **$99.95** #1110. Two additional brush heads **$29.95** #1111.



Hand brushing's motion misses plaque.

The counter-rotational movement cleans plaque.

ZONE OF CALM

Noise produces stress in the human organism. Today the volume of civilization nearly everywhere seriously interferes with our abilities to relax, read, sleep, concentrate or otherwise function at optimum efficiency. But you needn't be victimized by noise pollution. The Marsona Sound Conditioner electrically synthesizes a variety of pleasing natural sounds that help mask out and reduce the annoyance of unwanted noise. You can simulate the sounds of ocean surf, summer rain, mountain waterfalls. You control not only the volume but also wave pattern, wave or rain rhythm, the seeming nearness or distance of the source. The Marsona is the finest instrument of its kind on the market, solid-state electronics designed for continuous use (electrical draw only 5 watts). UL-listed. **$139.95** #1860.

A TIMELY RETURN TO THE CLASSICS

Back in the 1920's and 30's, the world's honored watchmakers were creating elegant timepieces that ultimately became valued works of art, fine heirlooms treasured for their imaginative yet functional design. Two of these classics are recreated with precision quartz movements accurate within a minute a year. The famous Retro Classic with silver brushed face curves to the wrist with contoured case, curved crystal. The antique-looking Moonphase revives a mystique that captivated moon-watchers eons before there were moon-walkers. The gold moon in a starry sky moves through a crescent window. Entrancing! All the delicate detail of the original is recaptured and a discreet date window added. Both styles **$59.95.** Retro Classic #1590 Ladies or #1600 Men's. Moonphase #1570 Ladies or #1580 Men's.

THE LIFESTYLE RESOURCE™

WORRY-FREE WATERING

Water grows things best when it arrives at regular intervals, a fact that sometimes gets by old Mother Nature. Now you can give your lawn or garden the advantage of timely watering. The new Rainmatic 2000 *automatically turns water on and off* up to four times a day, seven days a week, a different schedule every day. Lets you set any length watering interval. Easily programmed to set up your own customized watering schedule or use built-in program. One-touch manual override. Great for sprinklers and drip/trickle systems. Operates all season on four alkaline D-cells (not included). Weather-protected electronic circuitry. LCD readout with digital clock. Locks on to faucet. Only **$49.95** #1080 with one-year warranty and instruction manual.

AN AMERICAN CLASSIC

The Shaker bench, the Appalachian Quilt, the Adirondack chair: all are superb examples of native American design and craftsmanship that have endured over the years, virtually unchanged. The *very best* of the Adirondack chairs are still made in the Adirondacks, in Willsboro, N.Y., on the shores of Lake Champlain. Willsboro's skilled woodworkers have done more than just re-create the original Adirondack chair they've added refinements such as a fully contoured seat and 5″ wide arms for the ultimate in comfort. And the workmanship! Not a screw in sight - not even a capped screw, just solid maple pegs in the true yankee manner. The rich rock maple is an extremely dense and attractively grained wood, all sanded to a fine furniture finish - no sharp corners or splintering. Willsboro's artisans have incorporated a unique (and patented) tongue-in-groove folding mechanism that lets you fold the chair to a mere 10″ thickness; sturdy hardwood locking pins secure it in the open position. When each piece is completed, it is treated with Willsboro's special zinc-based, rot-resistant clear protective finish; leave them to weather naturally or they can be painted. There never has been an Adirondack chair like this...and we have a matching footrest for it, designed to create a chaise lounge, and an end table to go alongside - both constructed in the same fine tradition - an American classic. The chair is 38″ x 14″ x 20″, **$129.95** #1880. The footrest is 14″x 20″x 22″, **$69.95** #1890. The end table is 16″x 14″ x 24″, **$59.95** #1900.

38,000 HOUR LIGHTBULB!

The problem with bulb savers is they reduce light output by almost 50%. But the patented, UL listed Power Mizer has solved that problem. These little buttons are activated by a state-of-the-art microcontrol switch that limits voltage to the bulb to 85%, thereby prolonging the life of the filament. Light bulbs using Power Mizer lose so little perceptible light that a foot-candle meter is needed to calculate the light loss. Under normal conditions, the Power Mizer will extend the life of a regular one-way 100W bulb from 750 hours to 38,000 hours or about four years if used 24 hours every day! Many major hotels, corporations and universities use these types of devices in hard-to-reach places. They pay for themselves in light bulbs, labor and energy savings in no time. Set of 6 Power Mizers **$19.95** #1980; 12 for **$29.95** #1990.

HERE'S ONE FOR THE ROAD

Think how annoying it is to be stranded on a strange road at night. We thought about it too, and we put together this terrific auto emergency kit with all the high-quality essentials you need to get you moving again: the booster cables are heavy-duty, 100% copper with rugged alligator clamps; the 2 safety flares strike as easily as matches (and are approved in all 50 states.) You also get a can of Fix-a-Flat, a 12V spotlight, a hand-pumped gas siphon (no gas in your mouth), a water bag, repair wire, electrical/hose tape, a bungee (shock) cord, help flag, first aid kit, and moist towelettes to clean up with—all in a compact plastic case that stows easily in a corner of your trunk. Don't flirt with disaster—for the cost of a single service call—be prepared. **$39.95** #2020.

THE LIFESTYLE RESOURCE™

THE COLD BUSTER

Now a major scientific breakthrough —The Viralizer® System— prepares you for the onslaught of the cold season. It's the newest development of a concept pioneered at the Pasteur Institute in Paris. The cause of the common cold is the Rhinovirus family which lives and multiplies in the nose and throat, but cannot thrive in temperatures over 110°F.

The Viralizer is designed to deliver a gentle, controlled heat which penetrates the nose and throat, creating a hostile environment for cold viruses. After a pleasant heat treatment, the Vira-Spray dispenses either of two mild, over-the-counter, medicated sprays. Vira-Spray I is an analgesic, anti-bacterial spray. Vira-Spray II is a decongestant. These therapeutic sprays further discourage the stubborn cold germs so you're less likely to be re-infected or spread your cold to others. The Viralizer can produce effective relief by using it for only 3 or 4 minutes, several times a day. Proven in clinical tests 90% effective on eliminating the symptoms of upper respiratory infection in 24 hours or less, the Viralizer works without pills. Viralizer, is safe for children and adults, has been tested and recommended by doctors. The complete Viralizer® System includes 1 electric Viralizer with Vira-Spray I and Vira-Spray II plus a 3-pak refill of medicated sprays. **$39.95** #1690.

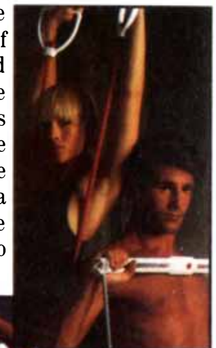
WHEN YOU'RE IN A HURRY

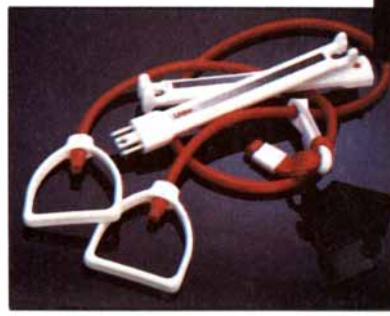
When you're running short of time, ask one of our phone operators about UPS Blue Label delivery. In most cases, we can get your shipment to you in three or four working days.

CALL TOLL-FREE 800-872-5200

THE LIFELINE GYM

The Lifeline gym is the most space efficient, time efficient home fitness device on the market today. You can simulate just about any of the exercises done on the expensive machines found in health spas and gymnasiums. The reason is latex—a unique stretchable latex resistance cable, similar to the type used to stop planes on aircraft carriers. With this latex cable, resistance increases with movement—your muscles are challenged through their full range of motion. These variable resistance exercises provide an excellent workout for a beginner as well as a challenge for a pro. The Lifeline Gym is so successful it is used by the Chicago Bears, The U.S. Ice Hockey team, and the UCLA track team to name a few. Adjustable for individual strength levels and for different body shaping and toning. The gym includes a lifting bar, resistance cable, stirrups, exercise belt, door attachment and a comprehensive fitness guide that outlines over 25 exercises. Everything packs into a compact carry case, so you can enjoy the benefits of a complete workout anywhere. **$49.95** #1840 Ladies or #1850 Men's.





The Lifeline Gym is so successful it is used by the Chicago Bears, The U.S. Ice Hockey team, U.S. Women's Volleyball Team, and the UCLA track team.

8K POCKET COMPUTER

Introducing the world's first double line 8K pocket computer, no larger than a credit card. With the revolutionary 8,000 character Platinum Business Card you can create a database: store up to 500 names, addresses and phone numbers, easily add, delete or edit information, program 6 currency exchange rates into the memory and also have a 4 function calculator. Then at the touch of a button, view it on the two line, easy to read, 40 character (the biggest of its kind) screen. Maintain a travel expense record, enter messages or appointments — then retrieve data quickly with the automatic alphabetical sort. Security code safeguards confidential data. Battery with memory backup included. **$49.95** #1820.

JEWELRY TAKES A SHINE

Brighten up your jewelry and restore the luster it had when you first wore it. Without the bother of taking it to a jeweler, or the difficulty of trying to clean it yourself. You'll welcome this remarkable, quick and easy process made possible by the science of ultrasonics. High-frequency sound vibrates ever so gently, creating unseen waves which act on a specially formulated cleaning solution. Makes jewelry come clean in minutes with no effort on your part. Everything gleams, sparkling new again—precious and semi-precious metals and stones. Even works well on dentures, using regular denture cleanser tablets. Complete twin tank system includes 6 oz. bottle cleaner solution, rinse tray, brush, dust cover. UL listed. **$36.95** #1630.

THE LIFESTYLE RESOURCE™ CREDIT CARD ORDERS CALL TOLL-FREE 24 HOURS **800-872-5200**

DEPT. SFAC08; 921 EASTWIND DR. SUITE 114: WESTERVILLE, OH 43081

	ITEM NO	QTY	DESCRIPTION	ITEM PRICE	TOTAL PRICE
SEND TO (PLEASE PRINT)					
ADDRESS					
CITY STATE ZIP					
☐ CHECK OR MONEY ORDER ☐ MASTERCARD ☐ VISA ☐ AMEX SIGNATURE					
ACCT. # EXP. DATE				SUB TOTAL	
				SHIPPING (see table at left)	
				TOTAL	

Shipping Charge covers UPS. handling and insurance for guaranteed delivery.

Up to $20 .... $ 3.95	$50.01 to $60 .... $ 7.95
$20.01 to $30 .... $ 4.95	$60.01 to $70 .... $ 8.95
$30.01 to $40 .... $ 5.95	$70.01 to $100 .... $10.95
$40.01 to $50 .... $ 6.95	Over $100 .... $12.95

UPS 2ND DAY AIR

UPS Second Day available for an additional $7.50 per order.

ORDER WITH CONFIDENCE

- *We ship most orders within 48 hours.*
- *Credit Card orders billed only upon shipment.*
- *Satisfaction guaranteed or return for prompt refund.*

THE MEDIA DEVELOPMENT GROUP

©1988 G9 ADVERTISEMENT

Seeing Red

Have astronomers glimpsed the bright birth of galaxies?

Where are the youngest galaxies, spawning multitudes of stars from primordial gas? One might expect protogalaxies to be abundant among the most distant objects seen from the earth: because of light's finite velocity, such objects represent the remote past. Of the most distant objects detected so far, however, nearly all are quasars—enigmatic point sources of light as bright as an entire galaxy.

Now a team of astronomers at the University of Arizona has recorded evidence for what might be an entire population of protogalaxies lying beyond the most distant quasars. The candidate protogalaxies, reported at the American Astronomical Society meeting in Austin, Tex., might lie as far away as 17 billion light-years and date from as early as two or three billion years after the big bang.

The investigators, Richard Elston, George H. Rieke and Marcia J. Rieke, made the observation with a camera incorporating an array of solid-state detectors sensitive to wavelengths of two micrometers, in the near-infrared region of the spectrum. Such infrared array detectors, vastly more sensitive than single-element detectors and photographic films, have become available only in the past three years and have transformed infrared astronomy.

High sensitivity to infrared light is needed to see objects at great removes of space and time. Because the universe is expanding, galaxies and other objects are receding from the earth at rates proportional to their distance. The recession is evident in a shift of their light toward the red (longer wavelength) end of the spectrum.

Last spring Elston and the Riekes coupled their camera to the 61-inch telescope at the University of Arizona's Steward Observatory and imaged a region of the sky well away from the obscuring dust in the plane of the Milky Way. After each exposure the observers photographed a second field offset from the first one and electronically subtracted the second image from the first to eliminate errors due to atmospheric variation and flaws in the detector array. The data from the 100 or so "flattened" images were then added together.

The resulting image showed several fuzzy objects only one ten-thousandth as bright as the background infrared glow of the sky. At visible wavelengths the objects were even dimmer. When the workers examined the same region of the sky with an optical detector mounted on a larger telescope (the 90-inch instrument on Kitt Peak), they detected only two of the objects, shining only one-twentieth as brightly at visible wavelengths as in the infrared.

The fuzziness of the objects suggested they are not nearby stars or distant quasars but galaxies. It was the sharp difference in their luminosity at infrared and visible wavelengths that led the Arizona group to propose that they may be protogalaxies at extremely high red shifts. "The prediction had been made," Elston says, "that primeval galaxies would look like what we found."

A protogalaxy, filled with short-lived, massive stars, is expected to shine brightly. Because the galaxy would also be rich in primordial hydrogen, however, its luminosity is expected to drop sharply at wavelengths shorter than the so-called Lyman limit, where hydrogen begins to absorb radiation. For nearby protogalaxies the Lyman limit would fall in the ultraviolet part of the spectrum, but it would be displaced redward for more distant objects. For protogalaxies at a red shift of between 6 and 25 (the number refers to the factor by which recession velocity increases wavelength) the discontinuity would fall between the near-infrared band and the visible band.

That is just where the luminosity of these new objects drops off sharply. Elston and other astronomers caution, however, that the identification of the objects as protogalaxies at high red shift is extremely tentative; the Arizona group plans follow-up observations that will include an attempt to record the objects' spectrum, which might yield a precise value for the red shift. Nevertheless, according to Harding E. Smith, Jr., of the University of California at San Diego, "It would be very exciting if it were true."

For one thing, the presence of two of these intriguing objects in the tiny patch of sky imaged by the Arizona group suggests that they are abundant: as many as 1,000 of them may populate each square degree of sky. The existence of many galaxies so soon after the big bang would complicate a widely accepted account of how density fluctuations first formed in primordial matter. The scenario, known as the cold-dark-matter model, holds that matter was evenly distributed until quite late in the history of the universe, well after the epoch in which protogalaxies may now have been spotted. "If all galaxies formed this early," says Simon D. M. White, a cosmologist at the University of Arizona, "the theory goes out the window." —*Tim Appenzeller*

MEDICINE

Decoy

One potential drug would lure the AIDS virus to a false target

As investigators gain a deeper understanding of the life cycle of HIV, the virus that causes AIDS, they are applying the knowledge in attempts to synthesize agents tailored to interrupt the cycle at specific points. In the most recent example of this method of drug design, workers have built on knowledge of how the AIDS virus locates and binds to certain cells of the immune system. They have found a way to interrupt the binding process, at least in a laboratory dish, by creating what are essentially decoys. The virus binds tightly to the decoys and is thereby prevented from attacking genuine immune-system cells.

The decoys are based on a molecule called CD4 or T4, a protein embedded in the membrane of some cells of the immune system. In the protein envelope of the AIDS virus there is a molecule called gp120, which recognizes and binds to the CD4 molecule, enabling the virus to invade the cell.

The binding between gp120 and CD4 is also important in later stages of the virus's attack. Once the HIV genetic material has been inserted into the DNA of the infected cell and activated, the infected cell itself produces gp120 molecules, some of which are incorporated in the cell membrane. There the gp120 can bind to the cell's own CD4, wrinkling and distorting the cell membrane until it ruptures. Alternatively, the gp120 can bind to the CD4 in the membrane of other, uninfected cells. The uninfected cells and the infected one then

THE NEW CHRYSLER LE BARONS. BEAUTY...WITH A PASSION FOR DRIVING.

The new Chrysler LeBarons have been called the most beautiful cars to come out of an American design studio in over a decade.

Others, however, say their true beauty is revealed on the open road. And not without reason.

Beneath their sleek, fluid lines breathes a 2.2-liter multi-point EFI turbo* that unleashes 146 spirited horses.

Plus all the right hardware—advanced front-wheel drive, quick-ratio rack-and-pinion steering, gas-charged shocks—to tame the most demanding roads.

The new Chrysler LeBaron convertible possesses a rare distinction. Not only does it offer exceptional beauty and luxury—it is designed and built as a true convertible from start to finish...for a smooth, quiet ride.

Chrysler backs every LeBaron engine, powertrain and turbo with a 7-year/70,000-mile Protection Plan.† And outer body rust-through for 7 years or 100,000 miles.

See the beauty. Feel the passion. The Chrysler LeBarons. Buy or lease at your Chrysler-Plymouth dealer.

Chrysler 7/70

Division of Chrysler Motors

CHRYSLER. DRIVING TO BE THE BEST.

*Optional. †See limited warranty at dealer. Restrictions apply. BUCKLE UP FOR SAFETY.

fuse to form an entity called a syncytium. Molecules of gp120 on the surface of the syncytium continue to bind to uninfected cells, bringing them into the syncytium. The result is that a single infected cell can disable as many as 500 uninfected cells.

Now, in five nearly simultaneous papers in *Science* and *Nature*, investigators working at a number of research institutions and pharmaceutical companies report independently that they have created fragments of the CD4 molecule that bind to gp120. When AIDS virus is added to a petri dish containing these fragments as well as cells that have CD4 in their membranes, the virus binds to the CD4 fragments instead of invading the cells or causing syncytia to form.

If CD4 survives a long series of animal tests and clinical trials, it will probably be administered in combination with drugs that attack the virus at other points in its life cycle. One example is zidovudine (AZT), which interrupts reverse transcription. That is the process in which the virus's genetic information, encoded in RNA inside the virus, is copied into DNA to be integrated into the genetic information of the infected cell. AZT is effective in patients with advanced cases of AIDS: the one-year mortality rate for patients taking AZT is about 10 percent, as against about 50 percent for similar patients who are not taking it. In addition, AZT has been shown in some instances to block or even reverse the effects of AIDS on the central nervous system.

The drawback of AZT is its toxicity. It attacks the bone marrow of about 40 percent of the patients who take it. A treatment that may avoid this problem has now been described in *The Lancet* by Robert Yarchoan of the National Cancer Institute and his colleagues at the NCI and at other institutions. Their treatment involves dideoxycytidine (ddC), which blocks reverse transcription much as AZT does but has a different toxicity: it attacks patients' peripheral nerves. In an initial trial in which patients alternated weekly between taking AZT and taking ddC, the alternating treatment was shown to be effective and to reduce the toxicity of both drugs. A similar drug, dideoxyadenosine, will soon undergo trials. Some alternation of reverse-transcription inhibitors could be the second ingredient, with CD4, in a therapeutic cocktail.

As for other stages of the virus's life cycle, investigators are exploring a wide range of candidate treatments. As Samuel Broder of the NCI says, "we now have more candidates that work in the test tube than can reasonably be tested in clinical trials." One agent, now entering animal trials, is castanospermine, which blocks the formation of HIV envelope protein. Other agents under consideration interfere with the assembly of new virus components in infected cells or with the "budding" of completed virus particles from those cells.

Not one of these drugs represents even a potential "cure" for AIDS. At best they might act as palliatives to extend the life of patients with late-stage cases of AIDS and to prevent patients in the earlier stages of the disease from progressing to later stages. Indeed, according to William Haseltine of the Dana Farber Cancer Institute, the two kinds of patients in which these agents could do the most good are those who have been infected by the virus but do not yet show symptoms of AIDS, and those who are at high risk of being exposed to AIDS (including fetuses and infants of mothers with AIDS).

The major problem now, according to Martin S. Hirsch of the Harvard Medical School, is in moving promising agents into clinical trials: "There is simply not enough funding available at the Federal level to get the job done. There are patients who should be in clinical trials, who are willing to be in clinical trials, who can't be." Until such trials are carried out, no amount of basic research on the molecular biology of the virus will lead to a therapy. —*Ari W. Epstein*

Dial-a-Doc

When is a medical "specialist" in the Yellow Pages a specialist?

Some people who let their fingers do the walking to find a physician in the Yellow Pages will soon have better information on which to base a decision. Beginning in March, Yellow Page telephone directories in several regions of the country will distinguish physicians who are certified by a recognized specialty board from those who are not.

The American West Directory Company of Costa Mesa, Calif., decided to include certification information supplied by the American Board of Medical Specialties (ABMS) after Julia M. Reade of the Massachusetts General Hospital and Richard M. Ratzan of the University of Connecticut Health Center published a study indicating that many self-designated specialists lack certification from a recognized specialty board. The study, published in the *New England Journal of Medicine,* showed that 12 percent of the specialists listed in the Yellow Pages in Hartford, Conn., were not certified by any of the specialty boards recognized by the ABMS.

It might be supposed that someone who advertises as an internist, for example, has some specialized training in internal medicine. Actually the law allows licensed physicians to describe themselves as they please. Some Yellow Pages do, however, already carry warnings to the effect that an advertised specialty does not necessarily reflect certification.

The law makes it difficult for the profession to police itself. For many years the American Medical Association opposed advertising, but it gave in after the Federal Trade Commission sued it in 1975 for unlawful restriction. The AMA's Council on Ethical and Judicial Affairs now allows advertising that does not mislead or deceive.

Finding a physician who is board-certified in a specialty can be confusing. The AMA's *Directory of Physicians* includes self-designated specialties as well as board certifications. Only 23 specialty boards are recognized by the AMA and the ABMS. Yet according to the ABMS's executive vice-president, Donald G. Langsley, there are about 70 self-designated "boards" that are not recognized, and some categories in Yellow Pages do not correspond to any recognized board.

Reade notes that many noncertified specialists did complete their specialty training but never took a certification examination. Of greater concern are those who advertise themselves as specialists in fields in which they have no particular training whatever. —*T.M.B.*

Setting the Pace

Are many pacemakers implanted unnecessarily?

In 1983 more than 120,000 pacemakers were implanted in the U.S. at a cost of some $12,000 each. Designed to maintain steady contractions of the heart muscle in patients whose heartbeat is arrhythmic or exceedingly slow, these small electrical devices are thought to be lifesavers for the millions who have them.

Not necessarily, say investigators

at the Albert Einstein Medical Center in Philadelphia. In the *New England Journal of Medicine* Allan M. Greenspan and his colleagues write that less than half (44 percent) of the pacemakers implanted in Philadelphia in 1983 were clearly justified on medical grounds; 36 percent were of questionable medical value and 20 percent had no medical justification whatsoever.

The study was based on an in-depth review of the medical charts of 382 Medicare patients who received pacemakers from January through June of 1983 at 30 hospitals in the Philadelphia area. An eight-member panel of experts, which included four board-certified cardiologists, reviewed the patient data and concluded that one in five pacemaker insertions was unnecessary. That held true for every hospital included in the study, regardless of its size or academic affiliation.

Pacemaker implantation is a relatively uncomplicated, low-risk procedure. Experts acknowledge that the pressures to implant the devices have been high. Physicians and hospitals stand to gain sizable financial rewards from the procedure; when in doubt, moreover, physicians are likely to recommend it for fear of malpractice suits.

In 1986 Medicare enacted a Quality Assurance and Utilization Review program and now requires advance authorization for every Medicare-funded pacemaker procedure. According to Roman W. DeSanctis, who is director of clinical cardiology at the Massachusetts General Hospital, "they are pretty strict about it." Although he is impressed by Greenspan's data, he thinks that inappropriate implantation has decreased significantly since 1983 and that the study must be repeated with 1987 data to be meaningful.

Greenspan intends to do just that. He agrees that the number of pacemaker implants has quite likely dropped as a result of Medicare's vigilance, and he is eager to know by how much. "One of my major concerns," he says, "is that pacemakers can be psychologically debilitating, especially when they are not really needed. For many older patients they signal the end of a productive life."

Should patients who have pacemakers they do not need rush out to have them removed? Absolutely not, say the experts. Although implantation carries a slight risk of infection, for the most part pacemakers are benign devices, unlikely to cause problems even when they are inappropriately implanted. *—Laurie Burnham*

Optrodes

Fiber-optic sensors may speed the analysis of human blood

In today's hospitals the analysis of a patient's blood requires at least 10 minutes, since a sample must be drawn from the patient and taken to a laboratory. In the future, however, patients may be monitored instantaneously and continuously by means of tiny fiber-optic sensors, called optrodes, that are inserted into the bloodstream.

Typically, an optrode has at its tip a cavity coated with a fluorescent dye. A beam of light travels down the fiber through this chamber and returns back up. When the dye comes in contact with a specific chemical—a naturally occurring component of blood, a drug or a pathogen—the resulting fluorescence alters the returning light beam in a way that can be measured by a spectrometer.

John I. Peterson of the National Institutes of Health, who invented one of the first biomedical optrodes, a *p*H probe, in 1976, notes that only one optrode-containing product has been marketed so far and that more research must be done to produce economic products. "But work is progressing," he says. Fiber-optic sensors, he explains, have several advantages over in vivo electronic sensors, which are also being developed. Optrodes are less vulnerable to corrosion in the warm, salty environment of the bloodstream than metal electrodes, they do not require electricity and are therefore safer, and they are immune to the distorting effects of electromagnetic fields.

Cardiovascular Devices, Inc., in Irvine, Calif., marketed the first optrode-based product five years ago. The machine monitors carbon dioxide, oxygen, *p*H and temperature outside the body, in blood pumped out of a patient by an artificial heart during open-heart surgery. The firm is now conducting clinical trials of a machine that monitors patients in intensive care with an optrode inserted into their bloodstream in a catheter.

Several companies, including Eli Lilly and Company and Abbott Laboratories, are trying to build a small, portable device that by means of a catheter can monitor the glucose levels of diabetics for extended periods. Alfred R. Potvin of Eli Lilly says workers hope ultimately to build optrodes that can be combined with implantable insulin pumps. Another promising device, according to S. Michael Angel of the Lawrence Livermore National Laboratory, would perform what is called a fluoroimmunoassay. In its simplest form an optrode is coated with an antibody that fluoresces when it comes in contact with its corresponding antigen. *—J.H.*

OVERVIEW

Star-crossed

NASA suffers new problems, both technical and political

An astrophysicist at the National Aeronautics and Space Administration, in the midst of describing observations of Supernova 1987A made from NASA balloons, airplanes and satellites, remarked: "This is the one area where we really feel as though we're accomplishing something around here."

Indeed, two years after the *Challenger* disaster NASA appears more star-crossed than ever. The agency's near-term prospects are clouded by technical problems it has encountered in redesigning the shuttle's solid-fuel boosters. NASA officials have announced that the problems will delay the resumption of shuttle launches, which had been scheduled for this June, by at least two months. Other experts predict the delay will be much longer.

A lack of direction from the White House poses a more serious and far-reaching problem. President Reagan had been expected to outline a broad new vision of America's future in space in his State of the Union Address on January 25, but he mentioned only the Strategic Defense Initiative. The following day the heads of NASA and of the departments of Commerce and Transportation were scheduled to unveil a new space policy drawn up by the Administration at a press conference, but White House officials canceled the meeting an hour before it was to begin.

In an interview with SCIENTIFIC AMERICAN William R. Graham, science adviser to the president, has provided a preview of the new policy. Graham noted that predictions in the press that the policy would emphasize colonization of the moon and a manned flight to Mars, possibly with the Soviet Union, were wrong. He said the Administration had deliberately avoided choosing a particular destination as a goal around which NASA could marshal its efforts. "Despite the success of the moon-landing program," he remarked, "that has shown itself not to be a useful long-term policy."

The policy also prohibits NASA from designing and developing technologies, and launch vehicles in particular, that it can buy from the private sector. "We don't want NASA to compete with private enterprise," Graham explained. Indeed, the Administration's policy seems based on the assumption that constraining NASA will stimulate more investment in space activities by the presumably more efficient private sector. This philosophy of "privatization," according to some observers, reflects the growing influence of the Commerce Department, which since 1984 has been authorized by Congress to promote commercial space ventures. Privatization, the Commerce Department has learned, does not always produce positive results immediately. In 1985 the department turned over the management of its Landsat remote-sensing satellites to a private company. Called the EOSAT Co., it is now operating under a large deficit.

One measure of the Commerce Department's influence is the support it has won for a relaxation of Federal limits on the resolution of civilian satellite images. The limits, imposed more than a decade ago at the urging of U.S. intelligence officials, have inhibited U.S. companies from developing satellites with a resolution better than Landsat's, which is 30 meters. France and the Soviet Union now offer satellite images with a resolution of 10 and five meters respectively. Graham acknowledged that the new space policy included a proposal to "broaden the range of considerations in remote sensing" and added: "We want to ensure that U.S. entities have the ability to compete with the systems of other countries."

The Commerce Department has also lobbied effectively for funding by NASA of an orbiting laboratory station designed by a private company, Space Industries, Inc. Called the Industrial Space Facility (ISF), it is intended to be an automated materials-processing laboratory that would be visited by workers only occasionally for maintenance. Maxime A. Faget, the company's president, estimates that the ISF could be built with off-the-shelf parts for about $700 million and launched by the shuttle by 1992. NASA's own much larger and more complex space station will cost at least $14 billion and will not be completely assembled until 1997 at the earliest.

Commerce officials have urged that NASA help to develop the ISF and then lease it after it is launched. Graham, although not confirming reports that the White House had backed this view, said: "You can imagine that with our development of something as substantial as the space station there should reasonably be some activity leading up to it." But NASA officials apparently fear that the ISF might delay—if not replace—the space station, whose funding has already been deeply slashed by Congress. James C. Fletcher, NASA's administrator, recently told Congress that NASA "does not now have needs that would justify a major commitment" to the facility.

Gregg R. Fawkes, director of the Commerce Department's office of commercial space programs, insists the ISF will enable the U.S. to take advantage of the "tremendous economic potential of microgravity materials processing in the near term." He accuses NASA officials of resisting the ISF because "they're most concerned about perpetuating their existence as a bureaucracy and trying to defend their budget and turf."

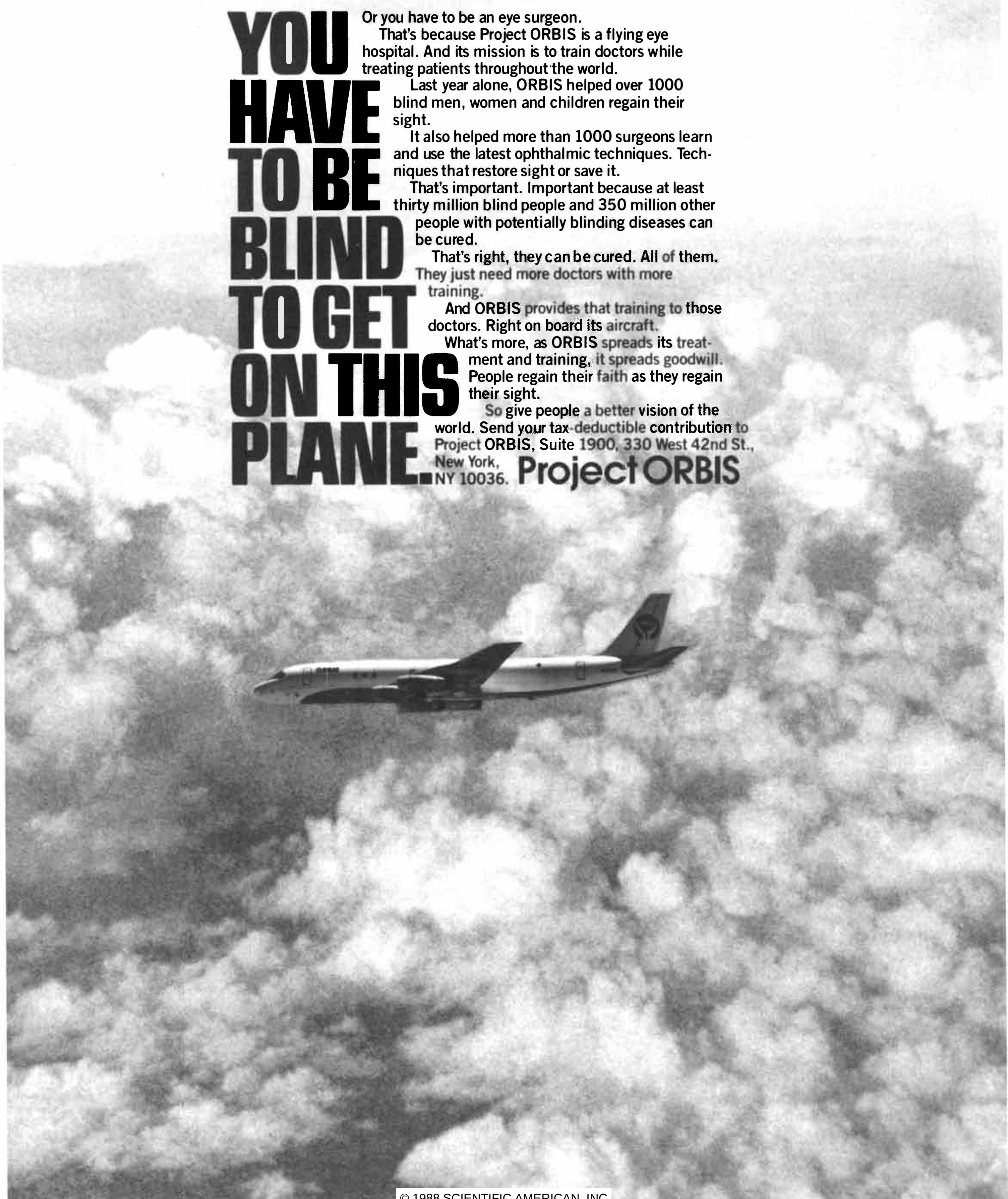
John E. Pike, a space specialist for the Federation of American Scientists, a private watchdog group, says NASA's fear of the ISF, which he calls "the Commerce Department's space station," is well founded. "We can't get funds for one space station," he says, "and now we're going to have two?" Pike disputes Fawke's views on the near-term potential for profitable ventures in space. He points out that France's remote-sensing company, the SPOT-Image Corporation, like the American EOSAT, is losing money. The communications and direct-broadcast satellite industries, although generally profitable, are nearing the saturation point, according to Pike. Of microgravity materials processing he says, "I think most people now doubt its value."

Pike faults NASA too for giving space science short shrift in its planning for the space station. He says the station will be more suited to materials processing and engineering research than to astronomy, earth sciences or plasma physics. The orbit of the space station and its exhaust, he notes, will preclude many observations. Thomas M. Donahue, chairman of the National Academy of Sciences' Space Science Board, points out, moreover, that the Office of Management and Budget has recommended that funds for two new scientific spacecraft, an X-ray telescope and a comet probe, be cut from NASA's 1989 budget. "That's a typical ordering of priorities," he complains. "Space science is the one customer NASA really has, but it doesn't get funded and all the engineering stuff gets in."

Donahue acknowledges that in at least one respect the prospects for space science have improved in the past year. Space scientists had complained that the delay-prone shuttle is inappropriate for launching space probes, which often have narrow "windows" of time during which they must be launched. Now NASA is planning to buy rockets for launching scientific satellites by the mid-1990's. Three scientific spacecraft scheduled for launching over the next three years—*Magellan, Galileo* and *Ulysses*—still must be launched on the shuttle, however. For these satellites, Donahue says, NASA's decision may be "too little, too late."

Donahue also maintains the Government should seek more cooperative ventures with other nations. "The U.S.S.R. has already preempted the high ground" in this area, he notes, by working with France, Germany and other countries. The U.S. could take a step in this direction by teaming up with the U.S.S.R. itself, Donahue says, perhaps in a project that would bring a sample of Martian soil back to the earth. Pike agrees, but he suggests such a project will succeed only if backed by a president who sees "a larger political purpose" in the space program.

Even U.S. officials trying to depict the U.S. space program in a more favorable light seem unable to avoid irony. Graham contended that "technologically there is no question that we're still ahead of the Soviets in most areas of space activity." *Skylab,* a manned station that orbited the earth in the early 1970's, was larger and "more capable" than the Soviet *Mir* station, Graham asserted. There is a second *Skylab,* he noted, "but unfortunately it is mounted to the floor of the Smithsonian's Air and Space Museum." —*John Horgan*


YOU HAVE TO BE BLIND TO GET ON THIS PLANE.
Or you have to be an eye surgeon.
That's because Project ORBIS is a flying eye hospital. And its mission is to train doctors while treating patients throughout the world.
Last year alone, ORBIS helped over 1000 blind men, women and children regain their sight.
It also helped more than 1000 surgeons learn and use the latest ophthalmic techniques. Techniques that restore sight or save it.
That's important. Important because at least thirty million blind people and 350 million other people with potentially blinding diseases can be cured.
That's right, they can be cured. All of them. They just need more doctors with more training.
And ORBIS provides that training to those doctors. Right on board its aircraft.
What's more, as ORBIS spreads its treatment and training, it spreads goodwill. People regain their faith as they regain their sight.
So give people a better vision of the world. Send your tax-deductible contribution to Project ORBIS, Suite 1900, 330 West 42nd St., New York, NY 10036.
Project ORBIS

SCIENTIFIC AMERICAN March 1988 Volume 258 Number 3

The Nonacoustic Detection of Submarines

Effective submarine-detection systems that do not rely on the vessels' sounds are not likely to emerge soon. Yet possible operating principles are under study, since a functioning system means a military advantage

by Tom Stefanick

Nuclear missiles are carried on board submarines of the U.S.S.R., France, Britain, China and the U.S. because they are relatively safe from attack as long as they are hidden underwater. The fact that the missiles are secure means that in a crisis between hostile nuclear powers there would be little incentive to launch the missiles to keep them from being destroyed in a first strike. Indeed, a nation that safeguards a capability to retaliate is not likely to suffer such a strike in the first place, because the threat of retaliation should deter an adversary from launching it. For these reasons secure sea-based strategic nuclear forces remain a key element of international security. Yet given the progress of technology, can submarines—including those armed with strategic weapons—remain concealed by the ocean in the future?

Submarines are not in themselves invulnerable; once they are detected antisubmarine-warfare (ASW) forces can get an accurate fix and attack them. Since World War I, when the first methods for detecting submarines were developed, ASW has grown increasingly sophisticated and has become an important part of any naval operation. Traditionally ASW has been carried out by specialized surface ships, aircraft, helicopters and attack submarines and has been directed primarily at the opponent's general-purpose submarines, which pose a significant threat to aircraft carriers, supply ships and other oceangoing vessels.

Over the past 20 years the U.S. Navy has redefined the wartime role of its ASW forces—particularly its attack submarines—to include threatening ballistic-missile submarines (SSBN's) of the Soviet Union and perhaps going so far as to carry out strikes against them at the start of hostilities. Such a strategy is meant to compel the U.S.S.R. to divert some of its naval forces, including its own modern attack submarines, from offensive to purely defensive missions. Moreover, actual successful attacks on SSBN's by U.S. ASW forces are seen as a way to pressure the Soviet leadership into terminating hostilities.

In this military context it is clear that the Soviet Union would benefit greatly from the unilateral deployment of an effective system by which to detect and track submarines: the task of protecting its SSBN's from U.S. attack submarines would be made considerably easier and it would be able to dispatch its own attack submarines to disrupt sea lines of communication between the U.S. and its allies. It might even be able to neutralize the U.S. sea-based strategic force, which accounts for twice as many nuclear warheads as the Soviet force. The U.S. would similarly gain a decisive military advantage if it deployed such a system: its threat to Soviet SSBN's would become considerably more potent and the sea lines of communication could be better defended against marauding Soviet attack submarines.

Of course, one does not need sophisticated military surveillance to find out where submarines are based. Even commercial earth-imaging satellites can pick out individual submarines in port. Between 35 and 45 percent of the U.S. SSBN's are in port on any given day. Soviet SSBN's are at sea only about 15 percent of the time, although many of the Soviet submarine-launched ballistic missiles can reach U.S. targets when fired from port. In the event of a crisis both sides can probably put most SSBN's out to sea within a few days.

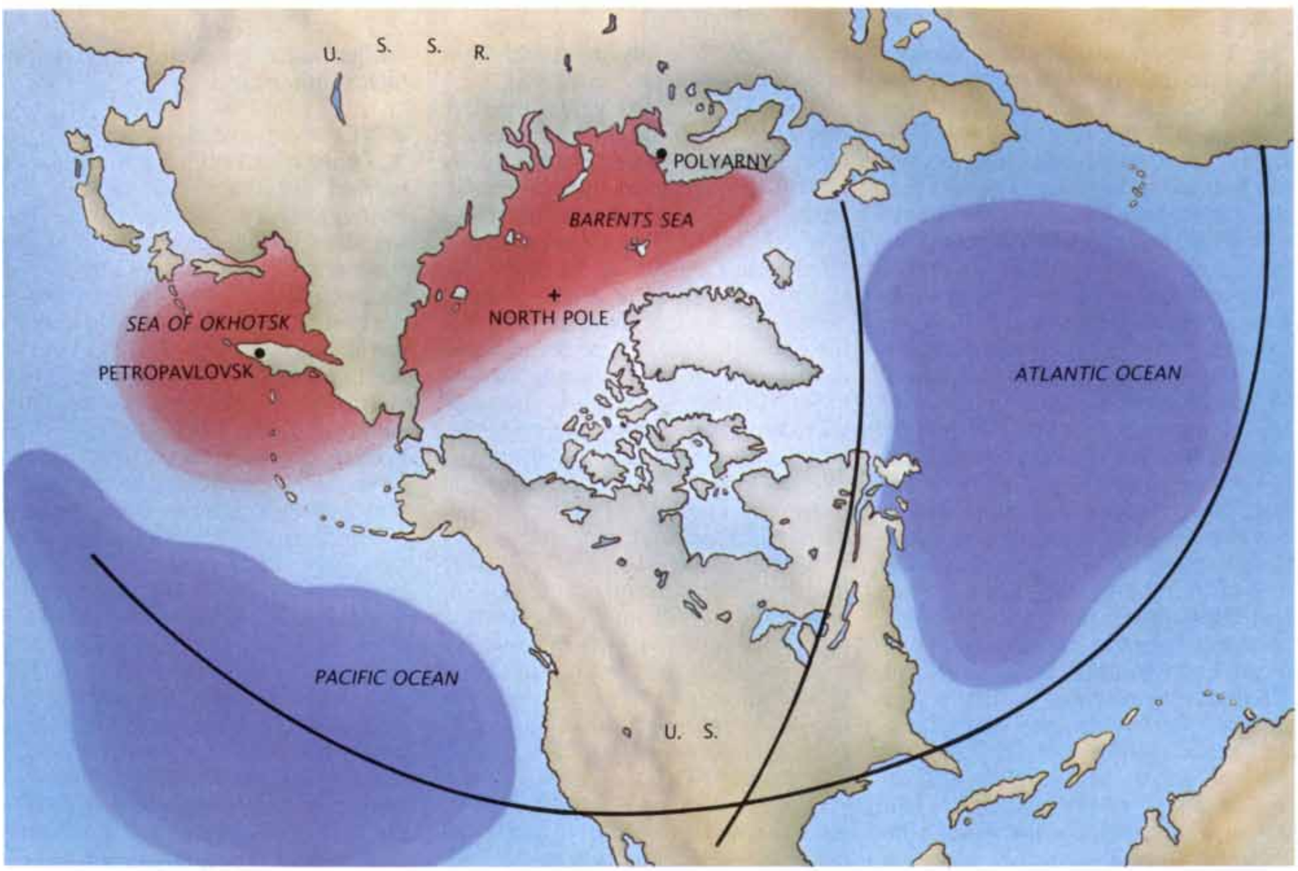
When they are at sea, U.S. SSBN's can roam submerged and hidden in expanses of the North Atlantic and North Pacific. Because the Poseidon C-3 missile has a range of 2,500 nautical miles, its warheads can strike targets in the Soviet Union from within an area of about three million square miles. The range of the Trident I C-4 missile expands the area to between 20 and 30 million square miles, since the new missile's range is 4,200 nautical miles. British and French SSBN's, which will carry hundreds of nuclear warheads by the mid-1990's, also patrol the North Atlantic. Most modern Soviet SSBN's patrol within several hundred miles of the Soviet coast in the Arctic and western Pacific, since even from such positions their missiles can hit targets in the continental U.S. Only two or three Soviet SSBN's, usually of the older *Yankee* class, are normally on patrol off the U.S. coast.

Regardless of their provenance, submarines have certain inherent characteristics, the most conspicuous of which is that they make noise as they propel themselves through water. It is because water is a very efficient medium for the propagation of sound that the commonest and most effective methods of detecting submarines have incorporated underwater acoustic sensors. There are two basic types of acoustic detection system, known as active and passive sonars [see "Advances in Antisubmarine Warfare," by Joel S. Wit; SCIENTIFIC AMERICAN, February, 1981]. An active sonar emits a pulse of sound (the familiar "ping" heard in motion pictures about submarines in World War II) and then listens for an echo produced when the sound is reflected off underwater objects. Unfortunately, in sending out powerful bursts of sound an active sonar reveals its location, making it easier for a hostile submarine to destroy or evade it. For this reason U.S. submarines rely largely on passive sonar when carrying out their missions.

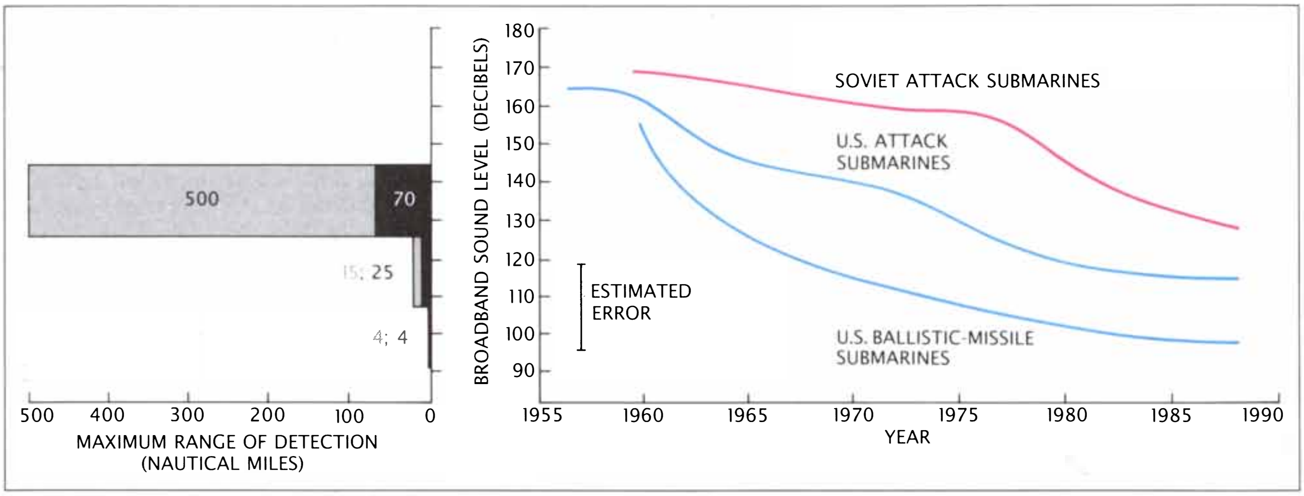
Passive sonar makes use of underwater microphones, called hydrophones, arranged in an array to monitor the sounds in the sea. By comparing the signals received by each hydrophone it is possible to determine the direction from which a particular sound originated. In addition, the sound can be compared with prerecorded sounds to determine what type of vessel may be producing it. Because passive sonar can be applied to detect, localize and identify a submarine only if the vessel makes enough noise, both the U.S. and the U.S.S.R. have implemented a number of techniques to make their respective submarines operate more quietly. They include precisely balancing rotating parts to minimize vibration; mounting machinery on sound-absorbing racks; relying on temperature-driven circulation rather than pumps to move nuclear-reactor coolant, and minimizing flow asymmetries in the wake near the propeller.

Another feature of submarines is that they are generally constructed of steel. Because massive amounts of steel disturb the earth's magnetic field locally, a magnetic-anomaly detector (a device sensitive to changes in the local geomagnetic field) can be employed to detect submerged submarines. Magnetic-anomaly detectors capable of registering the presence of submerged submarines as much as a few thousand feet away are standard equipment on U.S. and Soviet ASW aircraft.

Yet it is unlikely that such devices will be applied as a means of detecting submarines at much greater distances. (Short-range detectors based



PATROL AREAS for modern nuclear ballistic-missile submarines (SSBN's) of the U.S.S.R. (*red*) and U.S. (*purple*) reflect differences in geography and defensive tactics. Soviet submarine-launched ballistic missiles have a maximum range (*arcs*) that allows SSBN's based at Polyarny and Petropavlovsk to cover most targets in the continental U.S. respectively from areas in the Arctic and in or near the Sea of Okhotsk. The SSBN's thereby avoid having to run a gauntlet of U.S. and allied antisubmarine-warfare forces guarding the narrow passages to the Atlantic and Pacific oceans. In addition, because U.S. attack submarines could be ordered to hunt for enemy SSBN's in the early stages of a conventional war, it is prudent for Soviet SSBN's to remain near home waters, where aircraft and surface and underwater vessels of the Soviet navy can more easily protect them. U.S. SSBN's have unrestricted access to the North Atlantic and North Pacific and are quieter than Soviet SSBN's, and so they can afford to patrol unescorted in the ocean as long as the Soviet Union does not develop an effective means of detecting submarines from air or space.



NOISE LEVEL OF A SUBMARINE determines the maximum ranges at which it can be detected (*left*) by hydrophones, or underwater microphones, in water deeper than 4,000 feet (*gray*) and in water less than 1,000 feet deep (*black*). Both Soviet (*red*) and U.S. submarines (*blue*) have been made progressively quieter (*right*), but the U.S. submarine fleet still outperforms the Soviet fleet.

on high-temperature superconductors, however, may someday become useful in local ASW operations.) The strength of the magnetic-anomaly signal decays as the cube of the distance, and so even a large increase in the sensitivity of the devices yields only a small increase in the range at which a submarine can be detected. Moreover, background "noise" is often the dominant constraint on a magnetic-anomaly detector's capability. Airborne detectors face two sources of noise: naturally occurring concentrations of iron and random fluctuations in the geomagnetic field due to solar activity. The disturbances in the geomagnetic field can sometimes be so intense, in particular at higher latitudes (where Soviet SSBN's are likely to be), that they effectively mask the presence of actual submarines.

Given the advances in quieting submarines and the inherent limitations of magnetic-anomaly detectors, both the U.S. and the U.S.S.R. have established programs to study the application of other physical principles to detect submarines. In spite of the fact that it is not yet clear whether devices based on nonacoustic means of detection are even feasible, the military significance of such devices demands pursuing all possible technological avenues. This is particularly true for devices that could be based in space, allowing for oceanwide surveillance of submarines.

Although seawater is opaque to most electromagnetic radiation, blue-green light can penetrate a considerable distance through seawater. Hence one might attempt to detect the submarine optically by noting the way the hull of a submarine reflects or absorbs blue-green light. In analogy to the operating principle of radar, a laser could be made to emit an intense pulse of blue-green light so that a sensor could then detect its reflection. From the round-trip travel time of the pulse the distance between the laser and the object that reflected the laser light can be determined. On the other hand, the surface of a submarine hull may tend to absorb the laser pulse more than the surrounding water does, in which case it would be detected as a "hole" in the naturally occurring level of oceanic backscattering.

Many false targets are likely to be detected by such optical systems, since there are plenty of submerged objects other than submarines in the ocean, including marine animals such as whales. Nonetheless, such a laser-based system has apparently been applied by Sweden to detect submarines in national waters from overflying aircraft, although the system's effectiveness is not known. A space-based system would be limited by the far greater scattering and absorption of light in the atmosphere—particularly by clouds.

One can also envision potential detection methods that take advantage of other characteristic emanations from modern submarines. Particles of paint that slough off the outer surface of a submarine, minute quantities of radioactive substances that escape into the seawater through the reactor cooling system (in a nuclear submarine) or other effluents could leave a distinctive chemical trail indicating where a submarine has been. The detection of such trails would require direct measurement of very low concentrations of contaminants from samples taken at a number of depths throughout an area of the ocean. It is not known whether such detection of contaminants is possible, although even if it were, it would not provide the remote-sensing capability needed to track simultaneously many submarines distributed over a large area.

A somewhat more promising technique would seek to detect the heat emitted by submarines. A nuclear submarine relies on recycling water through a boiler, a steam turbine and a cooling system (whose heat "sink" is seawater) to convert the heat produced in a nuclear reactor into mechanical energy. As a result a nuclear submarine leaves warmer seawater behind it wherever it goes.

Although a submarine powered by a 190-megawatt reactor releases as much as 45 million calories of heat energy per second into the ocean, at a speed of five knots the dissipated energy increases the water temperature immediately behind the submarine by less than .2 degree Celsius. Because the warm water becomes increasingly diluted as the submarine travels farther away, the temperature differential quickly diminishes. It is probably about a hundredth of a degree one kilometer downstream of a submarine moving at speeds typical of normal operation.

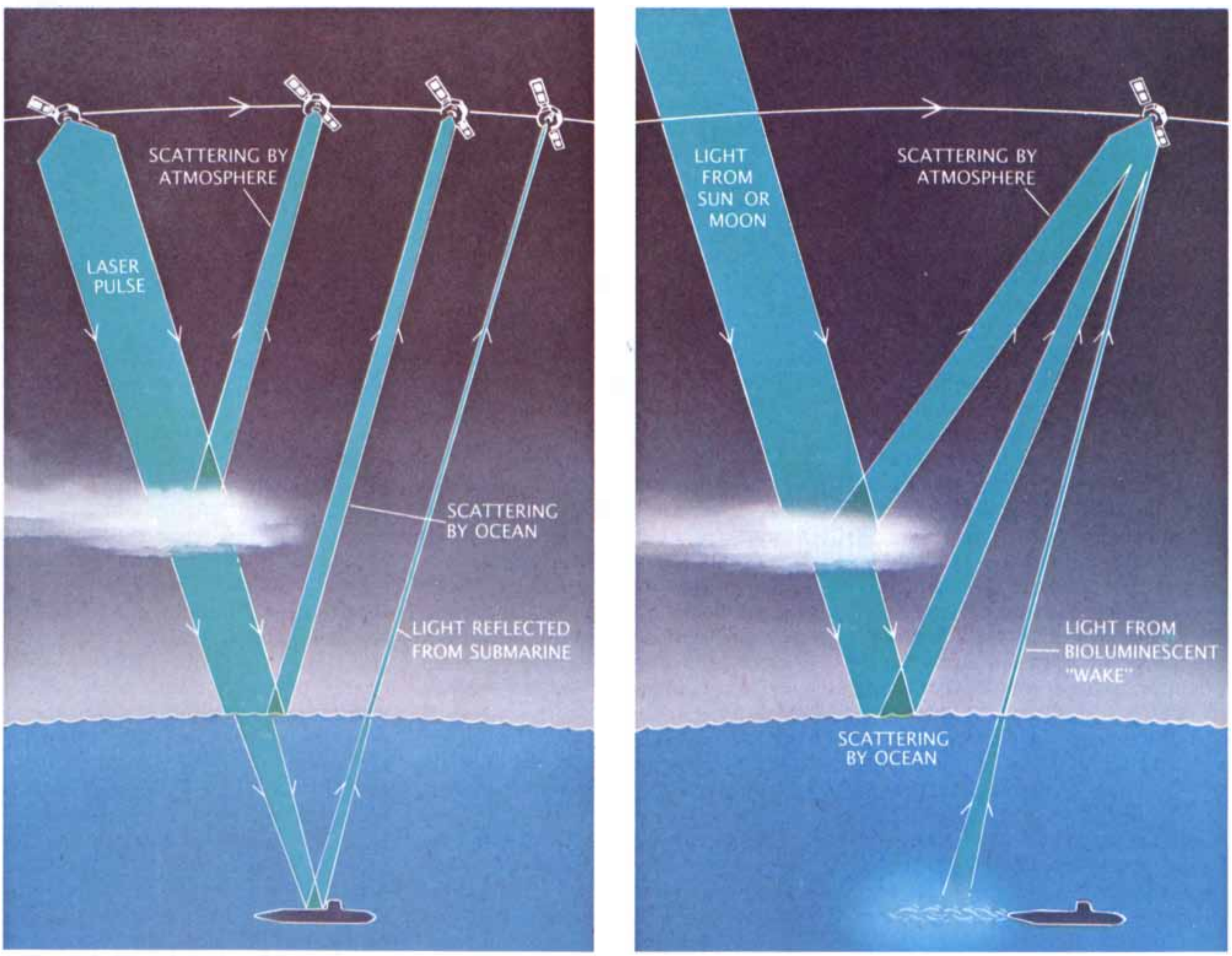
Moreover, the slightly warm water does not quickly rise to the surface of the ocean, where airborne or space-based sensors might be able to pick up the faint temperature increase. The reason is that the temperature of seawater increases (and its density therefore decreases) the closer the seawater is to the surface of the ocean. Hence the water warmed in the wake of a submarine rises only a few meters before it encounters water of the same density, and the further rise of the wake through buoyancy is suppressed.

An exception is the case of a slow-moving submarine in the Arctic, where the water temperature is nearly constant regardless of depth. In fact, Soviet submarines have been known to "ice pick": to rest against the underside of the Arctic ice cap. The maneuver enables a submarine to remain quietly in the same position without having to use its propellers to counter ocean currents; ice picking also allows a vessel to switch off its active sonar, which would otherwise be needed to avoid collision with protruding ice ridges. Unless the reactor is reduced to very low power, it seems likely that some heated water would rise to the surface of the ice, possibly giving away the submarine's position.

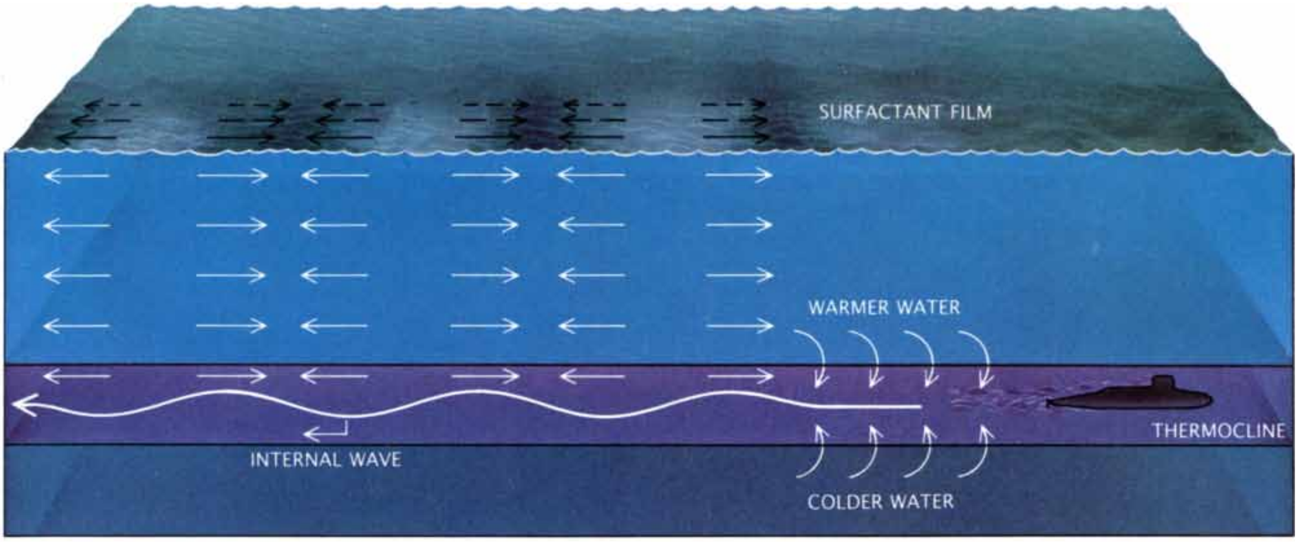
Methods for remotely detecting localized increases in water temperature would include the application of special sensors that measure the water's emission of infrared and microwave radiation, which increases as the temperature of the water is increased. Several satellites already carry such sensors. Yet the "signature" produced at the sea surface by heat from a submarine (if one exists) would be weak and exceedingly difficult to distinguish from the relatively strong infrared and microwave background radiation emitted by the atmosphere itself.

In addition to making noise, affecting the local geomagnetic field and producing heat, submarines also displace water. If a submarine travels at high speed near the surface of the ocean, it produces a telltale hump of water above it (sometimes called the Bernoulli hump) and a distinctive pattern of V-shaped waves, called Kelvin waves, behind it. (Kelvin waves are also seen in the wake of a moving surface ship.)

The size of a submarine's Bernoulli hump and Kelvin waves decreases rapidly with decreasing speed and increasing depth. For example, the height of the hump over a relatively



ABSORPTION AND SCATTERING OF LIGHT by the atmosphere and particularly by the ocean severely limit the strength of a submarine's optical signal, which could possibly be a reflected laser pulse (*left*) or the stimulated luminescence of marine organisms (*right*). In addition, light scattered either from the laser pulse or from natural sources constitutes an overpowering background "noise" from which the signal would have to be extracted by a space-based optical submarine-detection system.



CONVERGING AND DIVERGING CURRENTS (*horizontal arrows*) caused by internal waves sweep together surfactants (oils and organic fluids that are naturally found on the surface of the ocean) into "streaks" that might be detectable from the air or space. Internal waves are periodic motions of water near a thermocline: an ocean layer in which water temperature and density change sharply. The waves are generated whenever water is displaced upward or downward in the thermocline—as it can be when a submarine's turbulent wake collapses. Both the currents and the surfactant streaks can affect the amplitude and wavelength of wind-generated ripples, which in turn influence the emission and reflection of microwaves from the ocean surface.

large submarine traveling at 20 knots at a depth of 50 meters is about six centimeters, and the Kelvin waves are about two centimeters high. The Bernoulli hump of the same submarine traveling at a speed of five knots and at a depth of 100 meters is really not a hump at all, since it is only about a millimeter high, and the Kelvin waves are immeasurably small. Other hydrodynamic phenomena, such as the vortexes generated by a submarine's hull and control surfaces, might also disturb the ocean surface. Yet as long as a submarine commander takes reasonable precautions in limiting the operating depths and speeds of the vessel, submarine-generated surface displacements are not likely to be detectable.

There is an indirect effect of a submarine's hydrodynamics that could be detectable, however. The oceans are replete with organisms that emit blue or blue-green light whenever they are disturbed. Since a moving submarine creates turbulence in the water behind it, such organisms might outline its turbulent wake with their bioluminescence.

The possibility of detecting such a bioluminescent wake has attracted interest, because the phenomenon might potentially be detectable from the air or space overhead. Although light of blue-green color propagates well through ocean water compared with most other electromagnetic radiation, its intensity nonetheless is attenuated by a factor of two for every seven to 14 meters it travels. Hence light passing upward through 50 meters of water has an intensity that is only 1/140 of its original value by the time it reaches the surface, and light passing through 200 meters is reduced to less than a millionth of its original intensity.

If the stimulation of bioluminescence were limited to actual physical contact between the submarine and the light-emitting organisms, the risk of generating a detectable bioluminescent signal would be quite small, since the light would be generated at considerable depths. The concern is that the signal might somehow be generated closer to the ocean surface. Although the turbulent wake of a submarine could conceivably rise so that it would bring with it a bioluminescent signal to the surface, it is much more likely to collapse behind a submarine as a result of the stability of the ocean's stratified water densities. There is also a possibility that a bioluminescent signal could propagate toward the surface if organisms at different depths induced one another to emit light through their flashing. Relays of such empathetic responses among certain light-emitting organisms has been observed, but there is little evidence for ascertaining whether or not they might reveal the location of a submerged submarine. In fact, the geographic, seasonal and depth distributions of such organisms are not at all well known.

Even if these questions are thoroughly studied, a difficult problem remains in developing a bioluminescence-based detection system. The system must be capable of distinguishing between an anomalous bioluminescent wake and the ubiquitous background bioluminescence. There is in addition another nearly overpowering source of background noise: the blue-green component of sunlight or moonlight. Indeed, the scattering of sunlight from the ocean and atmosphere is so intense under normal daytime conditions that any space-based or airborne detection system of this type would have to be limited to nighttime use.

The oceanographic phenomenon that has stirred the greatest interest as a means of detecting submarines is internal waves: periodic variations in the density and temperature of water at depths near a thermocline, an ocean layer in which the density rises and the temperature drops sharply with increasing depth. The most pronounced thermocline is generally found in the top 200 meters or so of the ocean.

A volume of water in a thermocline that is displaced either downward or upward will experience a buoyant restoring force driving it back in the opposite direction. Hence when the thermocline is disturbed, oscillations in density and temperature are generated that propagate outward from the disturbance as waves. The natural frequency of such waves is known as the Brunt-Väisälä frequency, and its reciprocal (the wave's period) defines the fundamental time scale for oscillatory motions of the thermocline. The Brunt-Väisälä period varies greatly with time and location but is typically between 10 and 100 minutes. The displacement of water that results in internal waves can be caused by many things, including atmospheric-pressure variations, ocean currents or submarines.

By means of optical techniques it may be possible to measure the motion of layers of microorganisms that scatter light. Because the layers would move with the passage of an internal wave, such techniques could be applied to detect submarine-generated internal waves, although the distribution and properties of these biological scatterers have not been catalogued. Other possible detection methods under consideration rely on subtle surface effects associated with internal waves that may under certain conditions be rendered visible. An internal wave creates horizontal currents at or near the sea surface that are in phase with the wave. Above the crest of an internal wave there is a current moving in the direction opposite to that of the wave; above the trough of an internal wave there is a current moving in the same direction as the wave. As a result the surface currents converge and diverge above an internal wave.

Where the currents converge, surface films of oils or organic fluids known as surfactants are swept together, forming surfactant "streaks." Although the mechanisms involved are not entirely clear, the surfactant streaks along with the surface currents created by energetic internal waves can apparently produce large-scale patterns in wind-generated ripples on the sea surface by modulating the amplitude and wavelength of the ripples.

How could such patterns on the surface of the ocean be made visible? *Seasat,* an ocean-imaging satellite launched in 1978, revealed to the scientific community at large what Goverment investigators probably already knew: surface patterns associated with currents, seabed topography and internal waves many meters below the surface of the ocean can be imaged by a space-based microwave synthetic-aperture radar. Such a radar aims pulses of microwaves obliquely at the earth, off to the side of the satellite ground track and in a direction perpendicular to the direction of the satellite's motion [see "Radar Images of the Earth from Space," by Charles Elachi; SCIENTIFIC AMERICAN, December, 1982]. The pulses overlap to illuminate a wide swath of the earth (100 kilometers in the case of *Seasat*) with microwaves, some of which are then reflected back to the satellite.

A pulse's round-trip travel time establishes the range of an object being imaged. Objects having equal ranges are then distinguished by the reflected pulse's Doppler shift (the change in frequency caused by the motion of the spacecraft in relation to the object), since objects ahead of the satellite shift the microwaves to higher frequencies, whereas those behind it shift them to lower frequencies. Each point in the radar swath contributes backscatter returns to several different "snapshots" taken along the satellite's orbital path over a short period. If the objects being imaged do not move appreciably in that time, then the microwave radar's effective aperture (a measure of the resolving power of the radar that is normally given by the length of the radar's antenna) is much longer than the actual length of the radar's antenna.

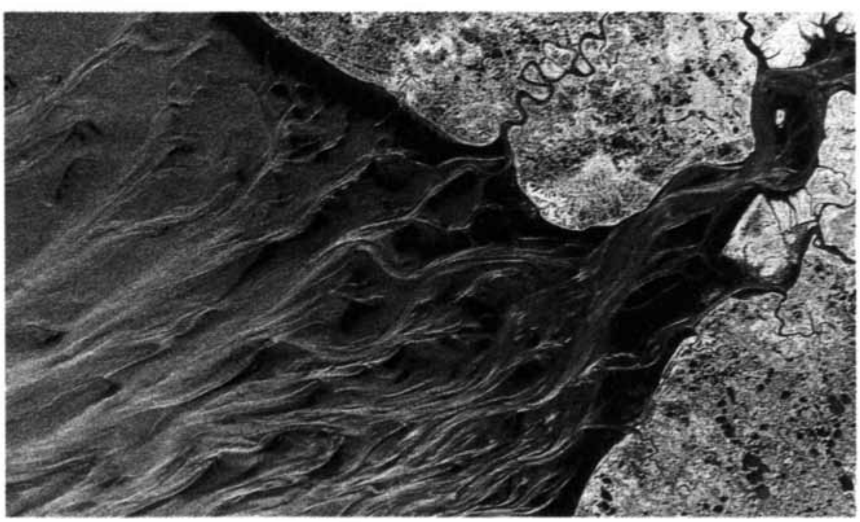


IMAGE MADE BY *SEASAT,* a satellite launched in 1978, shows a tangle of wisps at the mouth of the Kuskokwim River in Alaska that probably reflects an intricate pattern of currents, seabed topography and perhaps even internal waves many meters below the water's surface. The image was generated by processing data from a synthetic-aperture microwave radar (*see illustration on opposite page*) and has a resolution of 25 meters.

Although the *Seasat* images confirmed that under some conditions synthetic-aperture radar can detect modulations in ocean-surface roughness, the mechanism by which this occurs is not clearly understood. The explanation most frequently invoked is that the strength of the radar return from the sea surface can be modeled according to Bragg scattering theory. According to this theory, microwaves interfere constructively when they are reflected from structures that are regularly spaced by a distance equal to the wavelength of the microwaves and interfere destructively when they are reflected from all other structures. In spite of the fact that there are few experimental data, it is apparent that Bragg scattering alone is insufficient to account for the backscattering of microwaves from the ocean surface. Recent theoretical work has better predicted the results of experiments by assuming a model of ocean backscatter that is more comprehensive.

Beyond questions pertaining to the basic model of radar backscattering, the military utility of a synthetic-aperture radar would depend on the ability to form images and process them in real time, that is, immediately after the microwave echoes are received. Images of the sea surface can be obtained either through digital signal processing or through optical techniques that apply a series of lenses to process the raw radar data in an analogous fashion. While digital processing is more flexible, optical processing of radar data is much faster and may be adequate for a special-purpose, ocean-imaging satellite. In addition sophisticated pattern-recognition processing will be necessary to distinguish the internal-wave pattern generated by a submarine from the patterns generated naturally.

Another method of measuring the roughness of the sea surface is by means of passive microwave radiometry: measuring the microwave radiation emitted naturally by the surface. The sea-surface radiation in a given frequency range is a function not only of the surface temperature but also of the surface's emissivity, which in turn is a function of its roughness and the presence of surfactant films. Soviet research on microwave remote sensing suggests that combining active systems, such as radars, and passive systems can improve the ability to image the surface effects of underwater features.

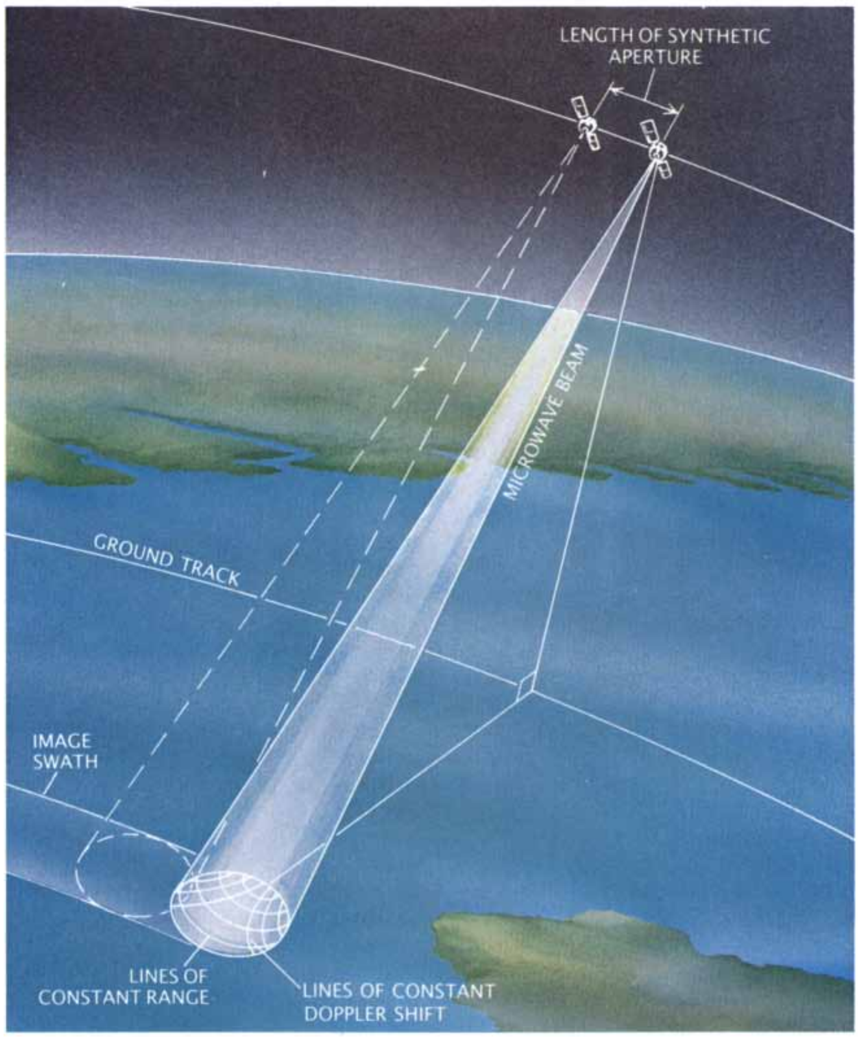
If submarine-related surface effects could be reliably detected by a space-based ocean-imaging radar—and I ought to emphasize that such an achievement is purely hypothetical at this time—there would quite likely be daunting problems in tracking a target in the presence of false targets and noise. Even such systems as magnetic-anomaly detectors and sonars, which have benefited from several decades of research and development, are severely limited by those problems. Moreover, potential countermeasures on the part of the opponent cannot be ruled out, although in order to invent cost-effective ones the opponent would have to understand the mechanisms by which a detection system operates. Countermeasures can be of several types, including changes in tactics, decoys and jamming (overloading a sensor with powerful artificial noise).

Because nonacoustic submarine-detection technologies are likely to be complex, they will probably be developed initially for use in aircraft, which can fly below clouds, make repeated passes over the same area and return to a base for repairs and maintenance. The amount of data processing needed for airborne systems, which cover smaller areas of ocean, is also significantly less than it is for space-borne systems. Since the Soviet navy puts a premium on defending its SSBN's from U.S. attack submarines in home waters, Soviet naval planners have an additional reason to apply advances in nonacoustic technology as soon as possible on a small scale in the Arctic and the western Pacific, near the U.S.S.R.

If the U.S. intelligence agencies have a good understanding of the basic principles of nonacoustic detection of submarines, they should be able to interpret relevant Soviet technological advances during these early stages of development and deployment. According to the former director of Naval Intelligence, Sumner Shapiro, it is "difficult to know" the general state of Soviet nonacoustic technology when it is in the planning and laboratory phases, but it "is easy to collect information. . .when [the Soviet Union] is operating and testing something."

It must also be kept in mind that the space platforms associated with an oceanwide surveillance system are large and highly visible. In order to have regular coverage of the North Atlantic and North Pacific, many satellites may be needed.

Based on the information available in the unclassified literature, a breakthrough in the means of detecting submerged submarines does not appear imminent. On the other hand, basic scientific questions regarding the possible detection of subsurface water motions remain unanswered. Without such answers one cannot absolutely dismiss, for instance, the possibility of applying microwave radars in detecting submarines. A broad research program, including basic unclassified research, is therefore crucial for assuring the Government and the public as well that sea-based strategic nuclear forces will continue to be secure—particularly if the U.S. and the Soviet Union negotiate deep cuts in strategic arms.



SPACE-BASED SYNTHETIC-APERTURE RADAR aims pulses of coherent microwaves obliquely at the earth's surface and records the reflected microwaves. The delay between the transmission of a pulse and the reception of its echo determines the range of an object. Objects at the same range are found along circular arcs centered on the point directly below the radar. An echo's Doppler shift (the change in frequency caused by the motion of the radar in relation to the reflecting object) serves to distinguish objects that have equal range, since objects that produce equal Doppler shifts are arranged along hyperbolic arcs whose focus is the point directly below the radar. The intersecting arcs act in essence as a coordinate system that pinpoints the location of objects.

Gravity and Antimatter

Newton and Einstein both maintain that an object's gravitational acceleration is independent of its mass and substance. Recent ideas challenge that notion; an antiproton experiment could provide a test

by Terry Goldman, Richard J. Hughes and Michael Martin Nieto

No one has ever dropped a single particle of antimatter. Yet most physicists assume that it would fall to the ground just like ordinary matter. Their arguments are based on two well-established ideas: the equivalence principle of gravitation and the quantum-mechanical symmetry between matter and antimatter. Today this line of reasoning is being undermined by the possibility that the first of these ideas, the principle of equivalence, may not be true. Indeed, all modern attempts to include gravity with the other forces of nature in a consistent, unified quantum theory predict the existence of new gravitational-strength forces that, among other things, will violate the principle.

Such effects may have been seen already in recent experiments. These effects are small in ordinary experience with matter in the earth's gravitational field. With respect to antimatter, however, these new forces could result in large anomalies. Hence an experiment to measure the gravitational acceleration of antimatter could be of great importance to the understanding of quantum gravity. We are members of an international team that has been formed to carry out such an experiment.

If the principle of equivalence is found to be violated, it will be a significant event in the history of physics, because the principle is the foundation on which the gravitational theories of both Newton and Einstein rest. The principle states that two objects fall with the same gravitational acceleration regardless of their mass or material composition. The idea was first expounded in this form by Galileo, who based his conclusions on experiments with inclined planes and on mathematical conjectures about the motion of projectiles. Einstein, recognizing that mass and energy are equivalent, extended the principle in his general theory of relativity to apply not only to objects with rest mass but also to all forms of energy, including light. Einstein's bold conjecture was verified in 1919 by Arthur Eddington, who measured the bending of light in the sun's gravitational field during an eclipse.

With the discovery of the positron, or antielectron, in 1932, a new question arose: Does antimatter obey physical law in the same way as ordinary matter does? The currently accepted answer came in 1957 with the CPT theorem of Gerhart Lüders, who proved that the mathematical operations that transform the description of a particle into a description of its antiparticle leave the laws of physics intact. For general relativity, then, gravity does not make any distinction between a particle and an antiparticle; all that counts is the particle's energy. Because the energy of an isolated antiparticle is the same as that of the corresponding particle, antimatter should fall to the ground precisely like ordinary matter.

This conclusion is valid if one believes Einstein's general relativity is the ultimate theory of gravitation. Recently, however, many physicists have proposed alternatives in which gravity may interact with aspects of matter other than energy, such as quantum number. Within these views the CPT theorem can only state that an antiapple would fall to an antiearth in the same way as an apple falls to the earth. It says nothing about how an antiapple falls to an earth of ordinary matter. In other words, one should not assume that the principle of equivalence applies to antimatter. Indeed, this should come as no surprise, because the two fundamental constructs on which the conventional argument was based, gravitation and quantum mechanics, have never been successfully joined in a single theory.

The unification of quantum mechanics and gravity has become the Holy Grail of 20th-century physics. No one has succeeded but, remarkably, the most realistic theories today all predict new types of gravitational interactions that might indeed cause an antiparticle to fall to the ground differently from an ordinary particle. We therefore proposed a few years ago to measure the gravitational acceleration of antiprotons. Such an experiment would provide an unambiguous test, if new gravitational interactions do in fact exist. Appara-

GALILEO stands at the center of this fresco, explaining the uniform acceleration of a sphere rolling down an inclined plane. Such experiments led him to reason that

tus for the experiment is now being built, and if all goes well, the results will be available early in 1991.

It has been only recently that investigators, guided by new experiments and concepts, have found certain anomalies suggesting that the equivalence principle may not hold true under certain conditions. For centuries, however, physicists had good reason to believe in the principle: it has survived many rigorous tests that proved its accuracy to extraordinary precision.

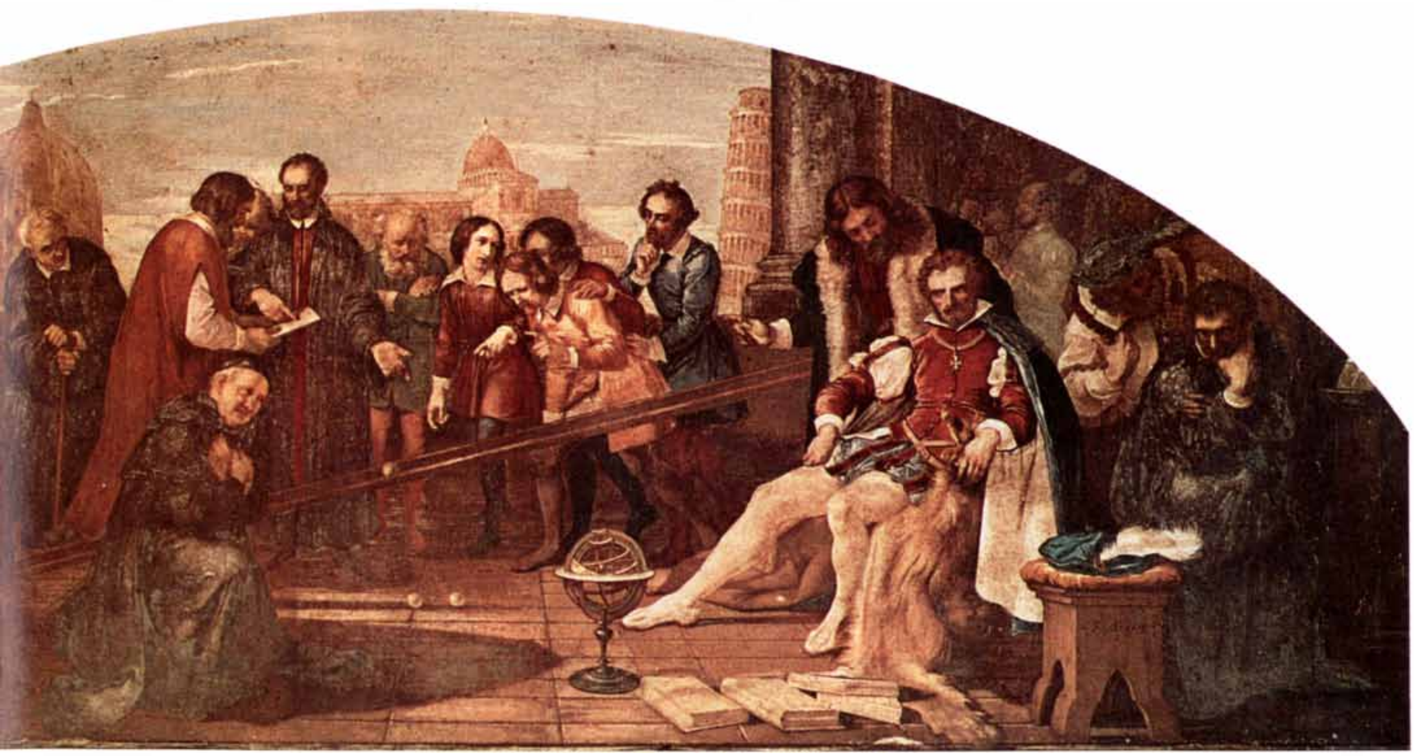
Newton himself tested the equivalence principle with experiments. To understand his experiment one needs to restate the principle somewhat. Newton introduced the concept of mass in two contexts. In his second law of motion, the force on an object is equal to the inertial mass times the acceleration. In his law of gravitation, the force of attraction between two objects is proportional to the product of their gravitational masses and inversely proportional to the square of the distance between them (and is therefore called an inverse-square law). The inertial mass is a kinematic quantity having to do with motion. The gravitational mass, on the other hand, is a "charge": an object feels a gravitational force in proportion to its gravitational mass, just as it would feel an electromagnetic force in proportion to its electric charge.

Although they are completely different concepts, Newton maintained that the two types of mass, inertial and gravitational, are equivalent. To test the idea he did experiments with pendulums. A pendulum's period of oscillation is given by the product of two factors: one depending on the length and another depending on the ratio of inertial mass to gravitational mass. Newton found that the period was always determined by the length-dependent factor alone and that the ratio of the two types of mass was always unity; in other words, gravitational and inertial mass are equal. In this way Newton verified the equivalence principle to a precision of one part in 1,000. (In the late 1820's Friedrich Wilhelm Bessel carried out experiments that increased the precision to one part in 60,000.)

The next major advance in testing the principle of equivalence was made by the Hungarian baron Roland Eötvös. In the 1880's this master geophysicist developed a torsion balance with which to probe the earth by measuring variations in the gravitational field. With the balance he could map local gravitational fields and thereby infer local mass anomalies, the most interesting being those associated with mineral deposits. The balance was so accurate that in spite of the long time needed for each measurement, it was a standard geophysical tool well into this century.

Eötvös realized that his torsion balance could be employed to test the principle of equivalence by placing objects of different materials at opposite ends of the balance. The net force acting on each object is a combination of the gravitational attraction of the earth, which is proportional to the gravitational mass of the object, and the centrifugal force due to the earth's rotation, which is proportional to the inertial mass of the object. If two different materials were put on the balance and the ratio of the gravitational mass to the inertial mass for one did not equal that ratio for the other, the Eötvös balance would rotate.

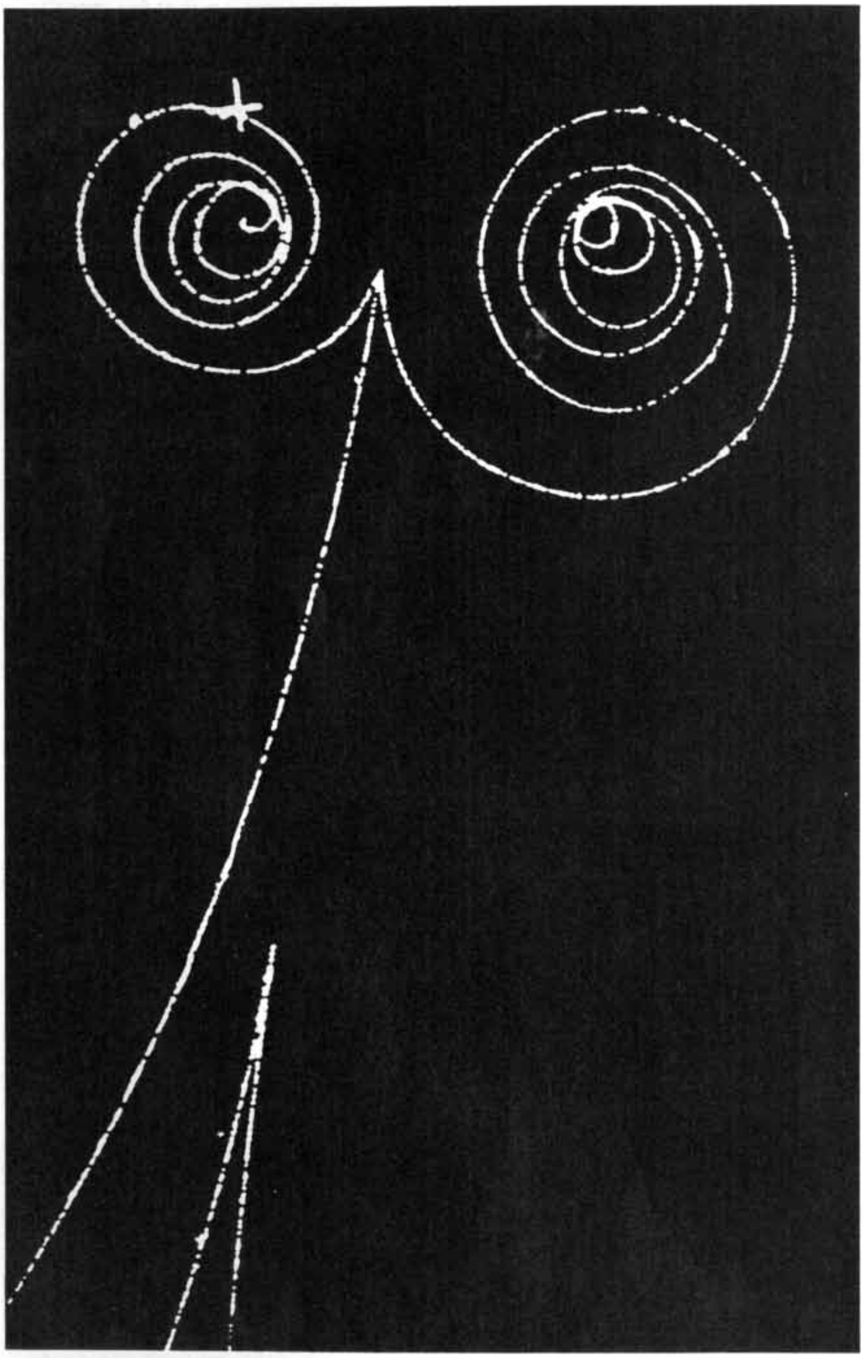
In 1890 Eötvös published results showing the equality of the gravitational and inertial masses of several substances to a precision of five parts in 10^8. He improved this work in 1909 and concluded that an appropriate limit for the precision of his experiment was five parts in 10^9. At less than this limit he found discrepan-



if objects of differing mass and substance were dropped from a height, they would strike the ground at the same instant. This principle of equivalence, fundamental to gravitational theory, is enshrined in popular mythology by the story in which Galileo dropped two stones from the Leaning Tower of Pisa, seen in the distance. A modern test of the principle will measure the gravitational acceleration of antiprotons. According to supergravity and string theories, antiprotons should fall faster than protons.

cies between different types of material, which he attributed to experimental error. (We shall say more about this below.)

In the 1960's and 1970's Robert H. Dicke and Vladimir B. Braginsky performed independent Eötvös-type experiments in which they measured the ratio of inertial mass to gravitational mass of objects in the gravitational field of the sun rather than of the earth, and they found the principle of equivalence was accurate to five parts in 10^{11} and 10^{12} respectively. A similar precision was obtained for the equality of gravitational mass and inertial mass of the earth and the moon from measurements, using corner mirrors left on the moon's surface by Apollo astronauts, of the lunar orbit around the earth as the earth-moon system itself orbits the sun. This last experiment, incidentally, demonstrated that the sun's gravity acts on the gravitational energy that binds matter in the earth—in short, that gravity attracts gravity.



ELECTRON-POSITRON PAIR leaves a bifurcating spiral trace in its wake in this cloud-chamber photograph. Created by a gamma ray colliding violently with a hydrogen nucleus, the electron and its antimatter partner have the same mass but opposite electric charge and so curve in opposite directions in the magnetic field of the chamber. Particles and their antiparticles have opposite quantum numbers; hence if there is enough energy available to conserve momentum and provide mass (in accordance with $E=mc^2$), they can be created in pairs because their net quantum number will be zero. A particle and an antiparticle can also annihilate each other in a burst of energy. Antiprotons are created in accelerators by smashing high-energy particles into suitable targets.

The idea that gravitational energy is itself subject to the force of gravity is a consequence of the revolution in physics that occurred at about the time of Eötvös' experiments. This was the formulation of the special and the general theories of relativity by Einstein. Because the present debate over the existence of new gravity-related forces results from attempts to extend Einstein's theory of gravitation, it is worth reviewing the history of this work in some detail.

The general theory of relativity emerged from Einstein's attempt to resolve a fundamental challenge that his earlier work on special relativity posed for Newton's concept of gravity. For although Newtonian theory is perfectly adequate for most practical purposes (such as sending people to the moon), it is unacceptable to the theoretical physicist because it assumes that gravity acts instantaneously over infinite distances and so violates special relativity's stricture limiting the velocity of everything—objects, energy, the propagation of a force—to the speed of light.

In classical relativistic field theory, forces are made to adhere to special relativity by the introduction of a field, which carries energy and momentum between interacting particles of matter (such as electrons and protons) at a speed no faster than that of light. It is the exchange of the energy and momentum carried by the field that produces the force experienced by the particles. For example, oscillating electrons in a transmitting antenna produce a field of radio waves, which propagate through space and exert a force on electrons in the receiving antenna.

With the advent of relativistic field theory, it was natural for physicists to study the possible forms a gravitational field could take. James Clerk Maxwell, for example, noting that both gravitation and Coulomb's law (which describes the force between electrically charged particles) are inverse-square relations, wondered whether his own theory of electrodynamics could be modified to describe gravity. Of course, some changes were necessary because the electrodynamic force produces repulsion between like charges, whereas gravity produces attraction. Maxwell was able to satisfy this condition by changing the sign of the field energy and making it negative, but he quickly realized that the resulting theory harbored a fatal flaw: with negative field energy a system would gain energy by gravitating, and so its energy would increase infinitely.

The first mathematically consistent, relativistic theory of gravitation was constructed in 1913 by Gunnar Nordström, before Einstein's general theory of relativity. Nordström's theory agreed with all gravitational experiments of the time. Einstein and Adriaan D. Fokker discovered that Nordström's gravitational field equations really described a spacetime that was curved—it was as though one had tried to describe the surface of a sphere in terms of a flat plane and then realized that the same surface could be described more naturally with spherical coordinates. Nordström's theory therefore unwittingly introduced the idea of a curved spacetime.

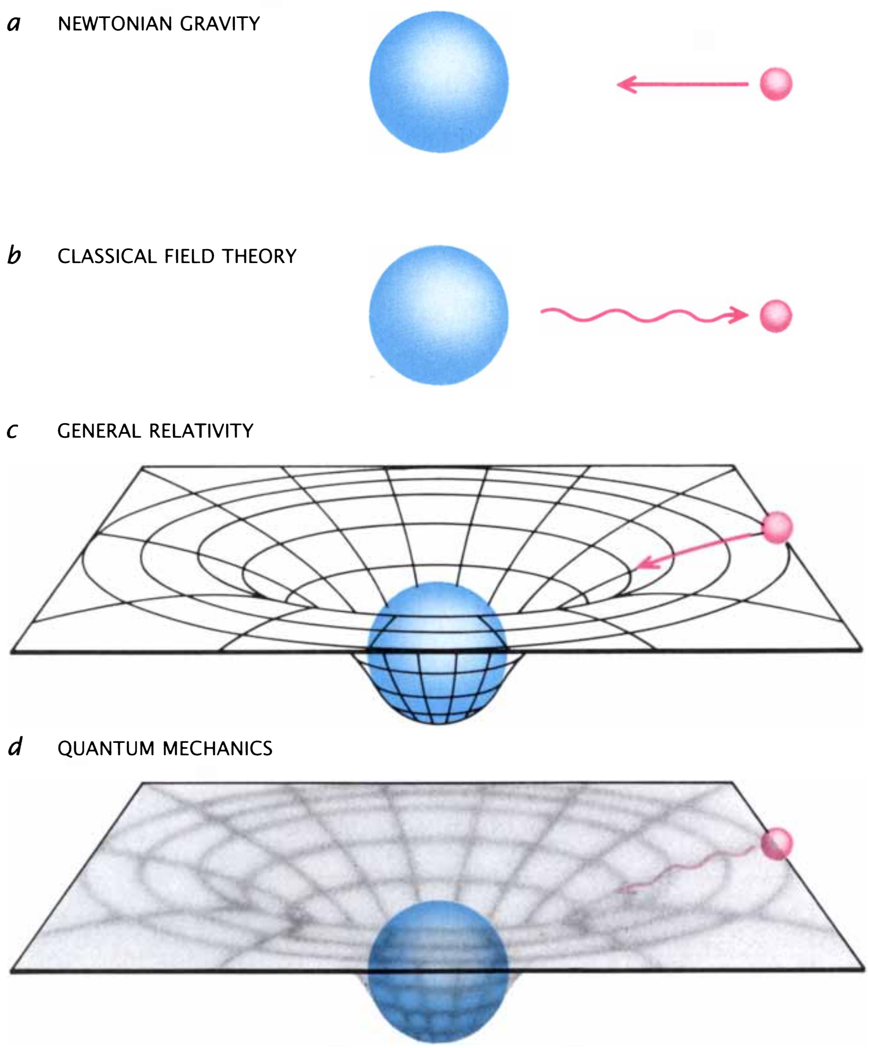
In Nordström's model the gravitational field acted only on objects that have a rest mass. But according to special relativity, energy is equivalent to mass. Therefore why should not energy also be subject to the force of gravity? It was this concept, which Einstein described as the happiest thought of his life, that led to the general theory of relativity.

Einstein's theory also resulted in a curvature of spacetime in the vicinity of massive objects. Moreover, because the theory describes a gravitational field that couples to energy and momentum rather than to mass, it predicted that gravity should deflect massless forms of energy, such as light. Einstein proposed that the effect could be tested by measuring an apparent shift in the position of stars near the limb of the sun during an eclipse. The prediction was confirmed in Eddington's 1919 expedition to Africa. Einstein's theory also accounted for the anomalous shift of the perihelion of Mercury and the red shift of light coming from stars or planets (later verified in the Pound-Rebka experiment).

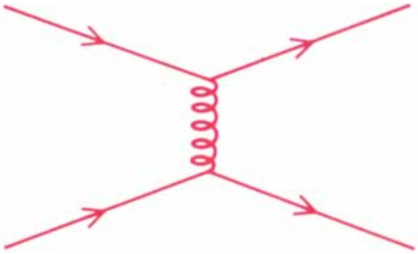
Theories of gravity such as Nordström's and Einstein's, in which the force manifests itself through spacetime curvature, are now known as "metric theories." In such theories force is manifested by a curvature of spacetime, which shapes the trajectory along which an object moves. Formulated in this way, the force is independent of the composition of the objects on which it acts. (Note that this is another way of stating the equivalence principle.)

The general theory of relativity, with its revolutionary and astoundingly successful concept of the universe, stands as one of the greatest intellectual achievements of this century. Yet its exalted status should not blind one to the fact that it is not unique in its ability to account for all experimental tests of gravity. Indeed, there is now evidence of gravitational effects that appear to violate the principle of equivalence, which, if the evidence is correct, cannot be explained by classical general relativity. These effects may, however, be consistent with more recent theories—work that has resulted from the effort to unify general relativity with that other great pillar of modern physics, quantum theory.

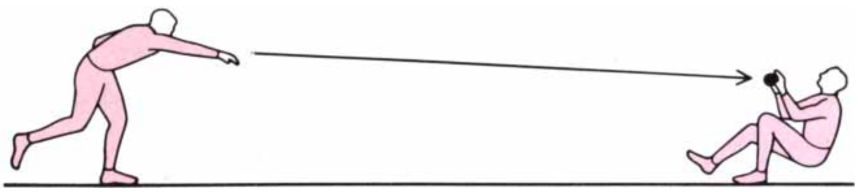
There is, in fact, a serious incompatibility between quantum mechanics and the principles of equivalence that underlie classical, non-



THEORIES OF GRAVITY describe the force between two masses. Newton, who related the magnitude of gravitational force to the objects' mass and the distance between them, assumed that the force acts instantaneously over distance (*a*). The special theory of relativity, however, proved that nothing moves faster than the speed of light. Classical field theory introduced the idea of a field that propagates force at finite speed (*b*). Einstein recognized that the field equations for gravitation describe a spacetime that is curved near massive objects. In his general theory of relativity gravity is manifested by the motion of objects along paths that follow the shortest possible distance in a curved spacetime (*c*). Quantum mechanics asserts that the path is indeterminate (*d*). This inconsistency between quantum theory and general relativity still plagues physicists.



RANGE OF FORCE	
PARTICLES WITH MASS	FINITE
PARTICLES WITHOUT MASS	INFINITE (INVERSE SQUARE)



QUANTUM FIELD THEORY introduced the idea of a particle that mediates force. Two interacting particles exchange a third particle that transfers energy and momentum from one to the other, rather in the way a thrown ball transfers energy and momentum from pitcher to catcher. Particles with mass act over a finite range. Massless particles, such as photons (light) and gravitons (carriers of gravity), act over infinite distance.

quantum theories of gravitation. For instance, according to the equivalence principle, an object's initial position and velocity determine a definite trajectory for a freely falling object in a gravitational field. In quantum mechanics, however, the object's path is indeterminate and probabilistic [see "Quantum Gravity," by Bryce S. DeWitt; SCIENTIFIC AMERICAN, December, 1983]. Therefore the equivalence principle cannot be an exact concept within a quantum description of gravity. One should not be surprised if a quantum theory of gravitation were to include interactions that violate the equivalence principle.

Quantum mechanics profoundly altered classical field theory. The classical view holds that energy and momentum are carried by a field. Quantum mechanics asserts that this energy and momentum exist in discrete units, called quanta, which can be described as particles. In quantum field theory, then, forces are said to occur through an exchange of such particles (rather the way the pitcher transfers energy and momentum to the catcher when he hurls a baseball). Electromagnetic forces, for example, are "mediated" through the exchange of photons, or light quanta. The strength of the resulting force is given by the particles' "coupling strength" to the matter that is being acted on.

The force-carrying particles have a definite rest mass (zero in the case of photons) and an intrinsic spin, or angular momentum, which can take integer or half-integer values. All the familiar forces—gravity, electromagnetism, the weak force responsible for radioactive decay, the strong force that binds atomic nuclei—are mediated by integer-spin particles, which produce forces with distance ranges determined by the inverse of the particles' mass. Forces mediated by massive particles, such as the weak force, act only over a finite range. Forces mediated by massless particles, such as electromagnetism and gravity, have an apparently infinite range, and the force diminishes in strength inversely as the square of the distance between the interacting particles.

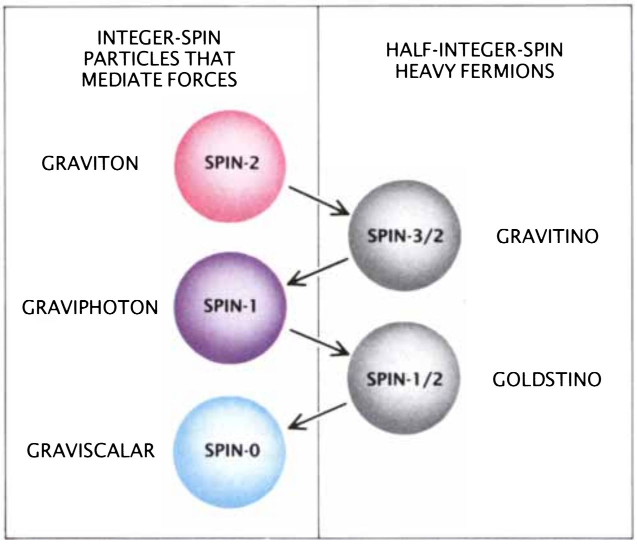
It is now known that the spin of a field is related to the nature of the force: fields with odd-integer spins can produce both attractive and repulsive forces; those with even-integer spins, such as scalar (spin 0) and tensor (spin 2) fields, produce a purely attractive force. Maxwell's electrodynamics, for instance, can be described today as a spin-1 field (the force is carried by the photon, which has a spin of 1). The force from this field is attractive between oppositely charged particles and repulsive between similarly charged particles.

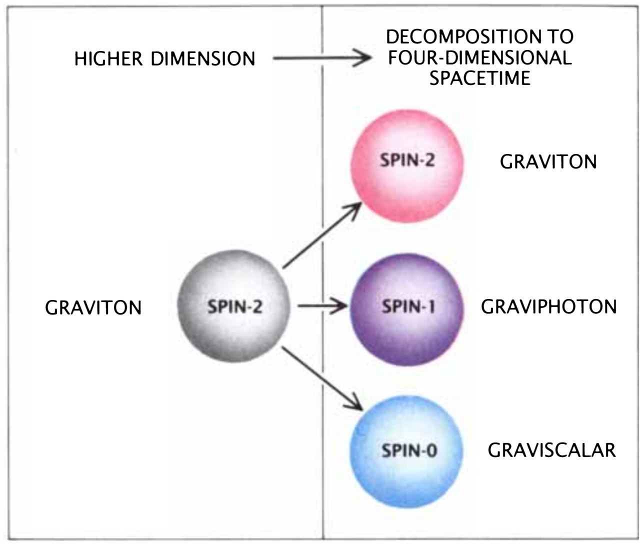
A theory of gravity, by the same reasoning, was expected to be based solely on scalar or tensor fields mediated by particles with even spin. Indeed, it has been shown that when general relativity, which is based on a tensor field, is recast as a quantum field theory, the gravitational force is carried by a massless spin-2 particle, called the graviton. Mathematically, however, the quantum version of general relativity is fraught with inconsistencies. This has led physicists to consider ways to extend general relativity in order to make gravitation more amenable to quantization.

One of the favored approaches for quantizing gravity is a class of theories known as gauge theories. These include theories widely believed to describe the strong and electroweak interactions, which are now the candidates for unification in the so-called grand unified theory. Gauge theories are based on a certain type of internal symmetry and are attractive to theorists because only a few initial parameters will allow one

		FORCE BETWEEN SIMILAR "CHARGES"	FORCE BETWEEN OPPOSITE "CHARGES"	EXAMPLES
SPIN	0 (SCALAR)	ATTRACTION	ATTRACTION	HIGGS PARTICLES, GRAVISCALAR
	1 (VECTOR)	REPULSION	ATTRACTION	PHOTON, GLUON, GRAVIPHOTON
	2 (TENSOR)	ATTRACTION	ATTRACTION	GRAVITON

INTEGER-SPIN PARTICLES mediate the familiar forces of nature. Particles of even spin produce only an attractive force, whereas particles of odd-integer spin produce an attractive or a repulsive force depending on whether the interacting matter has the same or opposite quantum numbers. For example, electromagnetic force is carried by the spin-1 photon; thus particles with the same charge repel each other and those with opposite charge attract each other. Likewise the graviphoton is expected to produce repulsion between matter and matter but attraction between matter and antimatter.





TWO THEORETICAL APPROACHES predict the existence of new gravity-related interactions that are remarkably similar. Supergravity theories that apply four or more "supersymmetry" operations (*arrows*) to the spin-2 graviton give rise to a series of new particles: the 3/2-spin gravitino, the spin-1 graviphoton, the 1/2-spin goldstino and the spin-zero graviscalar. The graviphoton and the graviscalar would mediate new forces. A separate class of ideas called metric theories, which describe forces in terms of spacetime curvature, make remarkably similar predictions: a spin-2 graviton in higher dimensions "decomposes" into a spin-2 graviton and one or more spin-1 graviphotons and spin-zero graviscalars in ordinary four-dimensional spacetime.

to calculate all phenomena within their purview [see "Supergravity and the Unification of the Laws of Physics," by Daniel Z. Freedman and Peter van Nieuwenhuizen; SCIENTIFIC AMERICAN, February, 1978].

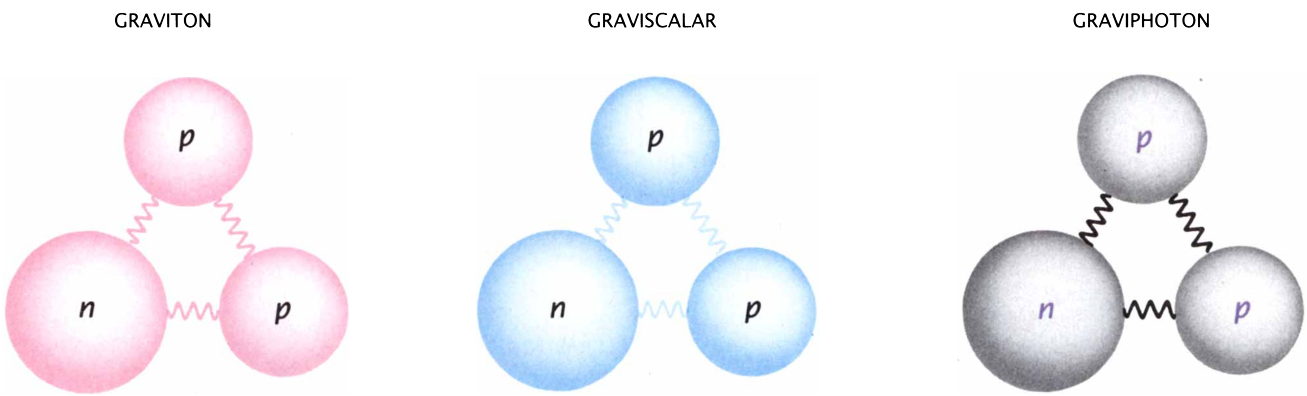
The success of gauge theories suggests that the mathematical inconsistencies in quantum general relativity might also be overcome by introducing what is now called a local supersymmetry. When general relativity is augmented by local supersymmetries, one finds (in most versions of such models) that there is a half-integer-spin particle partner for every integer-spin particle and vice versa, creating a kaleidoscopic cascade of new particles: the spin-2 graviton has a spin-3/2 partner, which has a spin-1 partner (the graviphoton), which has a spin-1/2 partner, which has a spin-0 partner (the graviscalar). (Some models describe more than one partner of each spin.) These new partners are like extra quantum states of the graviton, and their existence seems to ensure that the supergravity theories have reasonable (but perhaps still imperfect) properties for a quantum field. Indeed, all gauge theories of quantum gravity that are now being considered contain supersymmetric extra states.

The half-integer-spin particles in these so-called supergravity theories are expected to be extremely massive. Their rest-mass energy is expected to be about one trillion electron volts, or 1,000 times that of the proton. None of these predicted particles has ever been seen. Advocates of supergravity theories hope to produce them with new accelerators such as the Tevatron at the Fermi National Accelerator Laboratory, LEP (Large Electron-Positron) at the European laboratory for particle physics (CERN) and the proposed superconducting supercollider.

The integer-spin particles, on the other hand, are, like the graviton, mediators of forces and would generate new effects with a strength comparable to gravity—but with some notable differences. Both the graviscalar and the graviphoton are expected to have a rest mass and so their range will be finite rather than infinite. Moreover, the graviscalar will produce only attraction, whereas the graviphoton's effect will depend on whether the interacting particles are alike or different. Between matter and matter (or antimatter and antimatter) the graviphoton will produce repulsion; between matter and antimatter it will produce attraction. Thus for ordinary matter the graviscalar's attractive force will be more or less canceled by the graviphoton's repulsive force. Between matter and antimatter, however, both the graviscalar and the graviphoton will produce an attractive force and so will add up. It is evidence for this additional attraction that the antiproton experiment will seek.

Interestingly, similar effects are predicted by a separate group of theories that have approached the quantization of gravity from a completely different angle. Certain recent metric theories that contain more dimensions than the conventional four of spacetime also predict the appearance of the new particles. This work harkens back more than 60 years to the work of Theodor Franz Éduard Kaluza and Oskar Klein, who formulated a model of gravitation in higher-dimensional spacetime and then "projected" it into ordinary spacetime in an effort to produce a unified theory of gravity and electromagnetism [see "The Hidden Dimensions of Spacetime," by Daniel Z. Freedman and Peter van Nieuwenhuizen; SCIENTIFIC AMERICAN, March, 1985].

Ultimately unsuccessful, the Kaluza-Klein model languished for half a century, but in the past decade several theoretical physicists have considered it anew and begun to study what would happen if the model were extended to even higher dimensions. They found that a higher-dimensional spin-2 graviton would, when looked at in four dimensions, have several parts: a four-dimen-



VIOLATIONS of the equivalence principle would be allowed by new theories because the three particles that mediate gravity and related forces could couple to mass and energy with different strengths, as is suggested by varying intensities of color in this illustration. The graviton couples with equal strength to mass (*spheres*) and binding energy (*squiggles*). The graviscalar could couple to mass differently from the way it does to binding energy, and so elements with many bound protons and neutrons could experience a weaker force than those that contain only a few bound nucleons. The graviphoton could couple to internal quantum numbers such as baryon number (for example, protons and neutrons) rather than to mass or binding energy. One test is to compare the gravitational force on elements that have different proportions of protons and neutrons, because protons have a smaller inertial mass than neutrons but the same baryon number. Both graviscalar and graviphoton effects would be seen only over a finite range, which is predicted by theory to be between a few hundred meters and a few hundred kilometers.

sional spin-2 graviton, a spin-1 vector field corresponding to the graviphoton and a one-dimensional spin-0 scalar field corresponding to the graviscalar. The process is analogous to taking an arrow in three dimensions and projecting it onto a plane: two dimensions would define an arrow in the plane and the third, vertical dimension would define a point. (As with supersymmetry, some models have many partners of each spin.) Hence both the nonmetric supergravity theories and the higher-dimensional metric theories have strikingly similar consequences.

None of the quantum gravity theories has yet been shown to be mathematically consistent. Even so, Joel Scherk, not long before his death in 1980, recognized that the quantum gravity theories could have measurable physical consequences in the world of ordinary matter. For example, the graviscalar and the graviphoton are not massless and so do not behave according to the Newtonian inverse-square law. Hence one might seek experimental evidence of violations of Newton's law within the range of several hundred meters or kilometers over which the new particles are thought to exert their effect.

Experiments could also search for possible differences in the coupling strengths of the graviscalar and the graviphoton to different components of matter, such as binding energy or baryon number (the total number of protons and neutrons). The theories allow the graviscalar to couple to the binding energy with a strength different from the one with which it couples to the rest mass of elementary particles. For instance, the graviscalar force might be greater on a gram of hydrogen than on a gram of iron, because for every 56 unbound hydrogen atoms there is one iron atom containing 56 bound protons and neutrons. The graviphoton, on the other hand, must be coupled to some conserved quantum number of the elementary particles, such as the total number of baryons or quarks, or the sum of baryons and leptons (such as electrons). The graviphoton will therefore also produce a force that depends on the composition of matter. Thus both new forces can violate the equivalence principle.

The current theories, then, predict that two long-standing laws of physics will be overturned: the nonzero masses of the graviscalar and the graviphoton imply that, within their finite range, the Newtonian inverse-square law of gravitation will not be true, and the composition-dependent nature of the new forces means that the equivalence principle will also be violated.

For ordinary matter in the earth's gravitational field, the violations of the inverse-square law are expected to be minuscule. The reason is that the forces mediated by the graviscalar and the graviphoton will approximately cancel each other. A small residual effect may have been found by Frank D. Stacey, Gary J. Tuck of the University of Queensland and their collaborators in their measurements of local gravitational force made at various depths inside Australian mine shafts. After the gravitational effect of local geology had been accounted for, the measurements were inconsistent with the predictions of Newtonian theory. Instead the data were roughly consistent with the existence of a single repulsive force 100 times smaller than ordinary gravity with a range of hundreds of meters, or of both a repulsive and an attractive force with strengths about equal to ordinary gravity but canceling to one part in 100, and with ranges of up to 450 kilometers.

The Australian result, and more recent anomalous results reported by Albert T. Hsui of the University of Illinois at Urbana-Champaign and by Donald H. Eckhardt and his colleagues at the U.S. Air Force Geophysical Laboratory, are being tested independently by Mark E. Ander, Mark A. Zumberge and their colleagues at the Los Alamos National Laboratory, the Scripps Institution of Oceanography, the University of Texas at Dallas, the Amoco Corporation and the Scott Polar Research Institute. This past summer they measured gravity inside an existing borehole on the continental Greenland ice sheet, where the uniform composition of the surrounding ice helped to reduce errors in their analyses. The results should be announced within several months.

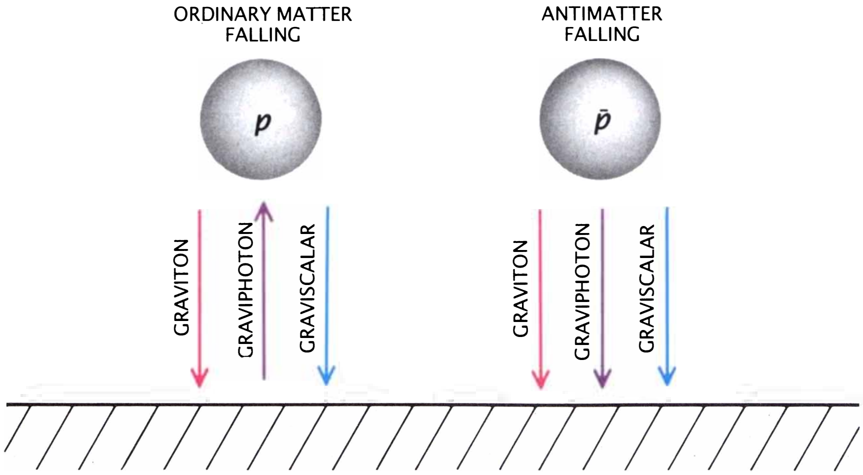
Recently there has also been much

excitement over the reanalysis of the 1909 Eötvös experiment, led by Ephraim Fischbach of Purdue University. The group found a correlation between the tiny discrepancies in Eötvös' results, mentioned above, and the ratio of baryon number to inertial mass of the various substances Eötvös had measured. The Purdue group suggests this correlation may be evidence for an entirely new "fifth force" of nature, although we think that because it is approximately the same strength as gravity, this force is a new aspect of gravity itself. The answer ultimately will come from experiment and theory.

Many new experiments have been mounted to test for composition-dependent effects. At this writing four results are in: two negative results from a group at the University of Washington led by Eric G. Adelberger and a National Bureau of Standards group led by James E. Faller, and two positive results by Peter Thieberger of the Brookhaven National Laboratory and by Paul E. Boynton and his colleagues at the University of Washington [see "Science and the Citizen," December, 1987].

All the experiments described so far test the acceleration of ordinary matter in the gravitational field of the earth. Recall, however, how things change if one replaces matter with antimatter. Such an experiment would constitute the extreme test of the equivalence principle. Here the graviphoton effect is attractive, as is the graviscalar force. Thus instead of canceling, the two effects would add up. Antimatter would then experience a larger acceleration toward the earth than matter.

In 1982, motivated by earlier discussions of this possibility, two of us proposed an experiment to measure the gravitational acceleration of antiprotons at the Low Energy Antiproton Ring (LEAR) at CERN. Since then many collaborators have joined from Los Alamos, Rice University, Texas A. & M. University, the National Aeronautics and Space Administration's Ames Research Center, the University of Genoa, the University of Pisa and CERN. The experiment would extract antiprotons from LEAR, cool them to just above absolute zero and then send them, 100 at a time, into a drift tube, where the time it takes for them to reach the top will be measured. Negative hydrogen ions, which have the same charge and nearly the same mass as the antiproton, will provide a time-of-flight measurement



NEW PARTICLES would cause matter to exert a force on antimatter different from the force it would exert on ordinary matter. The graviton and the graviscalar would produce attraction in both cases, but the graviphoton would produce repulsion for matter and attraction for antimatter. If the graviphoton and graviscalar effects are almost equal, they would nearly cancel each other for interactions between ordinary matter, but for interactions between matter and antimatter they would add up. Antimatter, then, could fall to the ground perhaps 14 percent or more times faster than matter.

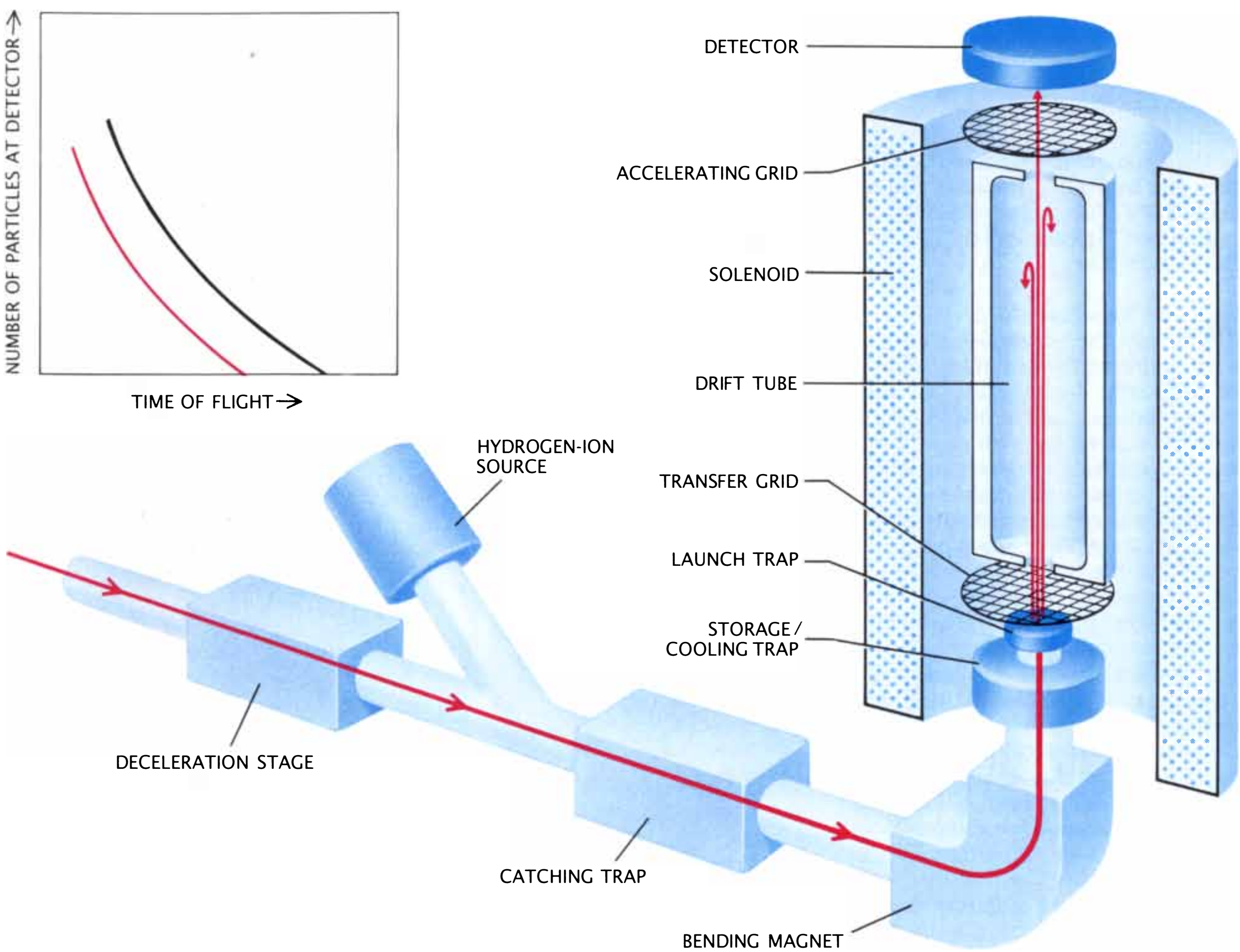
GRAVITY METER is carefully hoisted out of a deep borehole in the Greenland ice sheet a few miles south of the Arctic Circle. The half-million-dollar device measures local gravitational force at various positions to a depth of 1,600 meters with a precision of one part in 100 million. The meter, which is normally used for petroleum prospecting, is testing for the possible existence of new gravity-related interactions. Seen in the photograph are Casey Rohn of the University of Nebraska's Polar Ice Coring Office and James Wirtz of Amoco. Ted Lautzenhiser is inside the red winch-operating cabin.

for ordinary matter. The result will then be compared with the time of flight for antiprotons in order to determine whether the antiprotons are undergoing a larger gravitational acceleration. The drift-tube method stems from the pioneering work of Fred C. Witteborn and William M. Fairbank, both then at Stanford University, who in 1966 reported measuring the gravitational acceleration of the electron. Fairbank now hopes to undertake a state-of-the-art positron experiment, which would provide a complement to our antiproton experiment.

The antiproton gravity experiment is expected to achieve a precision of better than 1 percent. If there are indeed both vector and scalar interactions with coupling strengths close to the normal gravitational force and ranges of about 450 kilometers, the antiproton would fall with an acceleration 14 percent greater than that of ordinary matter. If the coupling strengths are larger than the normal gravitational force, the effect would be greater still.

Even if the experiment were to find no new effects, as the first measurement of the gravitational acceleration of antimatter it would extend experiments on gravity into new territory, just as the Eötvös-Dicke and the Pound-Rebka experiments did. These experiments got exactly the results expected by the prevailing theory of the day. Yet they were such beautiful and clear verifications of physics that they are classics—the kind of experiments that go into the textbooks.

But what if the antimatter experiment were to obtain a result that violated the classical understanding of gravitation? How would we convince ourselves and the rest of the community of physicists that the outcome was not an experimental error? When we discussed this difficulty with two members of the experimental team, Ron Brown and Nelson Jarmie, their eyes lit up. Their response says it all: "We'd love to have that problem." Whatever the outcome, all of us eagerly await the experimental results.



EFFECT OF GRAVITY on antiprotons will be measured at the Low Energy Antiproton Ring (LEAR) of the European laboratory for particle physics (CERN). Antiprotons are extracted from LEAR at an energy of two million electron volts, decelerated to between 10,000 and 20,000 electron volts and captured in the catching trap and storage trap, where they are cooled to 10 degrees Kelvin (–263 degrees Celsius). They are then launched, 100 at a time, up a one-meter-high drift tube. The antiprotons most useful to the experiment will have a starting velocity averaging four meters per second. As they drift upward the tug of gravity will slow them down. Hence the more energetic particles will reach the detector first and the less energetic ones will reach it later. There eventually will be a cutoff time after which no more particles will reach the detector because the slowest particles will not have enough speed to reach the region of the accelerating grid before their upward motion is overcome by gravity. The experiment will separately measure and compare the cutoff time both for antiprotons and for negative hydrogen ions (*black curve*), which have the same charge and almost the same mass as antiprotons. If antimatter were subject to a larger gravitational force downward than ordinary matter, the antiprotons would have a shorter cutoff time (*colored curve*) than the hydrogen.

Catalytic Antibodies

This new class of molecules couples antibodies' vast diversity with the catalytic power that makes enzymes invaluable for technology, medicine and basic research

by Richard A. Lerner and Alfonso Tramontano

Can antibodies be made to serve as enzymes? These two classes of proteins seem to have evolved for different tasks. Enzymes are distinguished by their catalytic ability, the ability to accelerate chemical reactions without being used up themselves. Antibodies are unique in their ability to recognize a diversity of substances.

Yet both classes of proteins exert their effects in much the same way: by binding to other molecules. Enzymes have a cleft or crevice on their surface, in which they bind chemical reactants in the course of transforming them. Antibodies have a specialized site that enables them to bind to molecules belonging to foreign organisms that have invaded the body, thereby marking the intruders for destruction by other components of the immune system.

The great diversity of antibodies is a reflection of the immune system's defensive function. In its ability to make perhaps 100 million different antibodies, each one capable of binding to a particular foreign substance, the immune system in effect anticipates the multitude and variety of possible invaders. The chemical reactions of life, in contrast, are stereotyped and invariant. Nature is satisfied with perhaps a few thousand enzymes, each one capable of catalyzing one or a few reactions.

Without a catalyst most biochemical reactions proceed with impossible slowness. Whether they take place in living organisms, in the laboratory or in industrial processes, these transformations depend crucially on the small array of existing enzymes. Yet for a reaction that is not biologically essential there is often no enzyme. The immune system, however, can make an antibody to almost any substance, and it is now possible to isolate a supply of pure "monoclonal" antibody that has a single target molecule. Is there any way to build on the basic similarity between antibodies and enzymes in order to give antibodies the ability to catalyze chemical reactions?

In our laboratories at the Research Institute of Scripps Clinic we have been learning about the structural details of an antibody's interaction with its target molecule. Recently this new knowledge, combined with other chemical intuitions, has suggested ways to harness for catalysis the energy with which antibodies bind to their targets. We have already produced the first catalytic antibodies, and the research could ultimately bear fruit in a virtually limitless variety of catalytic antibodies for use in biotechnology, medicine and investigations of protein structure and function.

Even though ordinary antibodies do not catalyze reactions in living organisms, they do show certain features suggesting a potential for catalysis. Working with Elizabeth D. Getzoff and John A. Tainer of Scripps and H. Mario Geysen of the Commonwealth Serum Laboratories in Australia, we found that antibodies can induce structural changes in their target molecules. We first elicited antibodies to a protein by injecting it into experimental animals. We went on to determine which parts of the original protein the resulting antibodies bound to most readily and how they gained access to those antigenic segments.

As a prototypical target protein we chose myohemerythrin (MHR), which serves in certain marine worms as an oxygen carrier. Like other proteins, MHR consists of a long chain of amino acids, folded into a specific three-dimensional form. The sequence of MHR's amino acids is known, and its three-dimensional fold was determined through X-ray crystallography by Steven Sheriff of the National Institute of Diabetes and Digestive and Kidney Diseases, Wayne A. Hendrickson of Columbia University and Janet D. Smith of Purdue University. The sequence information enabled us to synthesize peptides—short sequences of amino acids—mimicking segments of the protein. The structural information then told us which site on the protein each of the peptides represented.

We tested each peptide for its ability to interact with anti-MHR antibodies and took the result as a measure of antibody binding to the corresponding site on the folded protein. The resulting antigenic map of MHR showed that every site on the protein is the target of a few antibodies, but the most reactive sites are those at which the protein's structure is flexible and its amino acids form a convex surface, which presumably fits the concave binding pocket of one or another antibody.

We then examined the antibodies' binding preferences in more detail. To find out which amino acids are crucial to binding at a given site, we synthesized peptides mimicking the antigenic site at every amino acid but one, where a different amino acid was substituted. Again we tested the peptides for their reactivity with anti-MHR antibodies, noting the effect of each substitution. If a substitution reduced the reactivity substantially, the original amino acid must be involved in antibody binding.

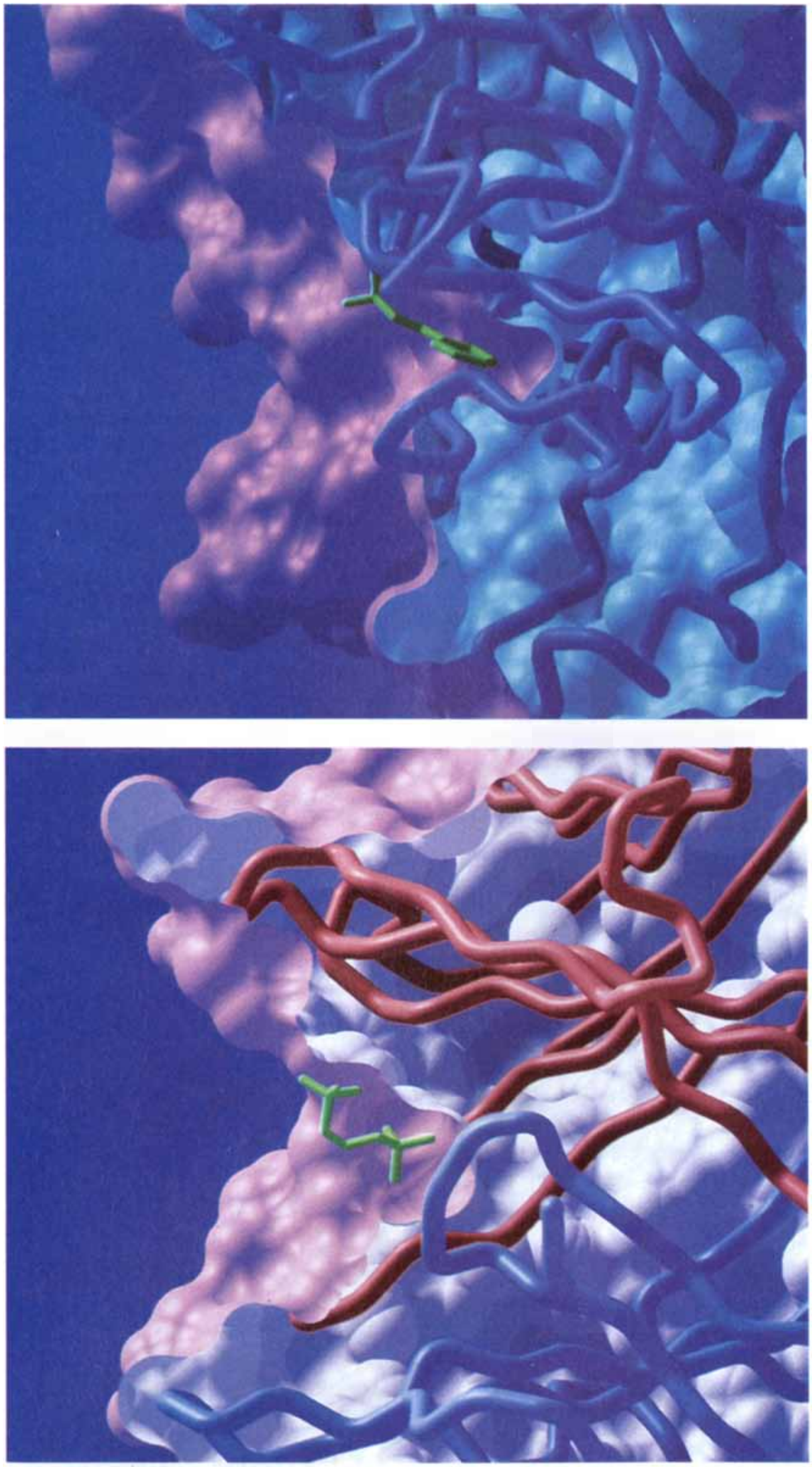
To our surprise, not all the amino acids crucial to binding can be found on the surface of the protein, where they are directly exposed to antibodies. In one sequence of amino acids representing an antigenic site, for example, antibodies must interact with the amino acids valine, tyrosine and glutamic acid to bind effectively. Replacing any one of the three sharply reduces the reactivity of the peptide. Now, the three-dimensional structure of MHR shows that only valine and glutamic acid are available at the surface of the protein; tyrosine (and in particular its bulky side chain) is buried in the protein's interior. It is hidden under the glutamic acid and another nearby amino acid, lysine, which are held together by a weak electrostatic bond. The tyrosine residue would seem to be inaccessible to an antibody interacting with the surface of MHR.

As the antibody binds, then, the tyrosine must somehow be presented on the protein surface. If the structure in which tyrosine is buried is the protein's stablest (lowest-energy) configuration, then the bound antibody must hold the protein in a relatively high-energy form. The antibody must expend energy to do so; the source of the energy must be the antibody's binding strength.

Proteins are dynamic molecules; their bonds forever stretch, twist and vibrate. It may be that the tyrosine side chain occasionally rotates out to the surface of the protein, breaking the weak bond between glutamic acid and lysine, and is stabilized there by the bound antibody. Alternatively, the initial interactions of the antibody with the protein may foster the cleavage of the bond. The buried side chain could then rotate outward to interact directly with the antibody.

In either case, the results say something of fundamental importance about antibodies and their binding ability: some antibodies may bind most readily to high-energy states of their target molecules. In stabilizing these states the antibodies in effect overcome bonds or forces that exist in the targets' low-energy form. In this, antibodies are something like enzymes. Enzymes also alter bonds in their substrates, or target molecules. To be sure, the covalent bonds broken by enzymes are much stronger than the ones overcome by antibodies. Yet enzymes bind to their targets no more strongly, on the whole, than antibodies do. If binding energy is responsible for the catalytic ability of enzymes, might antibody binding also be put to work so that instead of simply labeling their targets, antibodies actually transformed them chemically?

The way enzymes work suggests a means of doing just that. The effect



ACTIVE SITE is a binding pocket in both an enzyme (*top*) and an antibody (*bottom*). Both kinds of molecules bind to other substances. In so doing enzymes catalyze a chemical change in their target substances; antibodies merely tag their targets (molecules characteristic of foreign organisms) for recognition and destruction by other parts of the immune system. The authors have found that the binding energy of antibodies can be directed so that they too can transform their targets chemically. The computer images, showing the enzyme chymotrypsin and a specific antibody cut away to reveal the proteins' backbone (*blue at top, red and blue at bottom*), were made by Arthur J. Olson of the Research Institute of Scripps Clinic using the MCS graphics program developed by Michael L. Connolly of Scripps. Each binding pocket is filled by a small molecule (*green*).

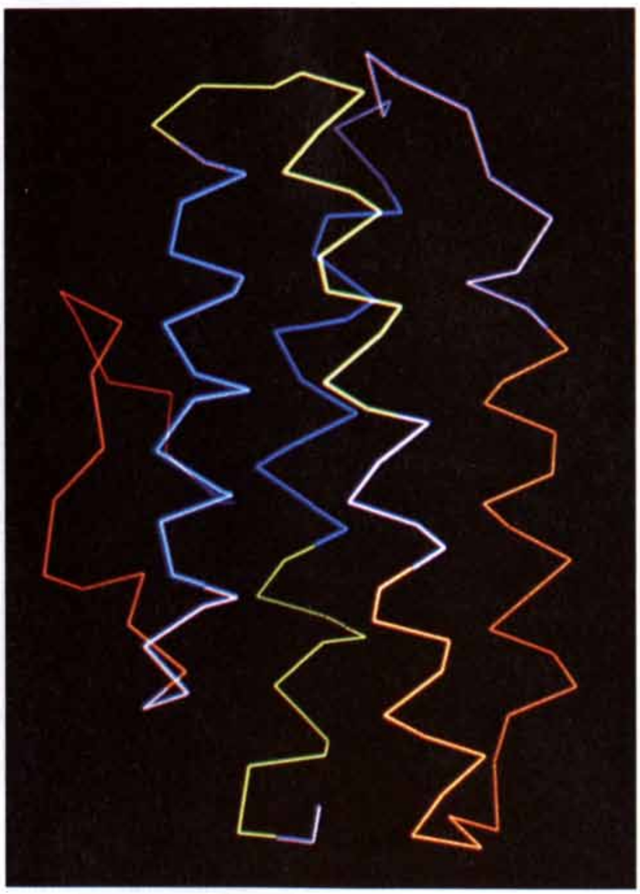
of an enzyme can be understood in terms of the energetic demands of a reaction. Chemical processes can be described by energy surfaces in which stable molecules are defined by deep wells. For one molecule to be transformed into another, its atoms must travel across the energy surface from one well to another. The atoms must first gain energy until they reach a crest and then lose energy to fall into the stable product well. The highest point on the reaction path corresponds to a dynamic, unstable transition state in which bonds are only partially formed or broken. The transition state exists for just a fleeting instant during the journey from reactants to products.

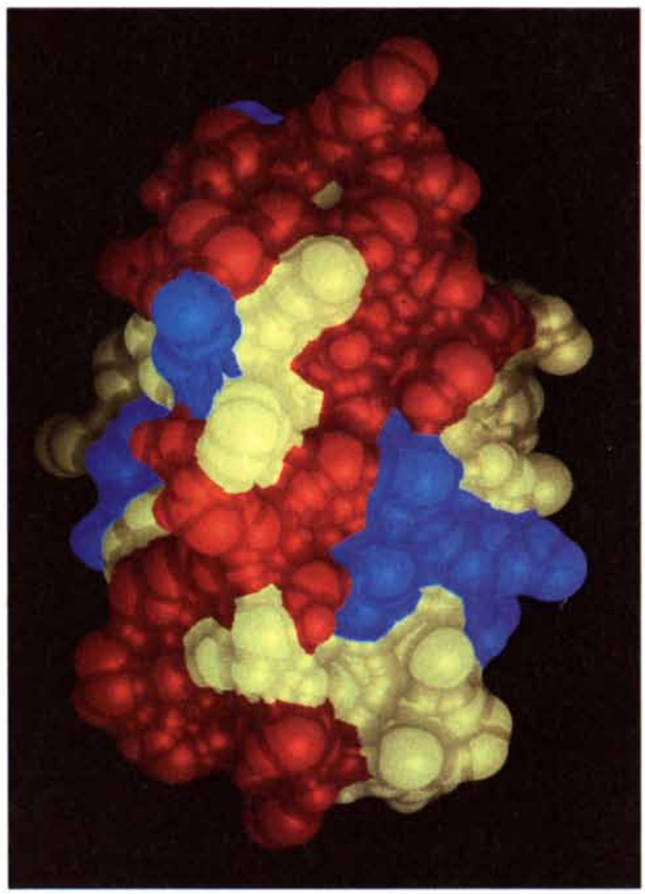
The difference in height of the points on the energy surface representing the starting materials and the transition state is the reaction's activation energy. It is the energy barrier that must be surmounted before the reaction can coast to completion. The higher a reaction's activation energy is, the slower it proceeds. An enzyme speeds up a reaction by lowering its activation energy: changing the topography of the energy surface to provide a pathway crossing a smaller energy "hill."

In 1946 Linus Pauling suggested one way in which an enzyme might lower a reaction's energy barrier: by binding most strongly not to the reactants but to the transition state. The transition state is thereby stabilized, and as a result less energy is needed to form it; the reaction is accelerated, often by factors of several billion. The enzyme's effect is catalytic because the products diffuse away from it, enabling it to bind and transform molecules of the substrate repeatedly.

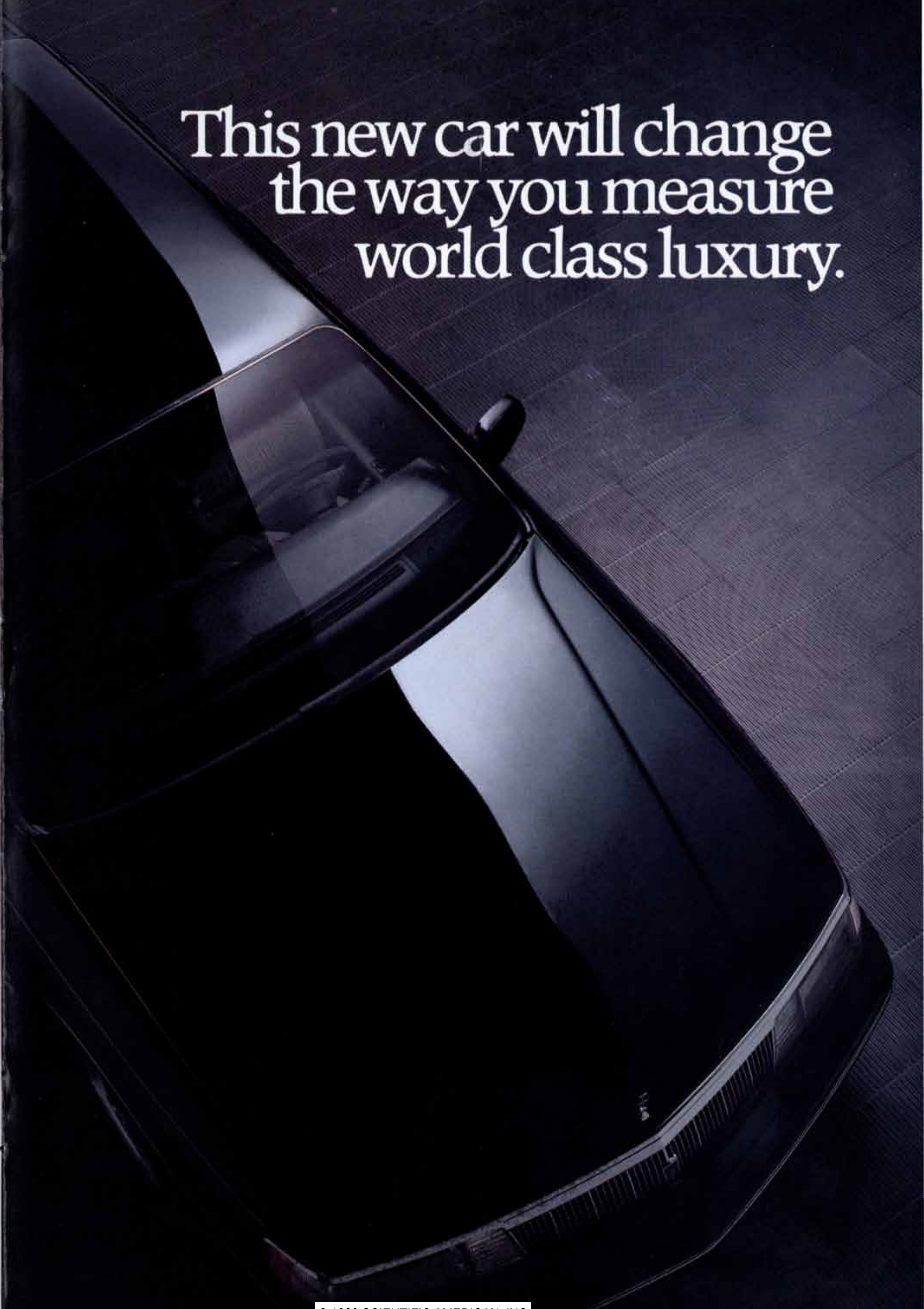
In Pauling's scheme the fundamental difference between the actions of enzymes and antibodies is that whereas enzymes bind most readily to high-energy activated states, antibodies bind to low-energy structures. Many years ago William P. Jencks of Brandeis University proposed that if one could develop an antibody to a transition state (a high-energy structure), that antibody might have a catalytic effect on the corresponding chemical reaction.

An effort to develop antibodies that bind to a transition state faces a practical problem, however. To elicit antibodies one needs an antigen, which is injected into an experimental animal to induce an immune response. Here the true antigen is unavailable: the transition state itself is so unstable that for practical purposes it does not exist. An answer to the dilemma came from another of Pauling's proposals. He predicted that, given a reaction for which an enzyme exists, a stable substance might mimic the transition state in shape and charge. Such a transition-state analogue would bind very tightly to the enzyme, inhibiting its catalytic action by filling the binding pocket and preventing it from binding to its true substrate. Over the past





MYOHEMERYTHRIN (MHR), the oxygen-carrying protein of a marine worm, was mapped to show antibodies' binding preferences. Peptide molecules mimicking short segments in the protein chain (*left*) were synthesized. The degree to which antibodies elicited by the whole protein reacted with each peptide was mapped onto the protein surface (*right*). The most reactive regions are red; regions of intermediate reactivity are yellow and the least reactive regions are blue. The antibodies preferred sites at which the surface is convex and the protein's chemical groups are highly mobile. Both traits may ensure a good fit between the binding pocket and the surface. The computer images were made by Elizabeth D. Getzoff and John A. Tainer of Scripps.


This new car will change
the way you measure
world class luxury.

LINCOLN

What a luxury car should be.

The world's most advanced luxury-car steering system. The problem with most power steering systems is that they don't compensate for a basic trait of the automobile—constantly changing speeds. An automobile carving through a turn at 55 mph ideally doesn't need as much power assist as one turning at 35 mph. Or at 10 mph. Continental's speed-sensitive variable-assist steering solves the problem by constantly monitoring vehicle speed, then adjusting power assist to match the driver's needs. Maximum assist at slow speeds for easy maneuvering, minimum assist at higher speeds

for proper road "feel" and precise cornering abi and graduated levels in between.

The world's most advanced luxury-car suspens system. It's a dual-damping electronically con- trolled air suspension. And it can actually modi Continental's over-the-road behavior to let it ex in both ride and handling. Sophisti- cated

sensors monitor lateral acceleration forces, the severity of road conditions, even the rate of acc eration or deceleration. When input from eithe the road or the driver calls for a stiffer suspensi Continental answers the call with a nine-fold increase in shock damping. Continental's perso ality changes from smooth-riding luxury car to

INTRODUCING THE NEW LINCOL

Once again, world-clas in miles, instead

C O N T I N E N T A

responsive sports sedan (or back again) in fractions of a second.

The world's most advanced braking system. Continental was engineered on the belief that how quickly and confidently you reach zero is every bit as important as how quickly you reach 60. That's why Continental is equipped with a four-wheel-disc Anti-lock Brake System (ABS). A locked wheel is less effective in stopping an automobile and virtually useless in steering one. Continental's ABS modulates brake-line pressure to keep each wheel at the threshold of lock-up. In doing so, ABS helps Continental's driver maintain steering control while simultaneously providing up to 40% shorter stopping distances under almost all road conditions.

The world's most advanced luxury car. That statement may raise a few eyebrows. But the creators of this new front-wheel-drive Continental feel the assertion is justified. For Continental truly is a bold new approach. All the comfort traditionally associated with Lincoln matched to a road integrity and commitment to quality (exemplified by a **6-year/60,000-mile powertrain warranty***) that earn it a position in the front ranks of the world's most respected motorcars. Understandably, you still might not be totally convinced. If that's the case, a comprehensive Continental Portfolio is available free by calling **1-800-251-3673** or by filling out and mailing the attached card.

After that, you may be moved to the final proof. A test drive. At your Lincoln-Mercury dealer.

*Certain restrictions and a deductible apply. Ask to see a copy of this limited warranty at your dealer.

CONTINENTAL.

luxury is measured of kilometers.

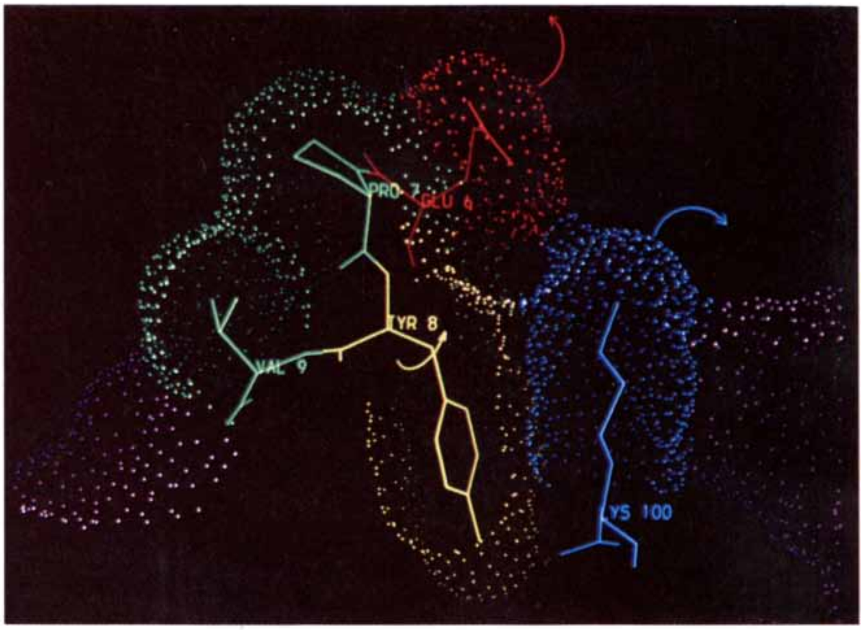
20 years a number of compounds have been synthesized that behave the way Pauling predicted.

Such a transition-state analogue, then, might serve as an antigen that would induce antibodies capable of recognizing the actual transition state, stabilizing it and perhaps acting on the substrate as a true catalyst. We began our exploration of the possibility by focusing on a reaction known as ester hydrolysis, in which a water molecule attacks a chemical group known as an ester to produce a molecule of an acid and a molecule of an alcohol. An ester group consists of a central carbon atom bound to two oxygen atoms and another carbon. The central carbon is doubly bonded to one of the oxygens, with which it forms the acid product; the other oxygen, which is destined to become part of the alcohol product, links up with another organic group. In the hydrolysis reaction the bond between the central carbon and the oxygen of the alcohol group is broken and a new bond forms between the carbon and the oxygen in water.

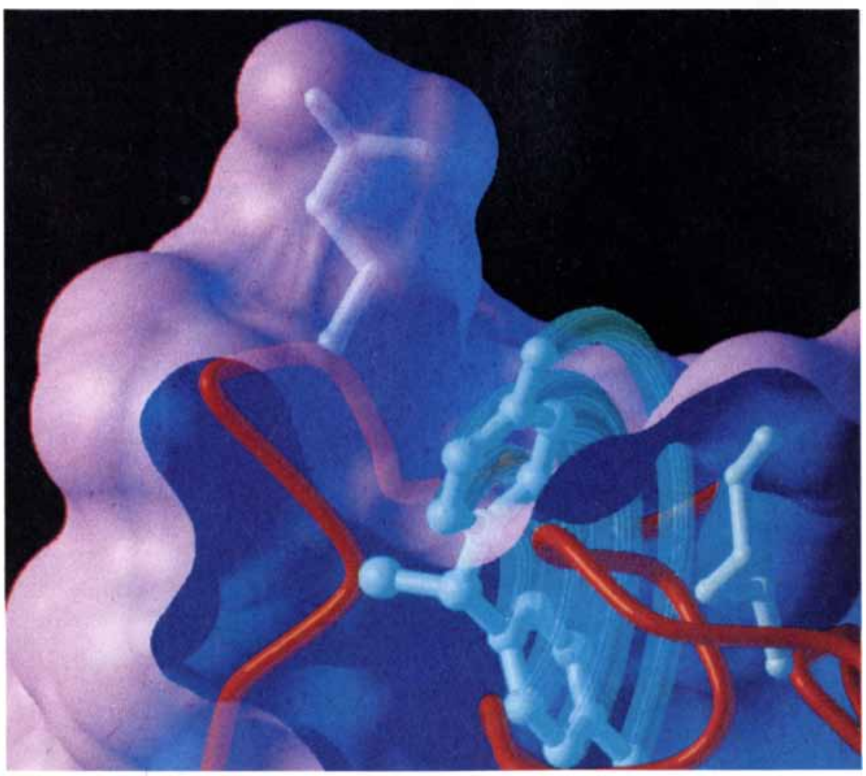
The four atoms of the ester bear little charge and lie in a plane. As a water molecule interacts with the ester, however, the reaction passes through a transition state in which the central carbon is surrounded by a tetrahedral arrangement of three oxygens, some of them carrying an electric charge, and a carbon. In the transition state the bonds not only are reoriented but also are stretched to perhaps 120 percent of their normal length. The distinctive features of the transition state indicate that the ester itself cannot serve as an antigen for inducing catalytic antibodies. The ester would elicit antibodies capable of recognizing and stabilizing only the starting material of the reaction. Such antibodies would increase rather than lower the reaction's energy barrier.

What is needed is a transition-state analogue. Substituting a phosphorus for the central carbon in the transition state's tetrahedral ensemble of atoms yields a stable compound known as a phosphonate ester. The distribution of charge on the oxygen atoms of the molecule resembles that of the transition state. In addition the phosphorus-oxygen bonds are about 20 percent longer than ordinary carbon-oxygen bonds, which enables the analogue to mimic the elongated bonds of the transition state.

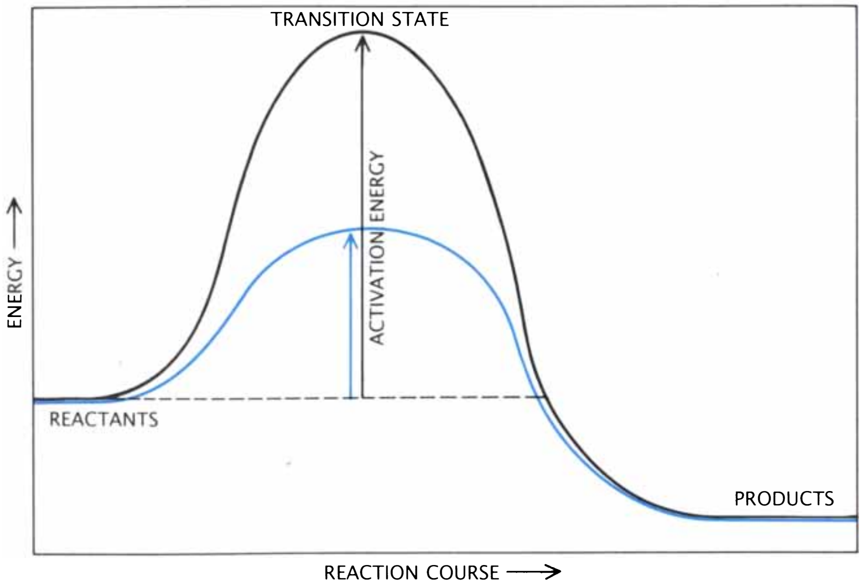
We synthesized such an analogue, coupled it to a carrier protein and



ANTIGENIC SITE on myohemerythrin includes several amino acids, three of which are essential to the ability of a matching peptide to react with anti-MHR antibodies. Two of these amino acids, valine (*green*) and glutamic acid (*red*), are on the surface of the protein, but the third one, tyrosine (*yellow*), is buried in the interior, under a weak electrostatic bond formed between glutamic acid and a fourth amino acid, lysine (*blue*). To interact with the tyrosine an antibody must somehow favor a rearrangement in which the bond is broken and the surface amino acids are parted, allowing the tyrosine to swing outward (*arrows*). The computer-graphics image was made by Getzoff and Tainer.



TYROSINE SIDE CHAIN ROTATES about the protein backbone (*red*) when an antibody binds to the antigenic site shown in the top illustration on this page. The rearrangement may occur through random thermal motion and simply be stabilized by the bound antibody; alternatively, the antibody might actively reshape the surface of the protein and induce the change in amino acid conformation. The image was made by Michael E. Pique, Getzoff and Tainer using Connolly's computer-graphics program MCS.

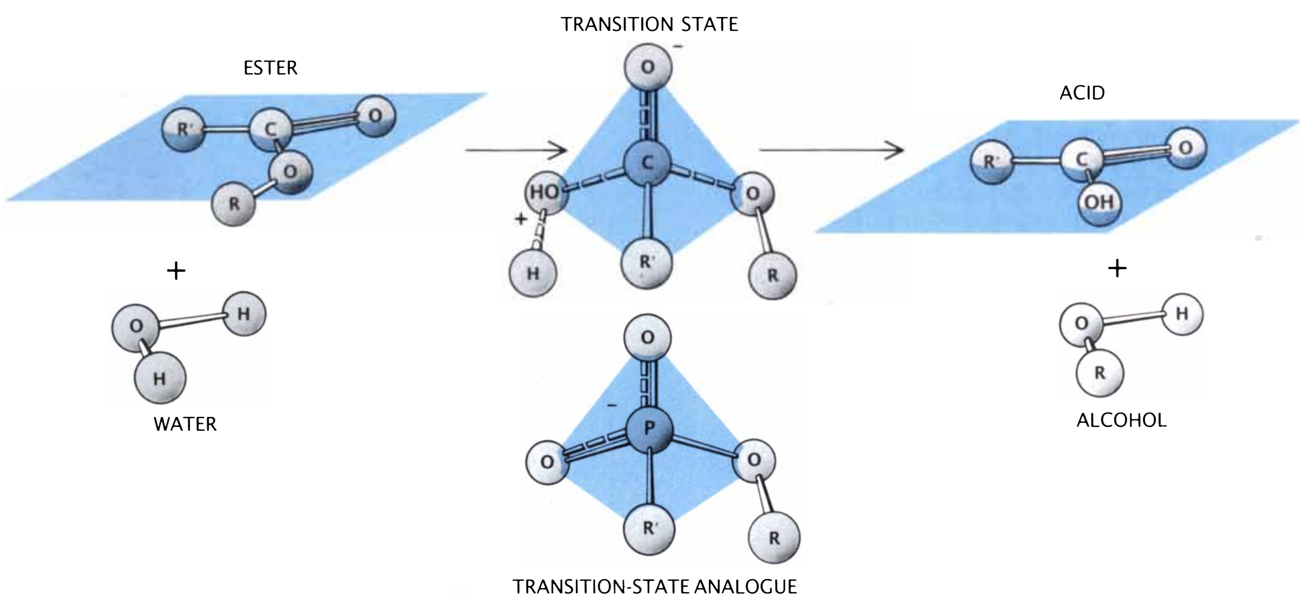


ENERGY PROFILE charts the energetic demands of a hypothetical chemical reaction, from reactants to products. From its starting point the curve rises by an amount representing the reaction's activation energy. It reaches a peak corresponding to the reaction's transition state, an ephemeral complex of atoms that has no resting state. Enzymes catalyze a reaction in part by binding to the transition state and thereby stabilizing it. The activation energy of the uncatalyzed reaction (*black curve*) is lowered (*colored curve*) and the process is accelerated, often by a factor of several billion.

immunized mice with the conjugate. We then extracted antibody-secreting spleen cells from the mice. We fused the spleen cells with tumor cells to generate clones of identical antibody-secreting cells. An antigen ordinarily induces the production of many different antibodies that bind to many different sites on the molecule, but each antibody-secreting cell (and the clone derived from it) makes just one kind of antibody. In order to find cells that were secreting antibody to the transition-state analogue and not, say, to sites on the carrier protein, we tested each monoclonal antibody for the ability to bind to the transition-state analogue.

Having identified monoclonal antibodies specific for the analogue, we tested their ability to catalyze the hydrolysis of the corresponding ester. Some of the antibodies had no effect; perhaps they were specific for a feature of the molecule that is not matched in the transition state itself. But we were gratified to find that other antibodies actually behaved as catalysts and accelerated the ester hydrolysis by a factor of about 1,000. As we expected, the phosphonate ester with which the antibodies had been induced inhibited their catalytic action, presumably by preempting the binding of the substrate. Like the recognition ability of ordinary antibodies, the catalytic activity of these molecules was highly specific: they catalyzed hydrolysis only for esters having a transition-state structure that closely matched the immunizing antigen.

At about the same time as we did our experiments Scott J. Pollack, Jeffrey W. Jacobs and Peter G. Schultz of the University of California at Berkeley carried out a somewhat different experiment based on the same principle. They began with an antibody specific for phosphoryl choline, a molecule containing a phosphorus atom tetrahedrally bound to four oxygens. The three-dimensional struc-



HYDROLYSIS OF AN ESTER passes through an unstable transition state whose distinctive shape and charge can be mimicked by a stable molecule. The ester group and the acid product, which inherits the ester's central carbon, have a planar geometry and carry no electric charge. (*R* and *R′* represent chemical groups that do not take part in the reaction.) The transition state is tetrahedral and is polarized: a partial negative charge is concentrated at one apex. A stable analogue, in which phosphorus takes the place of the central carbon in the transition state, mimics its geometry and approximates its distribution of charge.

ture of the antibody had been solved by David R. Davies and his colleagues at the National Institute of Arthritis, Metabolism, and Digestive Diseases, and it suggested that the binding pocket neatly accommodated the tetrahedral phosphate group.

The Berkeley team reasoned that this antibody might stabilize the transition state for a hydrolysis reaction and thereby catalyze the process. The group therefore set about designing a reactant that would give rise to a transition state resembling phosphoryl choline in charge and shape. To pass through a tetrahedral transition state containing four oxygens, the reactant had to contain a carbonate group (a carbon coupled to three oxygens). When the investigators synthesized an appropriate carbonate, they found that their antibody accelerated its hydrolysis many hundredfold.

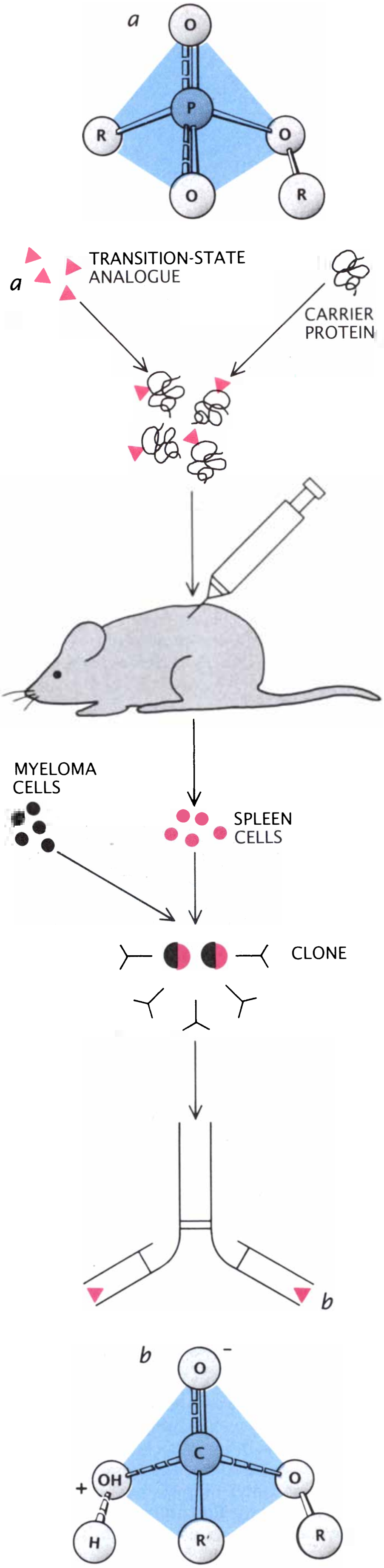
The Berkeley result supports the concept that stabilizing the transition state is the key to catalysis by antibodies. To demonstrate the broader principle, that it is possible to extract a catalyst of predetermined specificity from the immune system, one has to begin as we did, with an antigen. Our experiment points to a general scheme for developing catalytic antibodies. From study of chemical mechanisms one infers the shape and charge distribution of a reaction's transition state. Basic chemical considerations guide the design of a stable transition-state mimic, which serves to elicit an antibody with a complementary binding pocket. The development of catalytic antibodies, then, is largely a matter of designing and preparing the proper antigen.

The catalytic antibodies produced by careful antigen design can show not only chemical selectivity and catalytic efficiency but also a third property characteristic of enzymes: the ability to distinguish between stereochemically different forms of a molecule. In particular, molecules that contain a carbon atom bound to four different groups can exist in two forms that are chemically identical but are mirror images, related to each other as the left and right hands are related. The two enantiomers, or mirror-image forms, of such a chiral compound react identically with other, nonchiral substances. In a reaction between two chiral compounds, however, specific enantiomers may interact preferentially, just as there is a specific match between each hand and its respective glove.

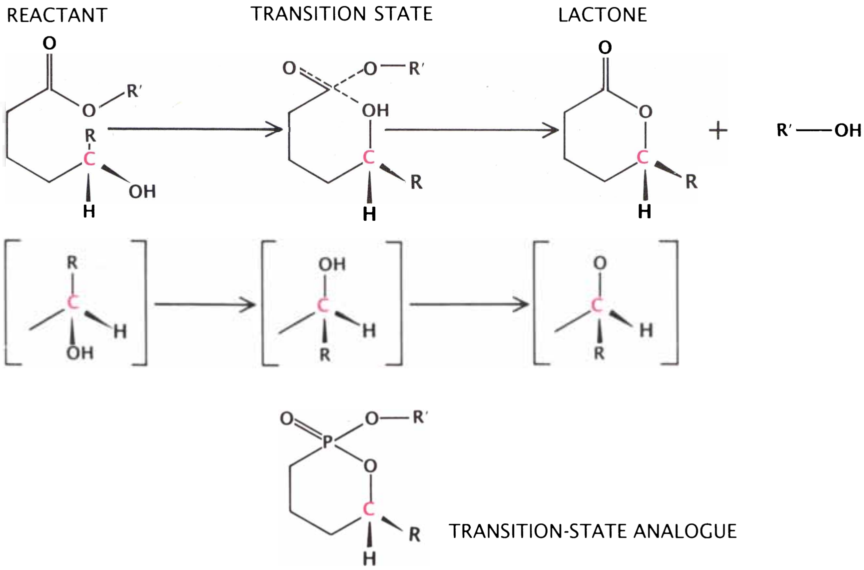
All but one of the amino acids are chiral, and so are the proteins they compose. Most living things build their proteins from only one enantiomer of each amino acid. Hence enzymes, like other proteins, exist in only one chiral form. In a reaction that has a chiral starting material or product, an enzyme often will selectively catalyze the process for only one enantiomer. One would expect catalytic antibodies—which are chiral proteins themselves—to show the same stereospecificity.

Working with Andrew Napper and Stephen J. Benkovic of Pennsylvania State University, we tested for stereospecificity by studying a reaction in which the substrate, transition state and product are all chiral. The reaction forms a bond between an oxygen at one end of a chainlike molecule and a carbon at the other end, transforming the starting material into a ring molecule called a lactone. A carbon atom in the middle of the chain is bound to four different groups and is chiral. It defines two enantiomers of the substrate, the transition state and the products. A stereospecific catalyst should preferentially bind to one enantiomer of the transition state, thereby transforming just one form of the substrate into one form of the product.

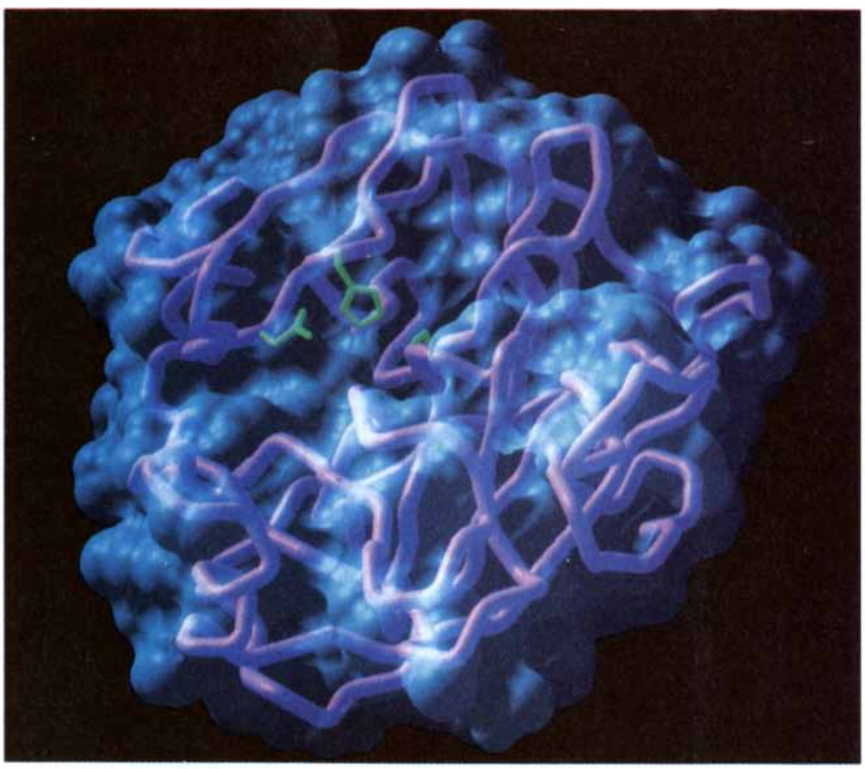
The reaction passes through a tetrahedral transition state resembling the transition state for ester hydrolysis. As before, we created a transition-state analogue by replacing the tetrahedral carbon with a phosphorus atom. We injected the analogue into experimental animals and obtained monoclonal antibodies. Although the analogue was a mixture of chiral forms, each monoclonal antibody recognized only one enantiomer. One of these antibodies had a catalytic effect. As expected, the catalyzed reaction consumed only half of



CATALYTIC ANTIBODY for a reaction (an ester hydrolysis in this case) is made by coupling a transition-state analogue (*a*) to a carrier protein and injecting the combination into an experimental animal. Antibody-secreting spleen cells are taken from the animal and fused with cells from a myeloma, a bone-marrow cancer. The hybrid antibody-secreting cells divide indefinitely, making it possible to obtain clones of cells, each hybrid clone secreting a monoclonal antibody that has a unique binding pocket. One then selects a clone making antibody specific for the analogue. The antibody may also be capable of binding to the transition state itself (*b*) and thereby catalyzing the reaction.



TEST OF STEREOSPECIFICITY for catalytic antibodies was a reaction in which a molecule containing an open chain of oxygen and carbon atoms is transformed into a cyclic compound called a lactone. The starting material, the transition state and the product all contain a chiral carbon (*color*): a carbon bound to four chemical groups that can have two mirror-image configurations (*top and middle*). The authors synthesized a transition-state analogue and made an antibody that would bind to just one form of the analogue. The antibody catalyzed the formation of only one form of lactone, which shows it was stereospecific: it distinguished between the chiral forms of the transition state.



TRYPSIN, a digestive enzyme that catalyzes the hydrolysis of proteins, speeds the reaction not only by stabilizing the transition state as an amide (carbon-nitrogen) bond is broken but also by acting chemically on the amide group, through amino acid side chains in the binding pocket. Three amino acids (*green*) form a "catalytic triad" in trypsin and several related enzymes. The side chains react with the bound protein to form transition states of lower energy than the transition state of the uncatalyzed hydrolysis reaction (*see illustration on opposite page*). The image, which shows the protein's surface and backbone, was made by Olson using Connolly's program MCS; it is based on X-ray crystallography by Robert Huber and his colleagues at the University of Munich.

the starting material, in which enantiomers of the progenitor molecule were mixed, and it yielded just one chiral form of the lactone product. The antibody's action, then, was stereospecific, presumably because its binding pocket was shaped to recognize only one form of the reaction's transition state.

Stereospecificity is likely to be a general property of catalytic antibodies, and it could suit them to a role in a number of industrial processes, including the synthesis of therapeutic drugs. Certain drug molecules contain one or more chiral centers, which give rise to several stereochemically different forms of the drug. Usually only one form of the molecule reacts properly with the drug's receptor site in the target cell. The wrong form can be useless or, if it reacts with unintended receptors in the body, even harmful.

The same considerations that enabled us to produce catalytic antibodies for reactions of simple molecules such as esters point the way to making antibodies for cleaving proteins and nucleic acids—the fundamental molecules of life and hence the principal materials of molecular biology and biotechnology. The structural bonds in proteins are amide bonds, which join a carbon in one amino acid to a nitrogen in an adjacent amino acid on the protein chain. In the transition state for amide hydrolysis the carbon at one end of the bond adopts a tetrahedral geometry much like the transition state in ester hydrolysis—a configuration readily mimicked with a phosphorus-containing analogue. An antibody elicited with such an analogue, which might also include a few of the amino acids flanking the target bond on the protein chain, might cleave a protein. Its action would be highly specific: it would hydrolyze only the bond residing within the amino acid sequence mimicked by the analogue.

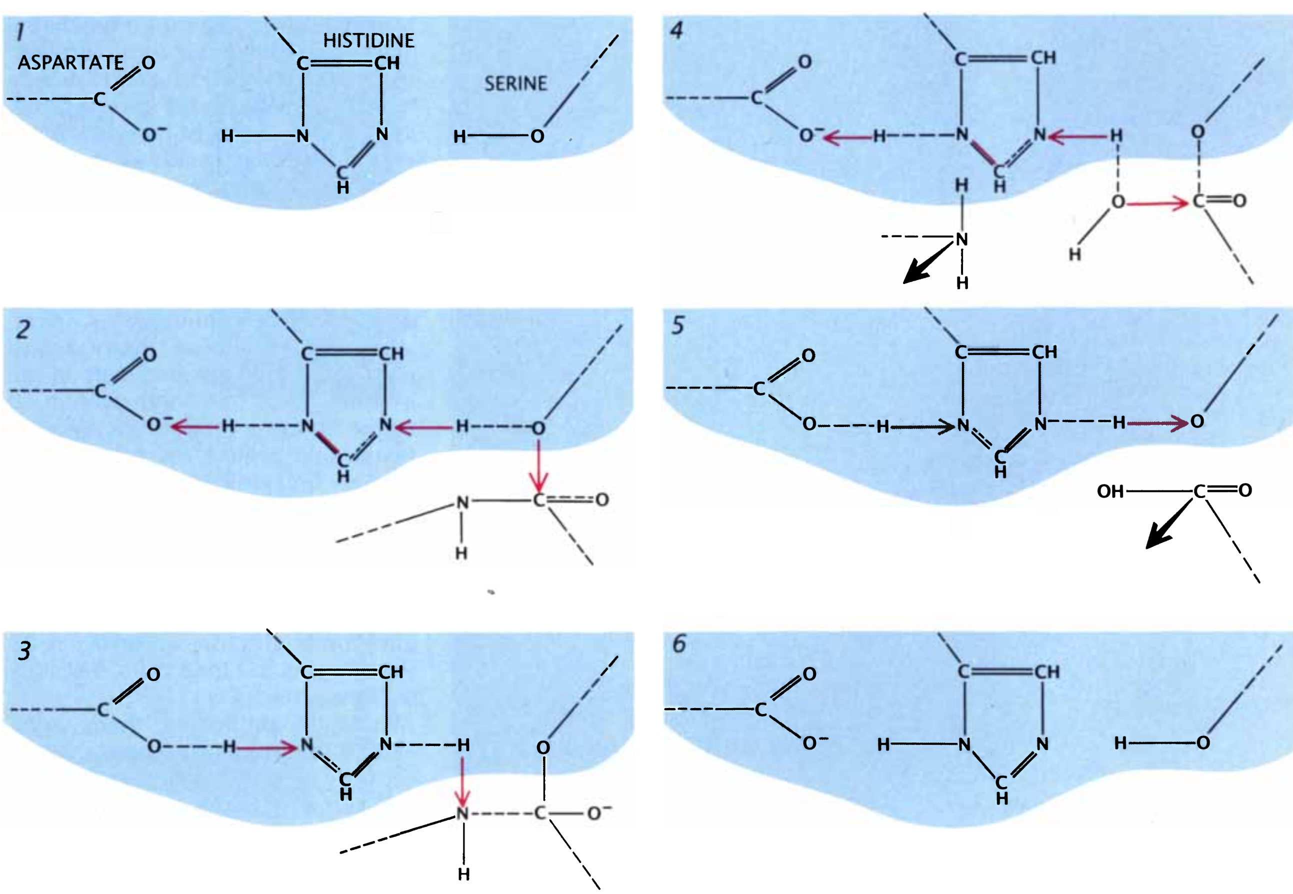
Amide bonds are exceedingly stable, however, and antibodies capable of accelerating their hydrolysis have yet to be devised. One difficulty may be that simple transition-state binding is not enough for accelerating reactions whose activation energy is very high. The binding pocket of the catalytic protein (whether it is an antibody or an enzyme) must also be able to intervene directly in the reaction, changing its chemical mechanism so that a molecule is able to cross the energy surface from the substrate well to the product well by an alternative pathway of lower energy. That is, the amino acids lining the binding pocket must take part directly in the reaction.

The action of the amino acids (or, more precisely, their side chains) is analogous to the catalytic effect of simple compounds or ions in solution. Unlike enzymes, these species are too small to enclose the substrate in a binding pocket, and yet they can act as catalysts by forming transient chemical bonds with the reacting atoms. A simple base, for example, can accelerate ester hydrolysis by removing a hydrogen ion from a water molecule. The resulting hydroxide ion reacts with the ester group much more readily than the water molecule does on its own. Alternatively, a small molecule with an affinity for carbon can substitute for the water molecule to break the carbon-oxygen bond, releasing the alcohol and forming a complex, or intermediate, with the remainder of the ester. A water molecule then displaces the catalyst and releases the acid product of the hydrolysis.

The amino acid side chains in an enzyme's binding pocket have an advantage over catalytic groups in free solution in that they need not rely on chance to bring them together with the target molecule. In many enzymes three or more groups interact simultaneously with the substrate. If those groups existed as small molecules in a solution, their juxtaposition would be quite unlikely, even disregarding the requirement that they all be oriented properly. By aligning active groups with one another and with the bound substrate, the binding pocket of the enzyme makes possible catalytic mechanisms that would be virtually ruled out otherwise.

For example, the "catalytic triad" of the amino acids aspartate, histidine and serine, precisely positioned within the enzyme binding pocket, is known to be crucial to the functioning of trypsin and other enzymes that break down proteins in the digestive tract. As the enzyme attacks an amide bond the three amino acids act in unison to dismantle the bond, step by step. First the carbon at one end of the amide bond is linked to the serine and the part of the protein ending with the nitrogen is released; then a water molecule reacts with the serine-substrate complex to release the remainder of the protein and restore the enzyme to its resting state [*see illustration below*].

Hence, by direct participation of side chains in the binding pocket, an enzyme can break down a reaction that ordinarily passes through one high-energy transition state into a series of simple steps having transition states of lower energy. It is convenient to describe this aspect of enzyme action as a process distinct from transition-state binding, but in fact the two functions are linked. The binding pocket of the enzyme may stabilize the subsidiary transition states, lowering the energetic

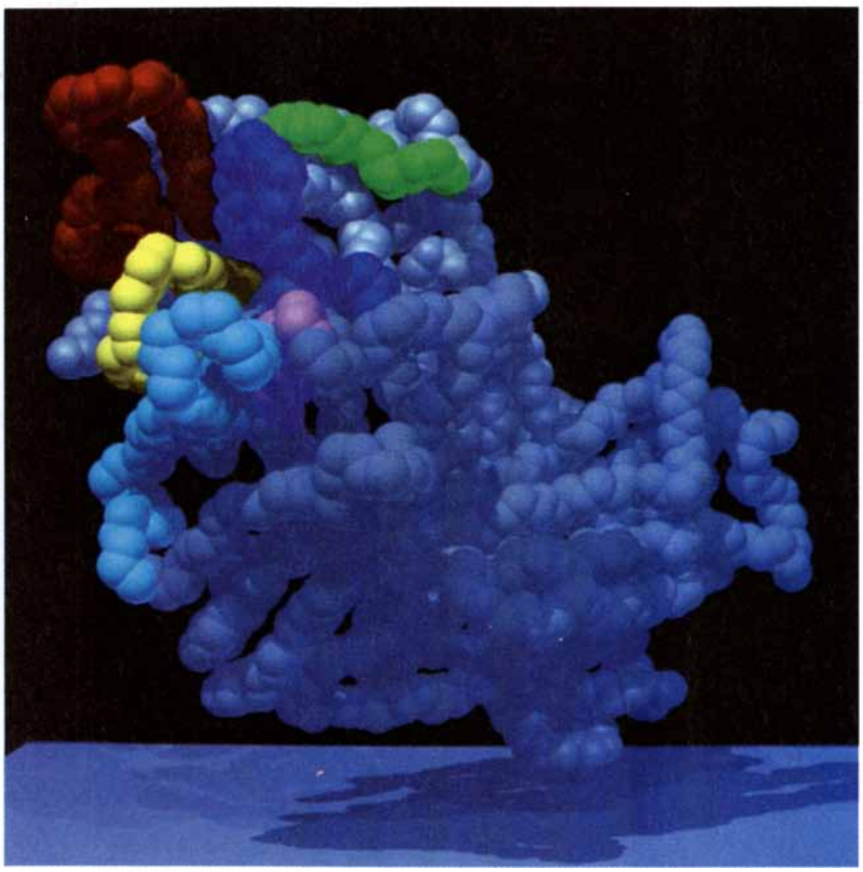


CATALYTIC TRIAD of aspartate, histidine and serine (*1*) acts in concert to break an amide bond of a protein in the active site of trypsin. The oxygen on serine is freed to attack the carbon as a proton (a hydrogen ion) is attracted from histidine to the negatively charged oxygen of aspartate, and another proton shuttles from serine to histidine (*2*). The first proton then returns and the second one attacks the nitrogen (*3*). As part of the protein is released, a water molecule takes its place, and the proton shuttle is played out again, this time forming a hydroxide group (OH) that attacks the bond between the carbon and the oxygen on serine (*4*). The bond is broken, the remainder of the protein is released (*5*) and the catalytic triad is restored to its resting condition (*6*).

cost of their formation, as well as taking part in them directly through amino acid side chains.

In antibodies too the binding pocket is lined with precisely oriented side chains, some of which could well play a role in catalysis. Can the immune system be coaxed into producing an antibody whose binding pocket not only would stabilize a reaction's transition state but also would take part in the reaction directly and alter its pathway? Antibody molecules have their greatest variability in the segments of protein making up their binding pockets; even antibodies that recognize the same antigen may have pockets containing different sets of amino acids. Through meticulous antigen design it might be possible to elicit an antibody bearing specific amino acids that could participate in a reaction. One might, for example, try to design not a static mimic of the transition state but a dynamic one: a compound that, in imitation of the transition state, would react chemically with the binding pocket of an appropriate antibody.

Indeed, the staggering variability of antibody binding sites opens the possibility of developing antibodies that would bring many different sets of catalytic side chains to bear on the same reaction. Each of hundreds or thousands of antibodies to the same transition-state analogue might catalyze the reaction by a slightly different mechanism, depending on the constellation of side chains in the binding pocket. We have already seen some evidence for a diversity of mechanisms in our catalytic antibodies for ester hydrolysis: whereas our first antibody accelerates the reaction by a factor of 1,000, other antibodies to the same transition-state analogue accelerate the reaction by a factor of as much as seven million.



"HYPERVARIABLE LOOPS" (*color*) of protein make up an antibody's binding pocket. Here one arm of a Y-shaped antibody molecule is shown; two of the antibody's polypeptide chains each contribute three hypervariable loops to form the pocket. The chemical makeup of an antibody's binding pocket might enable amino acids there to take part directly in a transition state as the antibody catalyzed a reaction. Because the composition of the pocket varies widely among the many millions of antibodies, there might be many different catalytic antibodies for a single reaction, each of them bringing a distinctive catalytic mechanism to bear on the process. The image was made by Pique.

Such a diversity of catalytic mechanisms could prove a boon for the study of protein catalysis. What specific features account for the remarkable efficiency of the enzymes found in nature? What features represent a minimum requirement for their activity? How might less efficient, evolutionarily more primitive enzymes have been structured? These questions have motivated detailed studies of existing enzymes. Study of the catalytic effect of antibodies whose binding pockets vary subtly in their makeup will provide an efficient new means of addressing such questions.

The catalytic antibodies devised so far by us and by others transform comparatively simple compounds. Much of the potential of these new protein catalysts for biotechnology and medicine depends on the development of antibodies that are able to act on proteins or nucleic acids. Existing proteases (protein-cleaving enzymes) are few in number and relatively nonspecific in their action: they cleave their target bond with little regard for its chemical surroundings. Catalytic antibodies might hydrolyze amide bonds that are resistant to the known proteases, and they might be much more sensitive to the specific amino acids flanking the target bond.

Such catalysts could be put to use in medicine—for example in vaccination. Current vaccines mimic a pathogen, such as a virus, in order to induce protective antibodies. An antiviral vaccine of the future might mimic just the transition state in the hydrolysis of one viral protein. It might induce catalytic antibodies that would protect the recipient by actively breaking down the invading virus. At the same time the antibodies would spare the host's own proteins. By the same principles one might stimulate the immune system of a patient with heart disease to produce antibodies that would break up the proteins in blood clots, forestalling heart attacks.

Catalytic antibodies, then, could extend the immune system's innate capacity to defend the body. They will certainly contribute to biotechnology and to basic research both in chemistry and in molecular biology. Those disciplines should profit directly, from an expanded molecular toolbox. They might also profit in unforeseen ways, from the opportunity to explore the full potential of protein binding pockets for fostering chemical reactions.

Special Introductory Offer

Save 20% on the most widely used biochemistry textbook in the world

BIOCHEMISTRY

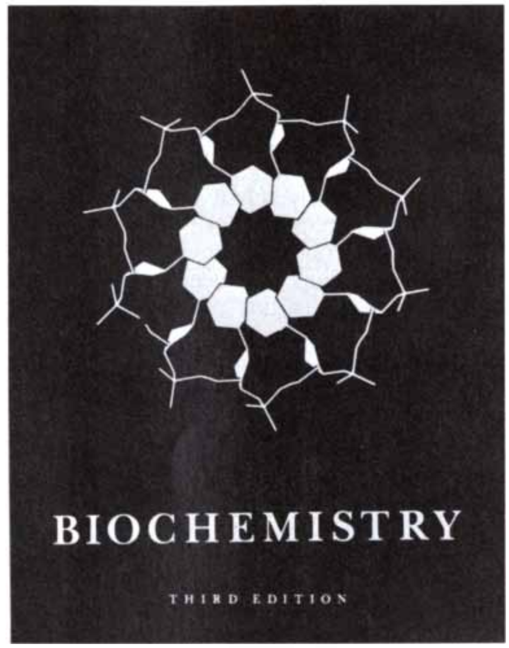
THIRD EDITION **Lubert Stryer** *Stanford University*

When Lubert Stryer's *Biochemistry* was first published in 1975, it became an instant success—earning widespread acclaim for its clarity, conciseness, accuracy, and coherence. The new Third Edition of Stryer's landmark textbook continues to set the standard by which all other biochemistry textbooks are measured.

The Third Edition is extensively updated to reflect the enormous advances made in biochemistry since the publication of the Second Edition in 1981. Stryer incorporates these advances while retaining historical perspective. The main goal, as in earlier editions, is to convey the enduring fundamentals of the discipline.

NEW

- A new Part I, *Molecular Design of Life,* introduces the flow of information from gene to protein and emphasizes the growing interrelatedness of molecular biology and biochemistry.
- The two final parts on *Genetic Information* and on *Molecular Physiology* contain a wealth of new material and reinforce the text's interdisciplinary approach.
- More than 300 new illustrations and 150 new problems have been added.



Available March 1988
1136 pages, 959 illustrations
ISBN 0-7167-1843-X, Regular Price $48.00 (tent.)
Special Introductory Price $40.00
This offer expires April 30, 1988.

"A major aid to the teaching of biochemistry world-wide."
ALEX GLAZER, University of California, Berkeley

W. H. Freeman and Company The book publishing arm of *Scientific American*

☐ YES! I would like to take advantage of this special introductory offer.

Please send me:

Quantity	ISBN	Title	Special Low Price	Amount
	1843-X	Biochemistry, Third Edition	$40.00	

TOTAL ______

Shipping and handling: Add $1.50 for first book plus $1.00 for each additional book ______

Sales tax (NY, NJ, CA, and UT) ______

FINAL TOTAL ______

Mail to:
W. H. Freeman and Company
4419 West 1980 South, Salt Lake City, UT 84104

☐ I enclose a check or money order payable to W. H. Freeman and Company

☐ MasterCard ☐ VISA Account #______

Expiration date____/____

Signature______
(Credit card orders must be signed)

Name______

Address______

City/State/Zip______

0317

Construction Cranes

These hoisting machines have diverged into various forms to meet the special needs of urban construction and are indispensable components of the building trade

by Lawrence K. Shapiro and Howard I. Shapiro

Cranes have been fundamental tools of construction since ancient times. In a modern urban setting, where construction heads upward, the outstretched booms of tall construction cranes lace the skyline, virtual icons of development and telling gauges of the economic health of a city. To the builder they are more than that, of course. They are the powerful, maneuverable machines that make building a modern skyscraper economically feasible.

Over the years the appearance and capabilities of these machines have changed dramatically but their function has remained essentially the same. Cranes employ cables and pulleys to raise and lower loads that would otherwise be too heavy to hoist aboveground; once those loads are lifted, cranes move them about horizontally and maneuver them into position accurately.

As long ago as the first century A.D. cranes were powered by men who raised and lowered ropes by shifting the weight of their bodies inside a circular treadmill. It was not until the 1800's, when the steam engine was introduced, that a significant advance was made on that power source. Today modern cranes rely on technologies such as microprocessors for load monitoring, high-strength steels for weight reduction in critical areas and advanced power transmission systems for fine-tuned control.

A crane is essentially a block-and-tackle system that is suspended from a boom and powered by its own engine. In the simplest arrangement a single rope runs from the hook to a pulley, called a sheave, at the head of the boom and down to a winding drum near the base of the boom. When the crane operator pulls on the hoist lever, the drum begins to turn; as the wire-rope hoisting cable spools onto the drum, the hook and its load rise accordingly.

There is no mechanical advantage in this simple system with a single line of rope; to lift an object, the force on the rope must slightly exceed the weight of the load. A mechanical advantage can be created, however, by adding sheaves and thus distributing the weight of the load to several lines of rope. If, for example, two sheaves are added to the hook block and three are added to the boom end of the hoisting system, the mechanical advantage is increased almost fivefold (a small percentage is lost to friction). This is accompanied by a matched reduction in the hook's speed of ascent (in this case the ascent is slowed by a factor of five). The relative advantages of each arrangement must be evaluated in advance; a job that demands rapid hoisting time will call for a crane with a simple pulley system. A job that involves the hoisting of heavy loads or the slow ascent of delicate pieces of equipment will call for a crane with multiple sheaves.

Modern Cranes

Modern cranes fall into two functional categories: those that are employed in industry and those that are employed in construction. Although the distinction between them is not absolute, the different work environments often call for cranes with different capabilities. Moreover, there is a fundamental disparity in the responsibilities imposed on the operators of the two kinds of machines. Industrial cranes often perform repetitive, well-defined functions within a known operating environment, whereas construction cranes must be capable of performing a variety of tasks under changing conditions.

A good example of an industrial crane is a port crane. It lifts containers of predetermined size and weight from one clearly defined area to another, and each lift deviates very little from the one preceding it. The geographic location, support conditions and load-handling requirements of the port crane can be determined in advance and a crane can be matched to the specifications of the job. In some cases the lifting process is so repetitive that it can be partly automated, minimizing the skills required of its operator.

Construction cranes, on the other hand, are standard production units that must operate under conditions that are anything but constant; only the machine's operating capabilities are defined in advance. For this reason heavy reliance must be placed on the skill of the people responsible for the planning and operation of these cranes. Also, site conditions and lift requirements may differ from one construction project to the next and in the course of construction at a single site as the work progresses from belowground to the roof. A tower crane (so called because it operates from the top of a towerlike structure) may operate in one environment at the start of a project, when it is sheltered from high winds, and in a different environment at the end, when at its full height it may be buffeted by winds of 100 miles per hour.

Construction cranes operating in U.S. cities today are extraordinarily diverse, a phenomenon that can be explained at least in part by their separate evolutionary histories in Europe and in North America. In Europe modern cranes were designed to function in cities where the buildings were not very tall, the streets were narrow and the population was dense. These conditions called for cranes that were quiet, relatively unobtrusive and safe. Europe's answer to those demands was the tower crane, which first appeared in large numbers during the 1950's. It is a tall, slender, latticed structure that rises above the building under construction and has the advantage of having both its operator and its working boom above the construction site. Moreover, because these cranes are equipped with electric drives, they are able to operate quietly, but with circumscribed speed, acceleration and lifting capabilities.

In the U.S., in contrast, the crane industry had its beginnings in the heavy industrial belt of the Middle West. Cranes began to proliferate there in the early part of the 20th century, manufactured by the same companies that produced mining shovels and excavating equipment. These early crane makers were not concerned about urban congestion but wanted a versatile machine that could be applied to a multiplicity of jobs, ranging from construction work to river dredging to soil excavation. For that reason they designed cranes that were mobile, rugged and powerful and were driven by internal-combustion engines. Such cranes were operated largely on the basis of operator discretion. A good operator could handle the machine and load with finesse, but in careless hands the great power of the crane could lead to its self-destruction.

In the 1950's high-strength steels were introduced at about the time that economic factors created a demand for taller residential buildings. These developments led to the appearance of the long-boom mobile crane mounted on either a trucklike chassis or a crawler base with caterpillar treads. From that point on, the mobile crane became an urban fixture, joined several years later by the tower crane. At present, construction cranes of all types are manufactured and distributed throughout the world. Mobile and tower cranes have both become commonplace in American cities.

Safety and economy dictate that the right crane be selected for a construction job. A building six stories tall obviously has markedly different requirements than one 60 stories in height. Other factors to be considered are the terrain of the site, the proximity of nearby buildings, pedestrian traffic and the unique characteristics of the structure being erected.

Mobile cranes (in contrast to tower cranes) have the advantage of being able to move about a job site under

MOBILE CRAWLER CRANE with tower boom is lowering its hook to the ground, where it will receive a bucket of concrete for the concrete-framed building beside it. The building, photographed when 12 stories tall, will stand 45 stories when it is completed.

their own power, which gives them valuable flexibility. They can be separated into two basic groups: those that have rubber tires and those that have caterpillar treads.

Rubber-tired cranes are relatively light in weight and are fitted with a truck chassis, and so they can drive from site to site. Smaller models can be driven fully assembled on public highways; larger ones can also be driven on the highway but must be partially disassembled so that their total weight is within legal limits. Although those limits vary from state to state and country to country (a situation that causes headaches for both designers and transporters of cranes), rubber-tired cranes with lift capacities as great as 300 tons are common and do travel legally in most jurisdictions. When rubber-tired cranes arrive at a site, they must be positioned and stabilized before they are operational. The outriggers need to be extended and the crane jacked up on cribbing, a timber or steel mat that distributes its weight.

Crawler cranes are more cumbersome to transport to a construction site, but they do not require outriggers and, unlike truck cranes, can travel about the site while carrying a load. For travel between sites, however, they must be loaded onto low-bed trailers or railroad flatcars, and they are usually disassembled to meet weight and width limitations. Caterpillar treads do provide an advantage: they spread the weight of the crane over a large area of ground. That explains why a crawler crane can move with a load on its hook.

HUMAN-POWERED CRANES, such as the one depicted here, were common lifting devices at waterfront docks in the 16th century. By walking inside paired treadmills called squirrel cages, one on each side of the crane's housing, workers were able to turn a winding drum, which spooled in the rope and raised the weight-bearing hook. The cranes were mounted on a center pivot and rollers and could rotate a full 360 degrees.

Variations in Boom Design

Truck cranes are available with one of two kinds of booms, the selection of which depends on the needs of a particular job. Telescopic booms, as their name implies, consist of multiple sections that nest inside one another and can be extended or retracted by the operator. Small rubber-tired cranes equipped with telescopic booms are commonly called cherry pickers. Highly mobile and available with lift capacities ranging from 15 to 80 tons, they are well suited to serve as general-purpose, quick-service machines. For that reason they complement the big cranes at virtually every major construction site. Larger telescopic models have lift capacities as great as 500 tons and booms that extend as far as 200 feet. Extension struts can be added to the boom to provide still greater height.

One drawback to telescopic cranes is the weight of the boom; the cantilevered boom (formed by overlapping boxlike sections) offers less than ideal structural efficiency so that large, heavy booms are necessary for strength. The extra weight makes the lifting capacity of these cranes low at long reaches. Another drawback to telescopic cranes is that per pound of lift capacity they are the most expensive of all construction cranes. Offsetting their initial cost, however, is the rapidity with which they can be put to work on arrival at a site. Because of that feature, in fact, some fleet operators offer them as a taxi service and will make lifts for several customers on the same day.

The other type of boom utilized on mobile cranes (both rubber-tired and crawler models) is the latticed boom. Latticed booms are formed by an assemblage of high-strength steel tubing or angle irons arranged as a three-dimensional truss. The boom, which consists of multiple sections pinned or bolted together, can be assembled to lengths as great as 350 feet and then increased by as much as 100 feet by adding a jib, a lightweight extension attached to the tip of the boom. A boom of that size has been used to construct buildings as tall as 46 stories.

Although it is possible to build latticed booms of greater length, it is generally not practical to do so. The strength of the boom is related to its length-to-width ratio. Since booms with cross sections greater than eight feet require special permits to be transported by truck, this dimension is a practical constraint on size.

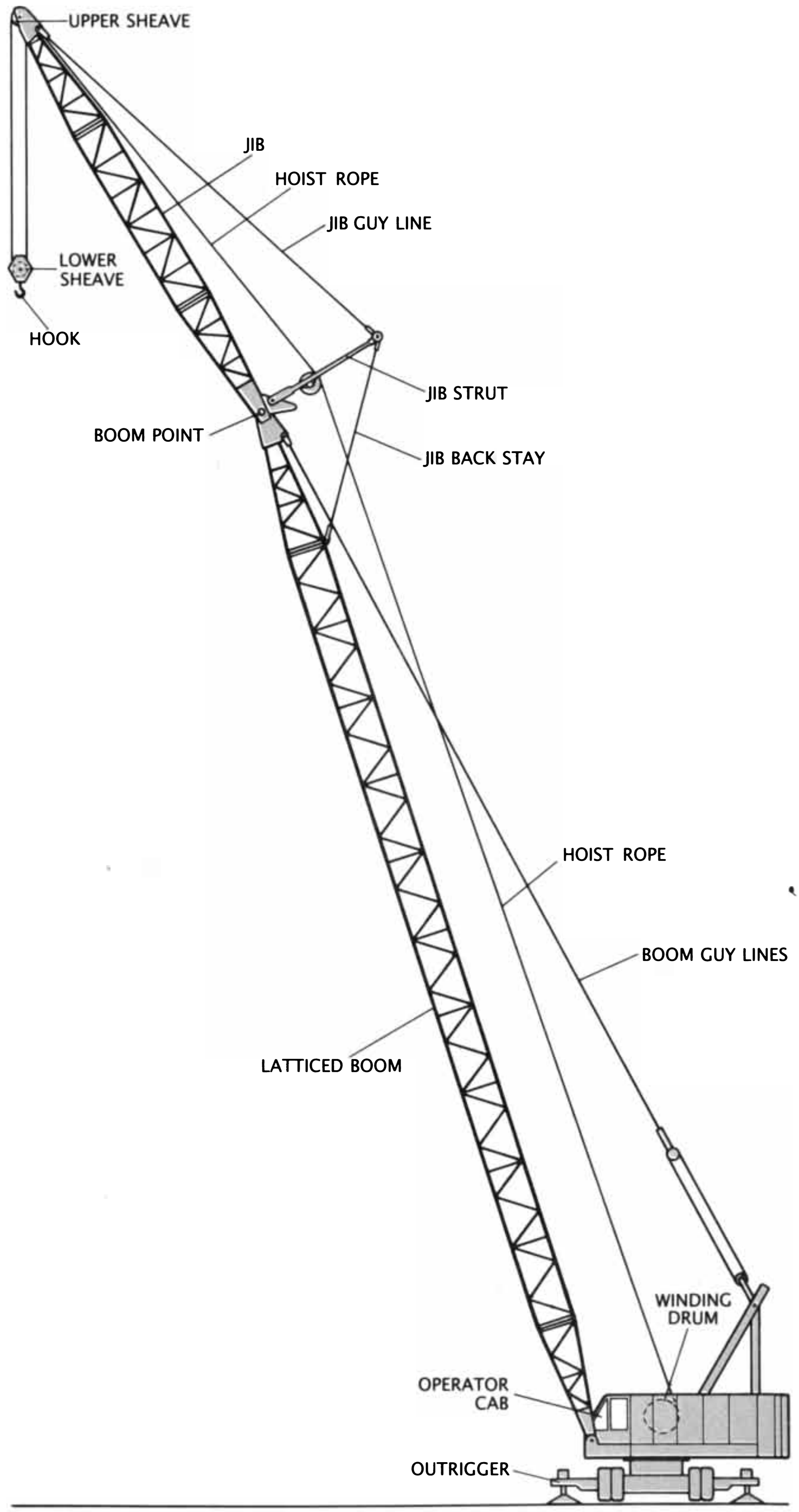
Latticed booms, unlike telescopic booms, require varying degrees of assembly once they arrive at the job site. Those that are short can be folded in half, driven to their destination and put into service within two hours of arrival. Longer booms must be disassembled and trucked in separate pieces to a site, where it may take up to five hours to reassemble them.

A recent variation on the latticed boom that has become very popular for urban use is a hybrid between the mobile crane and the tower crane; it is called a tower attachment. Mounted on crawler treads, it consists of a vertical tower section equipped with a latticed boom at its top. The tower runs parallel to the building face and places the working boom high in the air. This hybrid crane has most of the advantages of both crane types but its maximum height is limited to approximately 200 feet and its base occupies a large ground area. The average time needed to assemble this type of crane is about 10 hours, which is long for a mobile crane but is much shorter than for a tower crane. The expense of shipping and assembling these cranes, however, makes them generally uneconomical for assignments of less than a few weeks in duration.

Older latticed-boom mobile cranes are driven by internal-combustion engines that have geared mechanical drives. As in a car with manual transmission, power is provided to the winding drum by engaging a clutch, known in the trade as a friction. Hoisting speed is governed by the speed and torque characteristics of the engine and can be further controlled by slipping the friction.

By disengaging the friction and applying the foot brake the operator can suspend the load at any height; to hold the load for an extended period of time a pawl, or latch, is set into rachet teeth on the drum, preventing the drum from turning. When the load is to be lowered, the operator releases the pawl and controls the descent by slipping the friction, the brake or both. The hook falls freely if both friction and brake are released, and these free-fall winches can shave a significant amount of time off the return part of a lift when no load is on the hook.

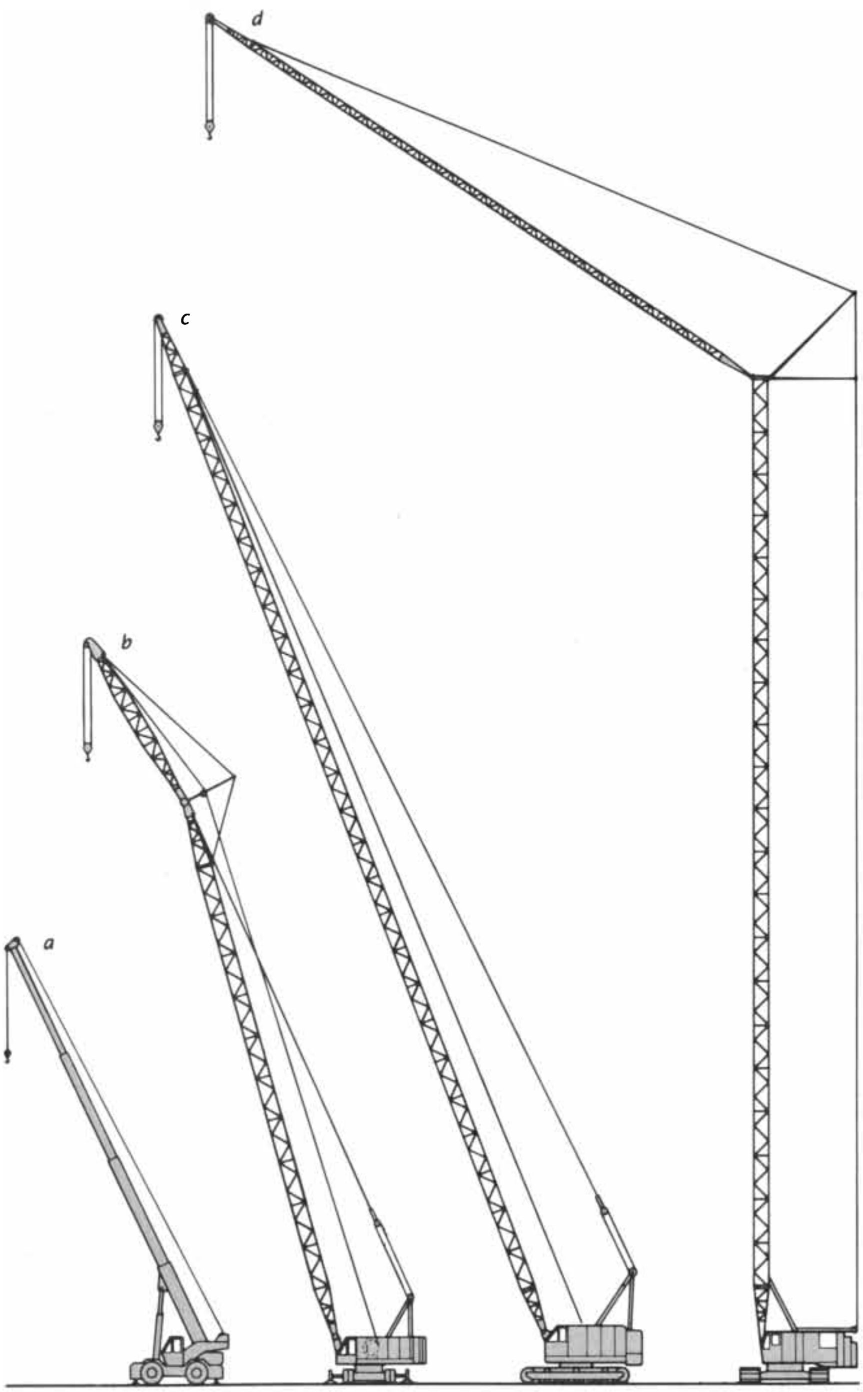
Some of the more modern mobile cranes have a hydrodynamic device called a torque converter instead of a



LATTICED-BOOM mobile cranes are found at many urban construction sites. By actuating various levers the operator can raise or lower the hook block, raise or lower the boom and jib and rotate the upper part of the machine on its base. Rubber-tired cranes such as this one are equipped with outriggers that provide stability during lifting.

mechanical transmission. The torque converter maintains constant power demand on the engine by adjusting hoisting speed to accommodate loads that differ in weight. For a given load, the operator can vary the power and hoisting speed by controlling engine speed. In order to raise a load off the ground, engine speed must be increased until the power output is sufficient to overcome the resistance of the load. Once a load has been lifted from the ground, it can be brought to a stop and held there by reducing the speed (and power) of the engine to match the resistance of the load. By slowing the engine still further, so that the load prevails, the hook can be lowered at a controlled rate.

The same thing happens in a car with automatic transmission: the car can be held in place on a hill if just the right amount of pressure is applied to the accelerator pedal; too little pressure and the car will roll backward, too much and the car will creep forward. A crane operator, like a car driver, can (and usually does) use the brake to achieve the same effect, namely to hold the load. The torque converter makes it possible to raise and lower loads more gently and with more control than was possible with the old mechanical systems.



MOBILE CRANES vary greatly in the length of their booms. In this drawing four representatives are drawn to the same scale. From left to right they are a rubber-tired cherry-picker crane with telescopic boom (*a*), a rubber-tired crane with latticed boom and jib (*b*), a crawler crane with latticed boom (*c*) and a crawler crane with tower attachment (*d*).

Tower Cranes

Tower cranes rise with the building they are helping to construct, giving their operators a bird's-eye view of the world below them. Like mobile cranes, they rotate and have a lifting hook to raise loads and move them radially, but in other respects they bear little resemblance to their mobile relatives.

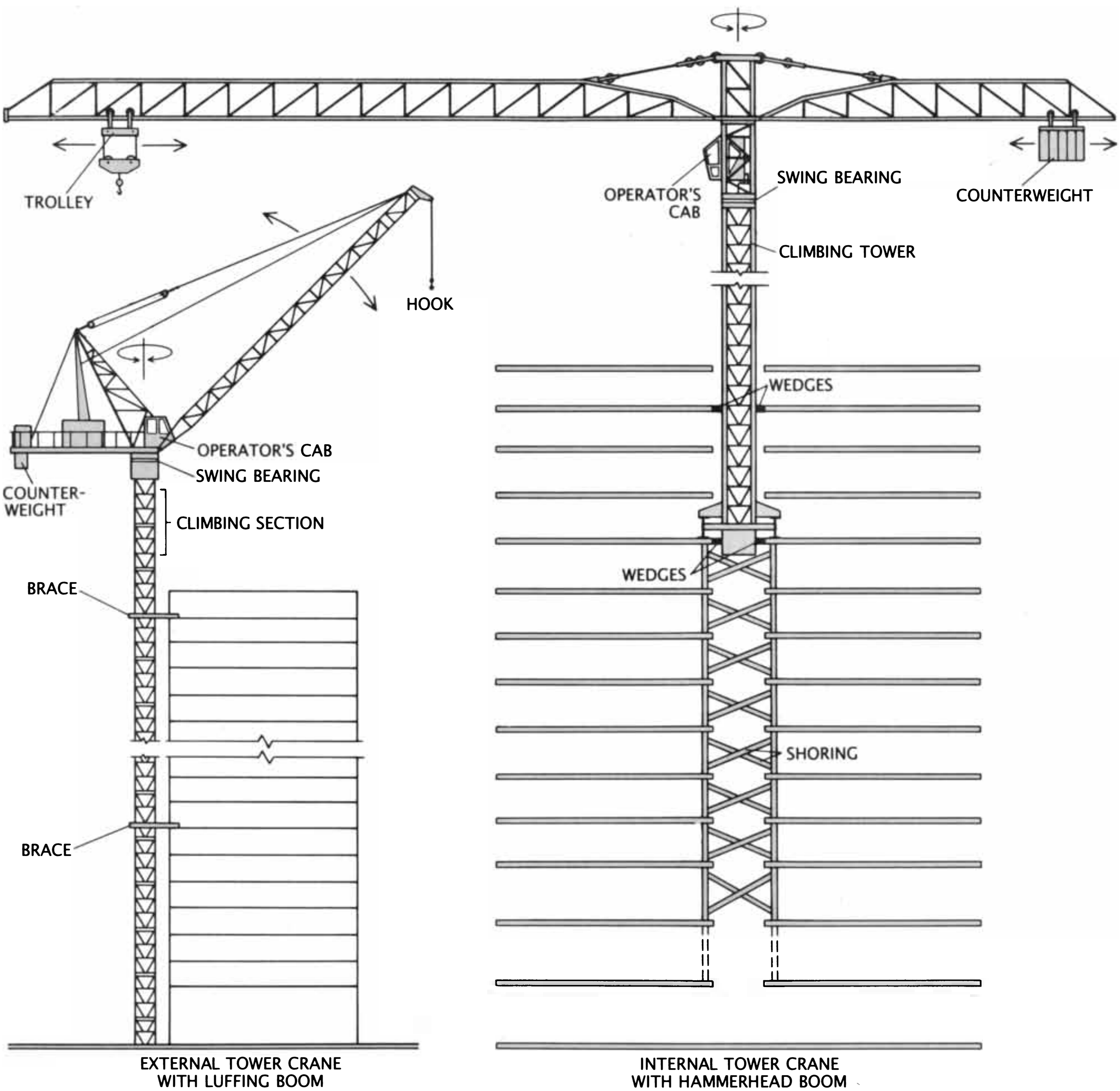
Tower cranes can be raised to almost limitless heights whether they are on the inside or the outside of a building. An internal tower crane has a fixed configuration: its support tower remains the same length throughout a construction job. It raises itself on hydraulic jacks that move it from floor to floor as the building grows in height [*see illustration on pages 78 and 79*]. Both the jacks and the building must support the weight of the crane, which typically is well over 50 tons.

Temporary openings slightly larger than the tower must be provided on each floor to enable the crane to slip through as it climbs upward. Overturning and twisting forces from the crane are passed to the building through upper and lower sets of wedges. Nevertheless, maximum rotational acceleration and the speed of the boom's swing are kept low by automatic controls in order to limit the inertia of the rotating mass and its twisting effect on the tower. Internal cranes have a drawback: they end up on the roof when the job is done and must then be removed. The crane components are disassembled piece by piece and lowered to trucks below, a difficult operation that may require several weeks to complete.

An external crane, on the other hand, is self-supporting and grows alongside the building. The base of the tower is anchored to a concrete foundation placed in close proximity to the building and initially the crane is freestanding. When the building reaches a height of about eight stories, the tower is braced to the frame of the building and begins to climb. That is accomplished by jacking up the upper part of the crane and adding a new tower section immediately below it. Climbing continues as the building rises, and braces are added at set intervals for lateral support.

Although the external tower crane is easier to dismantle than the internal crane (the climbing process can be reversed until the crane is low enough to be disassembled by a mobile crane), it is not suitable for every building. "Wedding cake" buildings, for example, may have tiered setbacks so distant from the tower of the crane that bracing is economically infeasible. In either case, the expense of setting up and removing tower cranes is often so great that they are usually found at jobs lasting for six months or longer.

Tower cranes vary somewhat according to their country of origin. Those made in Europe are powered by electric motors with stepped-speed and acceleration controls and sophisticated braking mechanisms; North American and Australian crane makers, catering to local preferences, produce mostly hydraulically driven machines powered by diesel engines. In all cases, however, tower cranes are equipped with numerous safety devices. These include fail-safe brakes that apply automatically in the event of a drive-system malfunction and speed and overload limiters to guard against careless operation. A tower crane with properly adjusted limiters, for example, will not carry loads that are too heavy to be



TOWER CRANES can be external (*left*) or internal (*right*). External cranes increase in length as the building rises; internal cranes do not increase in tower length but instead climb upward as the building grows. In both cases the cab rises above the skyscraper being constructed, giving its operator a bird's-eye view of construction from start to finish. Tower cranes can be designed with either a luffing boom (*left*) or a hammerhead boom (*right*). Hammerheads, although common, cannot be operated where obstacles, such as adjacent buildings, impede the full sweep of the boom. These two cranes are not drawn to the same scale.

safely handled. In addition the operator's levers are deadman switches: if released, they return to neutral and the brakes engage automatically.

An observer of an urban skyline will note that there are two types of tower cranes: those that have booms projecting at an angle above the horizontal and those whose booms lie flat. The first type is called a luffing boom, the second a hammerhead boom. A luffing boom projects at a vertical angle that can be altered to produce radial movement of the hook. A hammerhead is equipped with a trolley on the underside of the boom to carry the hook back and forth. Because hammerhead booms cannot be raised, however, they can be utilized only where there are no obstacles, such as adjacent buildings, to interfere with the boom's sweep. For that reason luffing cranes are commoner at congested sites or at sites where several cranes are operating simultaneously. Luffing booms greatly outnumber hammerheads in New York City, for instance.

Selecting the Right Crane

In selecting a crane, the type of building under construction must be taken into consideration. Today's skyscrapers belong to two framing categories: those with steel frames and those with concrete frames.

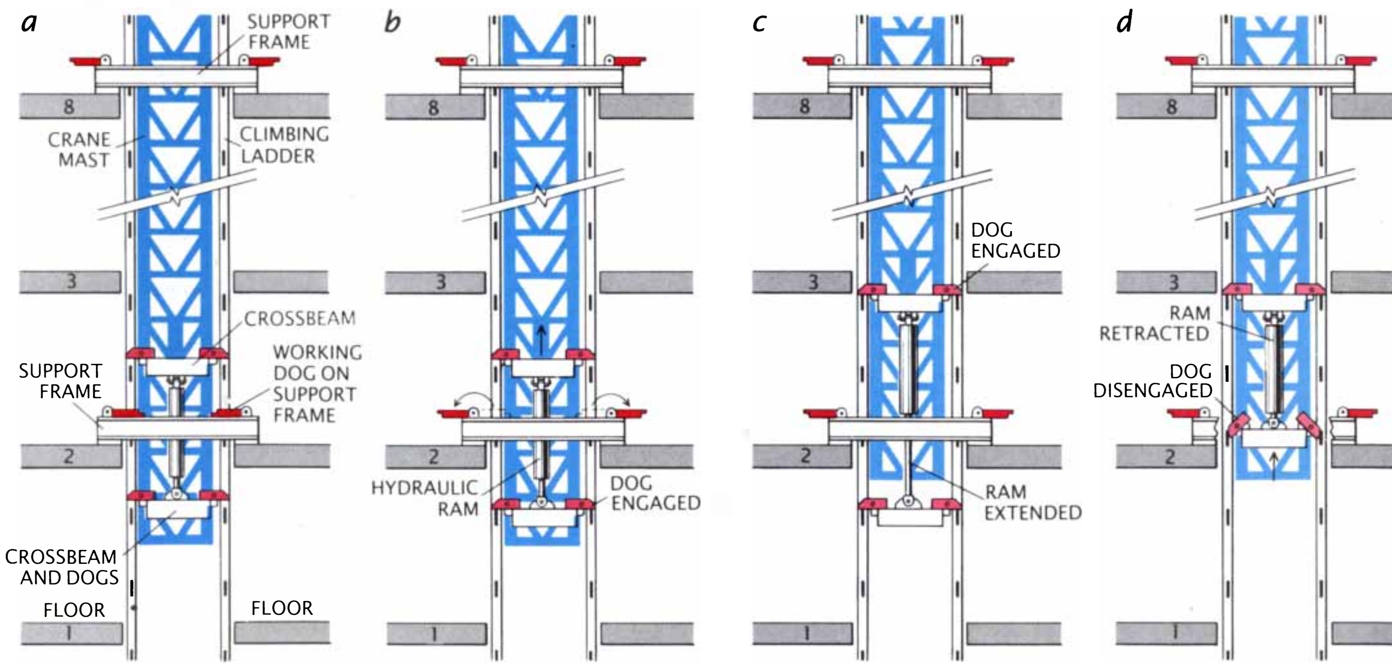
Steel-framed buildings are formed from pieces that vary tremendously in weight. Twenty-five-foot column sections, at the base of a 50-story building, for example, may weigh 25 tons each, whereas those at the top may weigh only a ton. Moreover, the demands on urban space today are so great that structural engineers must sometimes create immense spanning members to carry the weight of the building over an atrium, a meeting hall or even an adjacent building. These elements can weigh more than 100 tons each and may require two cranes working in synchrony to hoist them into place. In addition all the components of a steel-framed building must be sorted and tracked during the process of construction. This requires that they be "shaken out" when they arrive so that individually marked pieces can be identified. In urban areas where space is limited the shaking-out process is performed at the top of the building being constructed.

Crane selection at a steel-framed site is therefore determined largely on the basis of such parameters as typical and maximum weights to be hoisted, the accessibility of delivery trucks to the building, the location of the shaking-out site and the amount of space available for the crane itself.

Concrete-framed buildings impose a different set of constraints. Precast concrete, common in low and mid-rise buildings, is manufactured elsewhere and trucked to the site, and it is similar to steel in the way it is erected. For concrete that is poured in place (that is, at the construction site) a critical parameter is the speed with which concrete can be hoisted to the top of the building. A typical high-rise tower may require 250 cubic yards of concrete per floor, the equivalent of one million pounds.

Whereas the assembly of one story of a steel-framed building may require 25 lifts of steel, an equivalent story in a poured-in-place concrete structure may require as many as 150 lifts. If as few as 30 seconds can be eliminated from the hoisting time (60 seconds for a return trip), the amount of time saved will have a minor effect on steel erection but could save two and a half hours for each floor of a concrete building. When, as is typical in New York City, a poured-in-place concrete story must be completed in two days, the contractor's profitability depends on the speed of the hook.

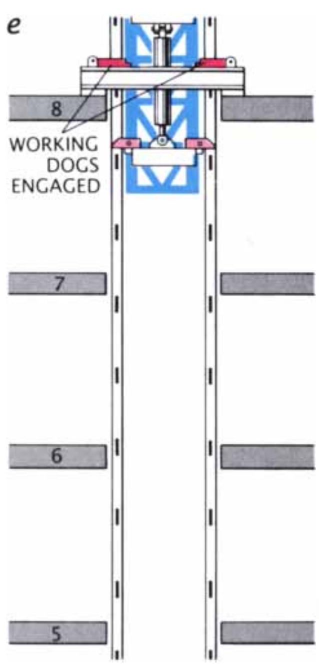
It is not unusual to see more than one type of crane at a construction site. At the earliest stages of a large



INTERNAL CLIMBING CRANES lift themselves by hydraulic climbing systems. This highly schematic drawing shows a pair of climbing ladders next to a section of the crane mast. The crane climbs upward along the rungs of the ladders, which are suspended from a support frame. It does so by alternately extending and retracting a hydraulic ram that has a crossbeam fitted with a pair of dogs at each end. When the crane is operating (*a*), the ram and both sets of dogs are retracted and the weight of the crane is carried by working dogs near its base. At the start of the lifting process the ram is partially extended by hydraulic pressure (*b*); the lower dogs are set into the rungs of the ladders, thereby transferring the weight of the crane to the ladders, and the working dogs are retracted. In the next step (*c*) the ram is fully extended and the upper dogs are engaged. The ram can

job a cherry picker may begin the job and then be replaced by a larger mobile crane, which can be positioned to carry out the heavy, complex framing of the building's base. When it is time to construct the lighter-weight, repetitively framed upper floors, a tower crane may be installed. Additional cranes may be brought to the site for finishing the building's exterior as scheduling (or other factors) usually require that the process be expedited. Materials that are very light in weight or compact (bricks, mortar, drywall panels, windows and so on) are more economically lifted in a temporary elevator erected on the outside of the building, thereby freeing the cranes for heavier work.

Many forces act simultaneously on a crane, threatening its stability and inducing stress. These include the weight of the load, the pressure of the wind, the weight of the crane itself and the inertia associated with the moving crane components and its load. The resulting effects must of course be constrained by the structural strength of the crane and the earth or building that supports it, but a crane must also have adequate stability to resist overturning. Indeed, mobile cranes are limited more by stability than by strength. The maximum permitted load is set at from 75 to 85 percent of the lifted load that would cause the mobile crane to tip over. Crane designers can maximize the lifting capacity of a machine by minimizing boom weight (for example by utilizing high-strength steel in a latticed boom to reduce self-weight in relation to lift capacity), maximizing the counterweight and providing the widest base possible.



then be retracted (*d*), raising the lower crossbeam and automatically disengaging the lower dogs. After the cycle has been repeated several times (five or six floors are climbed at a time in this way) (*e*) the working dogs are once again set in place and the crane becomes operational.

Counterweights on large mobile cranes can weigh as much as 75 tons. They are usually fabricated in sections that are removable and can be shipped separately from the crane itself. When counterweights are combined with the weight of the crane, engine and drive machinery, they add stability to the crane. (If too much counterweight is added, however, an unloaded crane can topple over backward.)

Outriggers that project from the base of a rubber-tired crane provide increased stability by extending the tipping fulcrum away from the body of the machine. Most crane outriggers are sturdy beams that telescope out from each end of the machine chassis. The outer ends of the outriggers are equipped with vertical jacks; when the outriggers are fully extended and the entire crane is jacked up, the crane is operational.

The operator of a mobile crane must exercise judgment, taking into account such factors as the length of the boom and the characteristics of the load, to allow for the effects of wind and inertia. The lifting of a large curtain-wall segment, for example, presents a substantial wind-catching surface and can therefore be carried out only when there is minimal wind velocity. When a crane swings, centrifugal forces project the load outward and stability is decreased; acceleration and braking induce lateral loading on the boom, causing sideways deflection and a corresponding increase in stress.

Tower cranes have automatic limits imposed on their rotational speed and acceleration to control the effects of inertia; mobile-crane operators, on the other hand, must exert that control themselves. But tower cranes are more sensitive to wind because of their fixed position. When major storms arise, the booms on mobile cranes can be lowered out of harm's way, but the booms of tower cranes remain in place and bear the brunt of the storm. As a result, tower cranes must be designed to withstand hurricane conditions. When storms do strike, the usual procedure is to let the boom swing freely, thus providing the least resistance and the smallest surface area to the wind. On days when the wind exceeds 30 miles per hour, all cranes generally cease operation, although most tower cranes are designed to withstand higher winds while working.

Crane Safety

Crane accidents at urban sites can be dramatic, and they are often followed by outcries that the public is inadequately protected. A crane that loses control of its load or topples over can wreak havoc on the streets below; pedestrians are sometimes killed or maimed. The hazards for construction workers are even greater: construction is a dangerous occupation and crane work is among its greater perils. Can anything be done to ameliorate the situation?

Most serious crane accidents are caused by overload, equipment misuse, excessive wear or damage to the hoisting ropes and failure to follow correct procedures for erection and dismantling (particularly in the case of tower cranes). A smaller but significant number of mishaps result from support failure, inadequate maintenance and collision between the boom and another object. Most of these accidents could be prevented through training programs for construction workers, crane operators and supervisory personnel.

Underlying many crane accidents, however, are errors committed in the planning stage. Safety is compromised when a crane is positioned incorrectly or is inadequate for the job. Crane operators have been known to make heroic efforts as they try to compensate for these inadequacies. Their efforts should be applauded, but the planning shortcomings that led to the selection of the wrong crane in the first place must be condemned. Because congested urban areas are associated with such high potential risk to the public, engineers must play an active role in evaluating sites, planning the lifting operations and selecting the right crane for the job.

Risks can never be totally eliminated from a crane operation (or from any operation in which human judgment is exercised), but they can certainly be minimized. To do so requires diligence on the part of the local authorities who enforce safety regulations and skill on the part of the people responsible for the deployment and operation of cranes.

How the Leopard Gets Its Spots

A single pattern-formation mechanism could underlie the wide variety of animal coat markings found in nature. Results from the mathematical model open lines of inquiry for the biologist

by James D. Murray

Mammals exhibit a remarkable variety of coat patterns; the variety has elicited a comparable variety of explanations—many of them at the level of cogency that prevails in Rudyard Kipling's delightful "How the Leopard Got Its Spots." Although genes control the processes involved in coat pattern formation, the actual mechanisms that create the patterns are still not known. It would be attractive from the viewpoint of both evolutionary and developmental biology if a single mechanism were found to produce the enormous assortment of coat patterns found in nature.

I should like to suggest that a single pattern-formation mechanism could in fact be responsible for most if not all of the observed coat markings. In this article I shall briefly describe a simple mathematical model for how these patterns may be generated in the course of embryonic development. An important feature of the model is that the patterns it generates bear a striking resemblance to the patterns found on a wide variety of animals such as the leopard, the cheetah, the jaguar, the zebra and the giraffe. The simple model is also consistent with the observation that although the distribution of spots on members of the cat family and of stripes on zebras varies widely and is unique to an individual, each kind of distribution adheres to a general theme. Moreover, the model also predicts that the patterns can take only certain forms, which in turn implies the existence of developmental constraints and begins to suggest how coat patterns may have evolved.

It is not clear as to precisely what happens during embryonic development to cause the patterns. There are now several possible mechanisms that are capable of generating such patterns. The appeal of the simple model comes from its mathematical richness and its astonishing ability to create patterns that correspond to what is seen. I hope the model will stimulate experimenters to pose relevant questions that ultimately will help to unravel the nature of the biological mechanism itself.

Some facts, of course, are known about coat patterns. Physically, spots correspond to regions of differently colored hair. Hair color is determined by specialized pigment cells called melanocytes, which are found in the basal, or innermost, layer of the epidermis. The melanocytes generate a pigment called melanin that then passes into the hair. In mammals there are essentially only two kinds of melanin: eumelanin, from the Greek words *eu* (good) and *melas* (black), which results in black or brown hairs, and phaeomelanin, from *phaeos* (dusty), which makes hairs yellow or reddish orange.

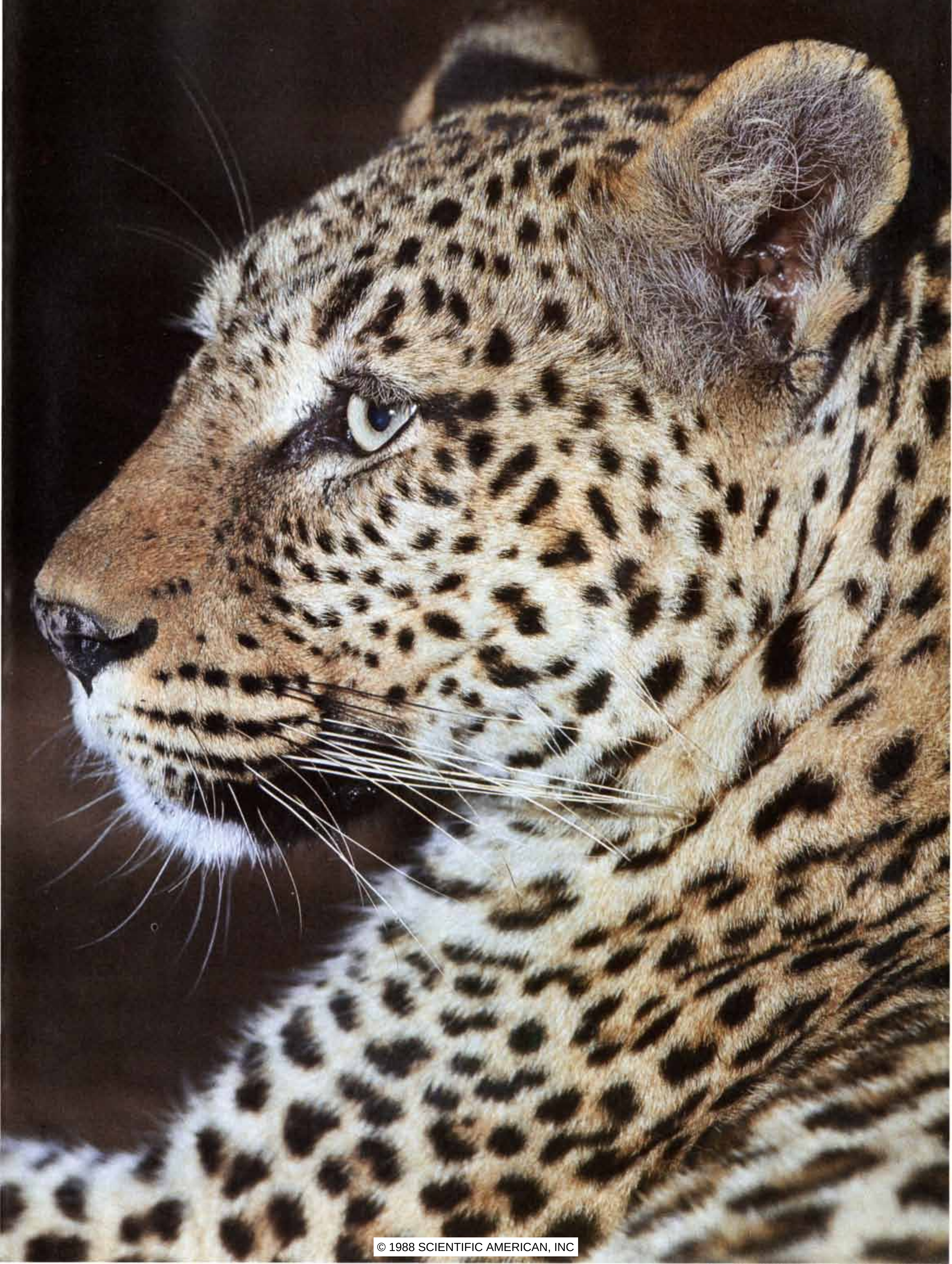
It is believed that whether or not melanocytes produce melanin depends on the presence or absence of chemical activators and inhibitors. Although it is not yet known what those chemicals are, each observed coat pattern is thought to reflect an underlying chemical prepattern. The prepattern, if it exists, should reside somewhere in or just under the epidermis. The melanocytes are thought to have the role of "reading out" the pattern. The model I shall describe could generate such a prepattern.

My work is based on a model developed by Alan M. Turing (the inventor of the Turing machine and the founder of modern computing science). In 1952, in one of the most important papers in theoretical biology, Turing postulated a chemical mechanism for generating coat patterns. He suggested that biological form follows a prepattern in the concentration of chemicals he called morphogens. The existence of morphogens is still largely speculative, except for circumstantial evidence, but Turing's model remains attractive because it appears to explain a large number of experimental results with one or two simple ideas.

Turing began with the assumption that morphogens can react with one another and diffuse through cells. He then employed a mathematical model to show that if morphogens react and diffuse in an appropriate way, spatial patterns of morphogen concentrations can arise from an initial uniform distribution in an assemblage of cells. Turing's model has spawned an entire class of models that are now referred to as reaction-diffusion models. These models are applicable if the scale of the pattern is large compared with the diameter of an individual cell. The models are applicable to the leopard's coat, for instance, because the number of cells in a leopard spot at the time the pattern is laid down is probably on the order of 100.

Turing's initial work has been developed by a number of investigators, including me, into a more complete mathematical theory. In a typical reaction-diffusion model one starts with two morphogens that can react with each other and diffuse at varying rates. In the absence of diffusion—in a well-stirred reaction, for example—the two morphogens would react and reach a steady uniform state. If the morphogens are now allowed to diffuse at equal rates, any spatial variation from that steady state will be smoothed out. If, however, the diffusion rates are not equal,

LEOPARD reposes. Do mathematical as well as genetic rules produce its spots?



diffusion can be destabilizing: the reaction rates at any given point may not be able to adjust quickly enough to reach equilibrium. If the conditions are right, a small spatial disturbance can become unstable and a pattern begins to grow. Such an instability is said to be diffusion driven.

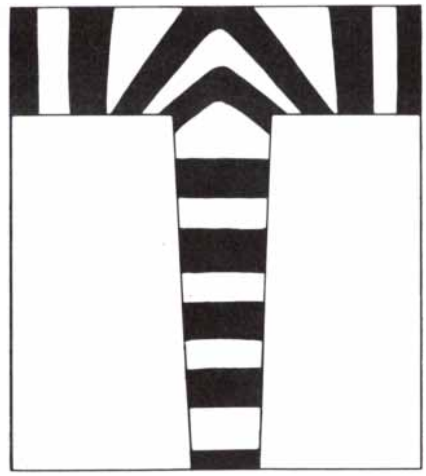
In reaction-diffusion models it is assumed that one of the morphogens is an activator that causes the melanocytes to produce one kind of melanin, say black, and the other is an inhibitor that results in the pigment cells' producing no melanin. Suppose the reactions are such that the activator increases its concentration locally and simultaneously generates the inhibitor. If the inhibitor diffuses faster than the activator, an island of high activator concentration will be created within a region of high inhibitor concentration.

One can gain an intuitive notion of how such an activator-inhibitor mechanism can give rise to spatial patterns of morphogen concentrations from the following, albeit somewhat unrealistic, example. The analogy involves a very dry forest—a situation ripe for forest fires. In an attempt to minimize potential damage, a number of fire fighters with helicopters and fire-fighting equipment have been dispersed throughout the forest. Now imagine that a fire (the activator) breaks out. A fire front starts to propagate outward. Initially there are not enough fire fighters (the inhibitors) in the vicinity of the fire to put it out. Flying in their helicopters, however, the fire fighters can outrun the fire front and spray fire-resistant chemicals on trees; when the fire reaches the sprayed trees, it is extinguished. The front is stopped.

If fires break out spontaneously in random parts of the forest, over the course of time several fire fronts (activation waves) will propagate outward. Each front in turn causes the fire fighters in their helicopters (inhibition waves) to travel out faster and quench the front at some distance ahead of the fire. The final result of this scenario is a forest with blackened patches of burned trees interspersed with patches of green, unburned trees. In effect, the outcome mimics the outcome of reaction-diffusion mechanisms that are diffusion driven. The type of pattern that results depends on the various parameters of the model and can be obtained from mathematical analysis.

Many specific reaction-diffusion models have been proposed, based on plausible or real biochemical reactions, and their pattern-formation properties have been examined. These mechanisms involve several parameters, including the rates at which the reactions proceed, the rates at which the chemicals diffuse and—of crucial importance—the geometry and scale of the tissue. A fascinating property of reaction-diffu-

MATHEMATICAL MODEL called a reaction-diffusion mechanism generates patterns that bear a striking resemblance to those found on certain animals. Here the patterns on the tail of the leopard (*left*), the jaguar and the cheetah (*middle*) and the genet (*right*) are shown, along with the patterns from the model for tapering cylinders of varying width (*right side of each panel*).



ZEBRA STRIPES at the junction of the foreleg and body (*left*) can be produced by a reaction-diffusion mechanism (*above*).

sion models concerns the outcome of beginning with a uniform steady state and holding all the parameters fixed except one, which is varied. To be specific, suppose the scale of the tissue is increased. Then eventually a critical point called a bifurcation value is reached at which the uniform steady state of the morphogens becomes unstable and spatial patterns begin to grow.

The most visually dramatic example of reaction-diffusion pattern formation is the colorful class of chemical reactions discovered by the Soviet investigators B. P. Belousov and A. M. Zhabotinsky in the late 1950's [see "Rotating Chemical Reactions," by Arthur T. Winfree; SCIENTIFIC AMERICAN, June, 1974]. The reactions visibly organize themselves in space and time, for example as spiral waves. Such reactions can oscillate with clocklike precision, changing from, say, blue to orange and back to blue again twice a minute.

Another example of reaction-diffusion patterns in nature was discovered and studied by the French chemist Daniel Thomas in 1975. The patterns are produced during reactions between uric acid and oxygen on a thin membrane within which the chemicals can diffuse. Although the membrane contains an immobilized enzyme that catalyzes the reaction, the empirical model for describing the mechanism involves only the two chemicals and ignores the enzyme. In addition, since the membrane is thin, one can assume correctly that the mechanism takes place in a two-dimensional space.

I should like to suggest that a good candidate for the universal mechanism that generates the prepattern for mammalian coat patterns is a reaction-diffusion system that exhibits diffusion-driven spatial patterns. Such patterns depend strongly on the geometry and scale of the domain where the chemical reaction takes place. Consequently the size and shape of the embryo at the time the reactions are activated should determine the ensuing spatial patterns. (Later growth may distort the initial pattern.)

Any reaction-diffusion mechanism capable of generating diffusion-driven spatial patterns would provide a plausible model for animal coat markings. The numerical and mathematical results I present in this article are based on the model that grew out of Thomas' work. Employing typical values for the parameters, the time to form coat patterns during embryogenesis would be on the order of a day or so.

Interestingly, the mathematical problem of describing the initial stages of spatial pattern formation by reaction-diffusion mechanisms (when departures from uniformity are minute) is similar to the mathematical problem of describing the vibration of thin plates or drum surfaces. The ways in which pattern growth depends on geometry and scale can therefore be seen by considering analogous vibrating drum surfaces.

If a surface is very small, it simply will not sustain vibrations; the disturbances die out quickly. A minimum size is therefore needed to drive any sustainable vibration. Suppose the drum surface, which corresponds to the reaction-diffusion domain, is a rectangle. As the size of the rectangle is increased, a set of increasingly complicated modes of possible vibration emerge.

An important example of how the geometry constrains the possible modes of vibration is found when the domain is so narrow that only simple—essentially one-dimensional—modes can exist. Genuine two-dimensional patterns require the domain to have enough breadth as well as length. The analogous requirement for vibrations on the surface of a cylinder is that the radius cannot be too small, otherwise only quasi-one-dimensional modes can exist; only ringlike patterns can form, in other words. If the radius is large enough, however, two-dimensional patterns can exist on the surface. As a consequence, a tapering cylinder can exhibit a gradation from a two-dimensional pattern to simple stripes [*see illustration on opposite page*].

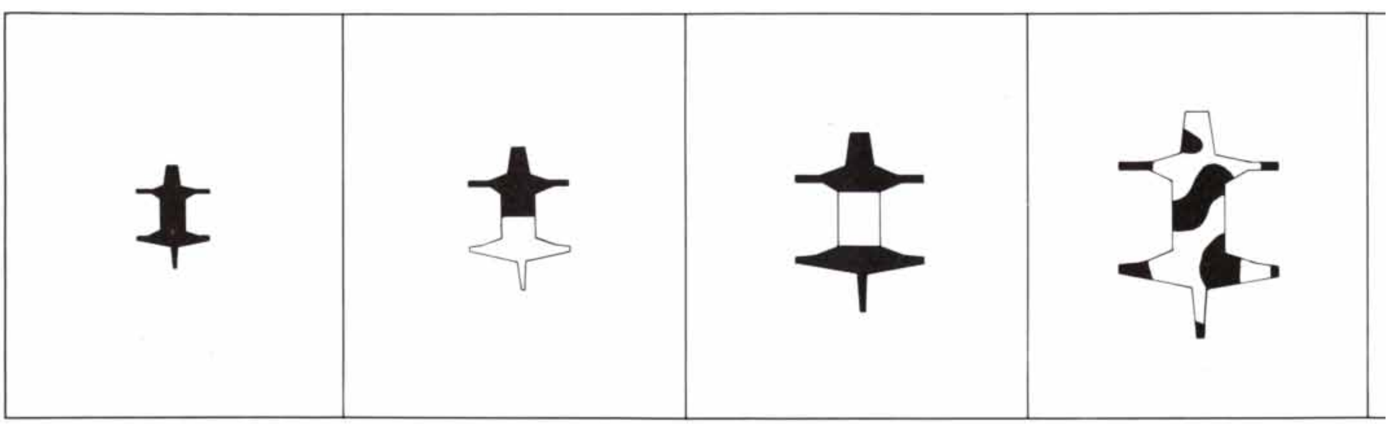
Returning to the actual two-morphogen reaction-diffusion mechanism I considered, I chose a set of reaction and diffusion parameters that could produce a diffusion-driven instability and kept them fixed for all the calculations. I varied only the scale and geometry of the domain. As initial conditions for my calculations, which I did on a computer, I chose random perturbations about the uniform steady state. The resulting patterns are colored dark and light in regions where the concentration of one of the morphogens is greater than or less than the concentration in the homogeneous steady state. Even with such limitations on the parameters and the initial conditions the wealth of possible patterns is remarkable.

EXAMPLES OF DRAMATIC PATTERNS occurring naturally are found in the anteater (*left*) and the Valais goat, *Capra aegagrus hircus* (*right*). Such patterns can be accounted for by the author's reaction-diffusion mechanism (*see bottom illustration on these*

How do the results of the model compare with typical coat markings and general features found on animals? I started by employing tapering cylinders to model the patterns on the tails and legs of animals. The results are mimicked by the results from the vibrating-plate analogue, namely, if a two-dimensional region marked by spots is made sufficiently thin, the spots will eventually change to stripes.

The leopard (*Panthera pardus*), the cheetah (*Acinonyx jubatus*), the jaguar (*Panthera onca*) and the genet (*Genetta genetta*) provide good examples of such pattern behavior. The spots of the leopard reach almost to the tip of the tail. The tails of the cheetah and the jaguar have distinctly striped parts, and the genet has a totally striped tail. These observations are consistent with what is known about the embryonic structure of the four animals. The prenatal leopard tail is sharply tapered and relatively short, and so one would expect that it could support spots to the very tip. (The adult leopard tail is long but has the same number of vertebrae.) The tail of the genet embryo, at the other extreme, has a remarkably uniform diameter that is quite thin. The genet tail should therefore not be able to support spots.

The model also provides an instance of a developmental constraint, documented examples of which are exceedingly rare. If the



SCALE AFFECTS PATTERNS generated within the constraints of a generic animal shape in the author's model. Increasing the scale and holding all other parameters fixed produces a remarkable variety of patterns. The model agrees with the fact that

two pages). The drawing of the anteater was originally published by G. and W. B. Whittaker in February, 1824, and the photograph was made by Avi Baron and Paul Munro.

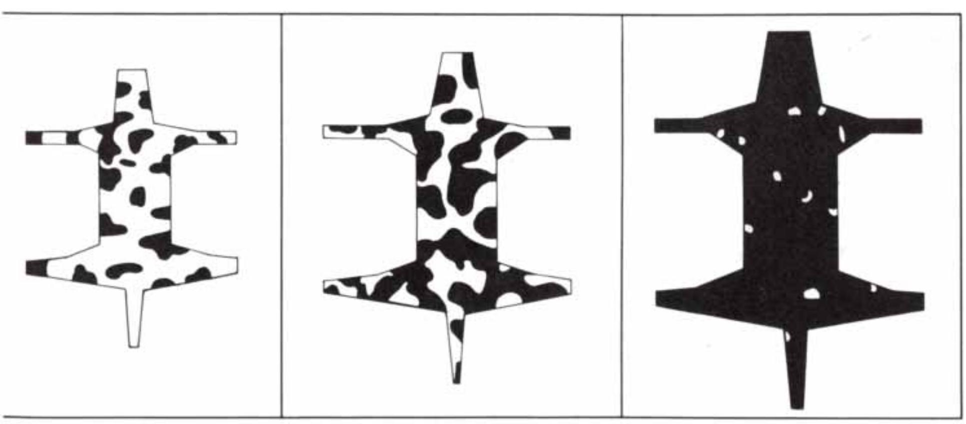
prepattern-forming mechanism for animal coat markings is a reaction-diffusion process (or any process that is similarly dependent on scale and geometry), the constraint would develop from the effects of the scale and geometry of the embryos. Specifically, the mechanism shows that it is possible for a spotted animal to have a striped tail but impossible for a striped animal to have a spotted tail.

We have also met with success in our attempts to understand the markings of the zebra. It is not difficult to generate a series of stripes with our mechanism. The junction of the foreleg with the body is more complicated, but the mathematical model predicts the typical pattern of leg-body scapular stripes [*see illustration on page 83*].

In order to study the effect of scale in a more complicated geometry, we computed the patterns for a generic animal shape consisting of a body, a head, four appendages and a tail [*see bottom illustration on these two pages*]. We started with a very small shape and gradually increased its size, keeping all the parts in proportion. We found several interesting results. If the domain is too small, no pattern can be generated. As the size of the domain is increased successive bifurcations occur: different patterns suddenly appear and disappear. The patterns show more structure and more spots as the size of the domain is increased. Slender extremities still retain their striped pattern, however, even for domains that are quite large. When the domain is very large, the pattern structure is so fine that it becomes almost uniform in color again.

The effects of scale on pattern suggest that if the reaction-diffusion model is correct, the time at which the pattern-forming mechanism is activated during embryogenesis is of the utmost importance. There is an implicit assumption here, namely that the rate constants and diffusion coefficients in the mechanism are roughly similar in different animals. If the mechanism is activated early in development by a genetic switch, say, most small animals that have short periods of gestation should be uniform in color. This is generally the case. For larger surfaces, at the time of activation there is the possibility that animals will be half black and half white. The honey badger (*Mellivora capensis*) and the dramatically patterned Valais goat (*Capra aegagrus hircus*) are two examples [*see top illustration on these two pages*]. As the size of the domain increases, so should the extent of patterning. In fact, there is a progression in complexity from the Valais goat to certain anteaters, through the zebra and on to the leopard and the cheetah. At the upper end of the size scale the spots of giraffes are closely spaced. Finally, very large animals should be uniform in color again, which indeed is the case with the elephant, the rhinoceros and the hippopotamus.

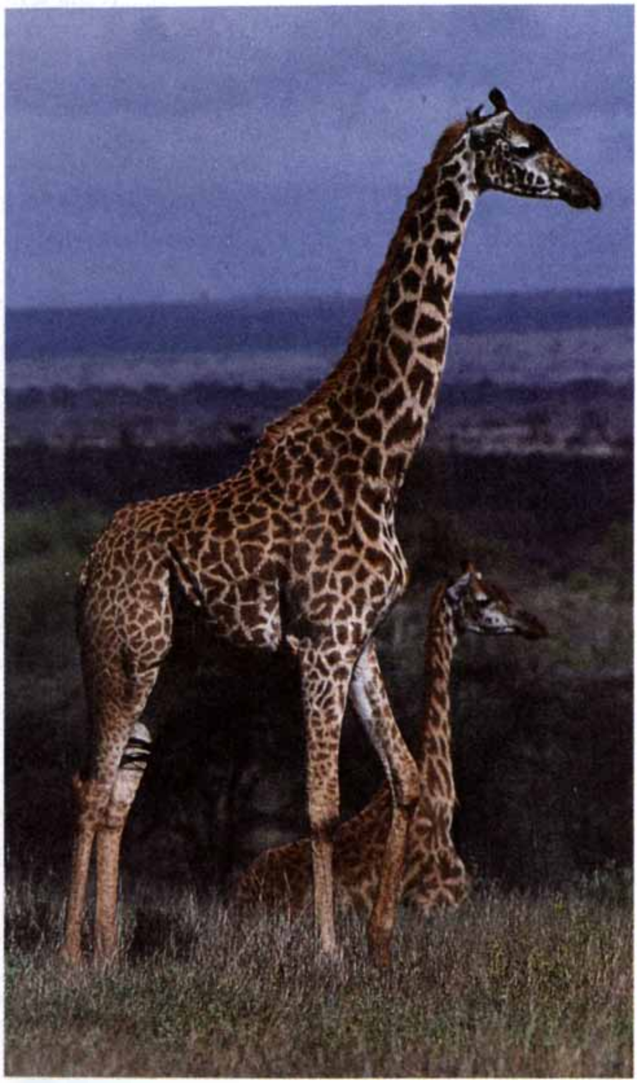
We expect that the time at which the pattern-forming mechanism is activated is an inherited trait, and so, at least for animals whose survival depends to a great extent on pattern, the mechanism is activated when the embryo has reached a certain size. Of course, the conditions on the embryo's surface at the time of activation exhibit a certain randomness. The reaction-diffusion model produces patterns that depend uniquely

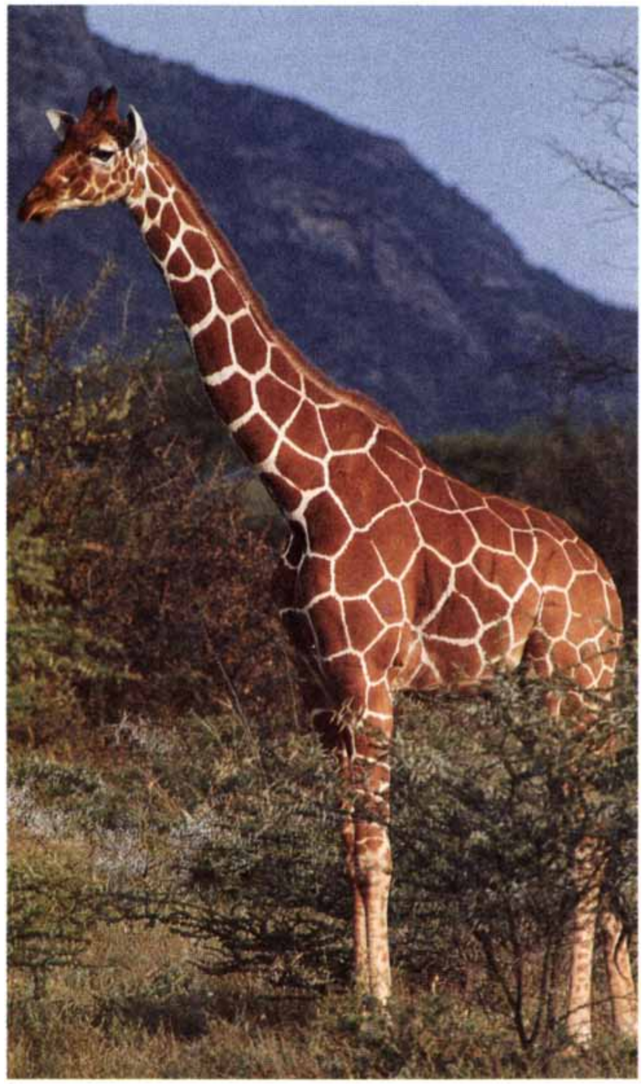


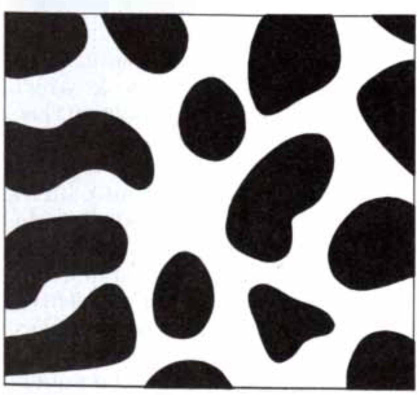
small animals such as the mouse have uniform coats, intermediate-size ones such as the leopard have patterned coats and large animals such as the elephant are uniform.

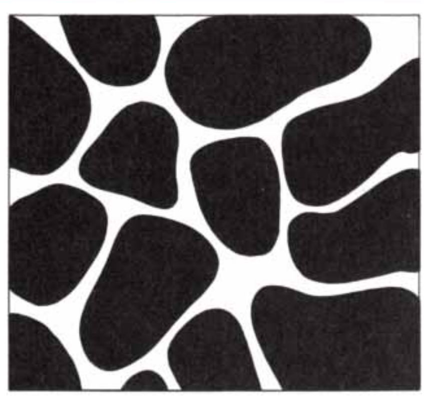
on the initial conditions, the geometry and the scale. An important aspect of the mechanism is that, for a given geometry and scale, the patterns generated for a variety of random initial conditions are qualitatively similar. In the case of a spotted pattern, for example, only the distribution of spots varies. The finding is consistent with the individuality of an animal's markings within a species. Such individuality allows for kin recognition and also for general group recognition.

The patterns generated by the model mechanism are thought to correspond to spatial patterns of morphogen concentrations. If the concentration is high enough, melanocytes will produce the melanin pigments. For simplicity we assumed that the uniform steady state is the threshold concentration, and we reasoned that melanin will be generated if the value is equal to or greater than that concentration. The assumption is somewhat arbitrary, however. It is reasonable to expect that the threshold concentration may vary, even within species. To investigate such









DIFFERENT GIRAFFES have different kinds of markings. The subspecies *Giraffa camelopardalis tippelskirchi* is characterized by rather small spots separated by wide spaces (*top left*); *G. camelopardalis reticulata,* in contrast, is covered by large, closely spaced spots (*top right*). Both kinds of pattern can be accounted for by the author's reaction-diffusion model (*bottom left and bottom right*). The assumption is that at the time the pattern is laid down the embryo is between 35 and 45 days old and has a length of roughly eight to 10 centimeters. (The gestation period of the giraffe is about 457 days.)

effects, we considered the various kinds of giraffe. For a given type of pattern, we varied the parameter that corresponds to the morphogen threshold concentration for melanocyte activity. By varying the parameter, we found we could produce patterns that closely resemble those of two different kinds of giraffe [*see illustration on opposite page*].

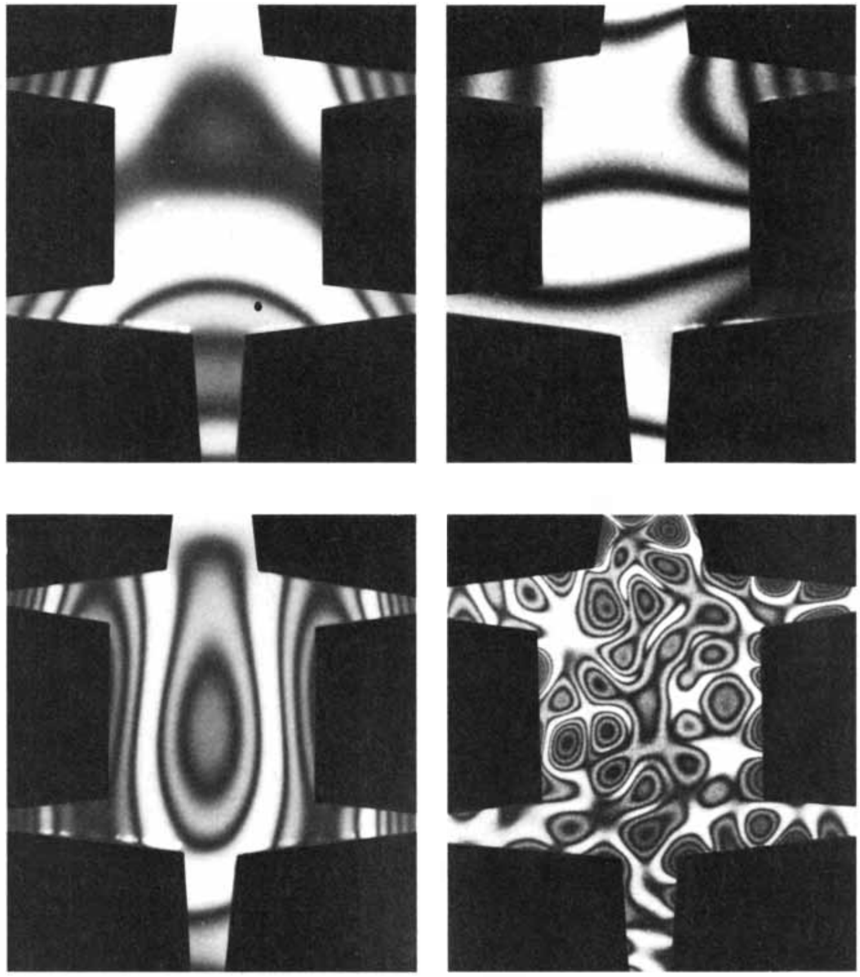
Recently the results of our model have been corroborated dramatically by Charles M. Vest and Youren Xu of the University of Michigan. They generated standing-wave patterns on a vibrating plate and changed the nature of the patterns by changing the frequency of vibration. The patterns were made visible by a holographic technique in which the plate was bathed in laser light. Light reflected from the plate interfered with a reference beam, so that crests of waves added to crests, troughs added to troughs, and crests and troughs canceled, and the resulting pattern was recorded on a piece of photographic emulsion [*see illustration at right*].

Vest and Youren found that low frequencies of vibration produce simple patterns and high frequencies of vibration produce complex patterns. The observation is interesting, because it has been shown that if a pattern forms on a plate vibrating at a given frequency, the pattern formed on the same plate vibrated at a higher frequency is identical with the pattern formed on a proportionally larger plate vibrated at the original frequency. In other words, Vest and Youren's data support our conclusion that more complex patterns should be generated as the scale of the reaction-diffusion domain is increased. The resemblance between our patterns and the patterns subsequently produced by the Michigan workers is striking.

I should like to stress again that all the patterns generated were produced by varying only the scale and geometry of the reaction domain; all the other parameters were held fixed (with the exception of the different threshold concentrations in the case of the giraffe). Even so, the diversity of pattern is remarkable. The model also suggests a possible explanation for the various pattern anomalies seen in some animals. Under some circumstances a change in the value of one of the parameters can result in a marked change in the pattern obtained. The size of the effect depends on how close the value of the parameter is to a bifurcation value: the value at which a qualitative change in the pattern is generated.

If one of the parameters, say a rate constant in the reaction kinetics, is varied continuously, the mechanism passes from a state in which no spatial pattern can be generated to a patterned state and finally back to a state containing no patterns. The fact that such small changes in a parameter near a bifurcation value can result in such large changes in pattern is consistent with the punctuated-equilibrium theory of evolution. This theory holds that long periods of little evolutionary change are punctuated by short bursts of sudden and rapid change.

Many factors, of course, affect animal coloration. Temperature, humidity, diet, hormones and metabolic rate are among some of them. Although the effects of such factors probably could be mimicked by manipulating various parameters, there is little point in doing so until more is known about how the patterns reflected in the melanin pigments are actually produced. In the meantime one cannot help but note the wide variety of patterns that can be generated with a reaction-diffusion model by varying only the scale and geometry. The considerable circumstantial evidence derived from comparison with specific animal-pattern features is encouraging. I am confident that most of the observed coat patterns can be generated by a reaction-diffusion mechanism. The fact that many general and specific features of mammalian coat patterns can be explained by this simple theory, however, does not make it right. Only experimental observation can confirm the theory.



STANDING-WAVE PATTERNS generated on a thin vibrating plate resemble coat patterns and confirm the author's work. More complex patterns correspond to higher frequencies of vibration. The experiments were done by Charles M. Vest and Youren Xu.

Fatness and Fertility

Loss of fat from dieting or exercise can lead to infertility that is reversible with fat gain. It is possible that fat tissue exerts a regulatory effect on the reproductive ability of human females

by Rose E. Frisch

Ever since the Stone Age, symbols of female fertility have been fat, particularly in the breasts, hips, thighs and buttocks—the places where estrogen, the female sex hormone, promotes fat storage. This historical linking of fatness and fertility actually makes biological sense; in fact, I propose that body fat, or adipose tissue, has a regulatory role in reproduction.

The evidence gathered in the past 15 years is strong. It indicates, for instance, that a female must store at least a threshold, or minimum, amount of body fat in order to begin and maintain normal menstrual cycles and hence have the ability to reproduce. Activities that reduce fat below the threshold, such as serious dieting and intensive exercise, can delay the age of menarche (the first menstrual cycle) to as late as 20 years. Such a loss can also "silently" halt ovulation—the midcycle release of an egg from the ovary—in someone who menstruates every month or cause frank amenorrhea: the absence of menstrual cycles. The resulting infertility is reversible by weight gain or reduction of activity or both.

A woman need not suffer from anorexia nervosa (the psychologically driven syndrome of self-starvation in which as much as a third of the body weight is lost) in order to induce menstrual irregularities. Even a rather moderate loss, in the range of from 10 to 15 percent below the normal weight for height, which is primarily a loss of fat, is sufficient.

Recent work has shown that the menstrual disorders associated with excessive leanness derive from the abnormal activity of the hypothalamus, the part of the brain that regulates reproduction as well as food intake and other basic functions. It has long been known that the hypothalamus receives information from higher centers of the brain. External factors such as temperature and stress can therefore affect the reproductive activity of the hypothalamus. It is now certain that nutrition and physical effort also have an effect.

It is not surprising that the reproductive function of the hypothalamus falters when a woman becomes too lean. Such a response would have given our female ancestors a selective advantage by ensuring that they conceived only when they could complete a pregnancy successfully. Reproduction, after all, requires energy, or calories: some 50,000 to 80,000 calories to produce a viable infant and then from 500 to 1,000 calories a day for lactation. (These figures are over and above the calories needed for other life processes.) In ancient times, when the food supply was scarce or fluctuated seasonally and when breast milk was a newborn's only food, a woman who became pregnant when she lacked an adequate store of body fat—the most readily mobilized fuel in the body—could have endangered both her own life and that of her developing fetus and newborn infant.

Indeed, one can speculate that females who continued to ovulate in spite of being undernourished left no viable offspring or did not survive themselves; they therefore left no descendants. Thus natural selection had its way: today most women have more than a fourth of their weight in fat (about 16 kilograms, or 35 pounds, representing some 144,000 calories) when they reach maturity. The main function of this fat store may be to provide energy for a pregnancy and about three months of lactation. Men, in contrast, have roughly from 12 to 14 percent of their weight in fat at maturity. Studies of very obese women (and animals) show that excessive fatness, like excessive leanness, is also associated with amenorrhea and infertility; the mechanisms are not yet known.

The discovery of a relation between fatness and fertility came about somewhat indirectly. Early in the 1970's I worked with Roger Revelle, who was then at Harvard University, gathering data on the heights, weights and calorie supplies of populations in Latin America and Asia in order to estimate the food needs of the world. In the course of analyzing the growth data I noted that a poor food supply delayed the age at which adolescent girls had their most rapid growth in weight, an event that precedes menarche; abundant calorie supplies advanced the age of the spurt. This was not particularly surprising, but another finding was: within each population the growth spurt, whether it was early or late, came when girls had achieved the same average weight. In other words, weight appeared to be the determining factor.

To explore the implied relation between weight and menarche, we then analyzed the height and weight data for 181 girls who had been followed from birth to age 18 (when growth was completed) in three comparable U.S. studies. We found that both the early and the late maturers had the same average weight (47 kilograms, or 103 pounds) at menarche; in contrast, the average height increased significantly with menarcheal age.

The finding that weight seemed to be important to the timing of menarche helped to explain many puzzling observations. For example, in the past century girls in the U.S. and Europe have attained menarche progressively earlier. On the average, American girls now begin to menstruate when they are 12.6 years old;

a century ago the age was 15.5 years. Revelle and I postulate that the earlier menarche is explained by the fact that children now become bigger sooner because they are better nourished and have less disease. The association between average weight at menarche and the timing of menarche in a population also explained why malnutrition delays menarche, why twins (who grow relatively slowly in utero and postnatally and hence take longer to reach the critical weight for menarche) begin to menstruate later than "singletons" in the same population and why high altitude (which slows the rate of growth) also delays menarche.

Knowing the average weights at menarche did not enable us to predict the timing of menarche for individual girls, whose weights at menarche vary considerably. I wondered if there was some weight-related factor that would make such prediction possible. Elegant animal studies by the late Gordon C. Kennedy of the University of Cambridge provided a clue. Kennedy found that puberty in the female rat is more closely related to body weight than to age. He also found that rats given different diets that altered their growth rates ate the same amount of food per unit of body weight at first estrus (the animal equivalent of menarche). He postulated that the feeding signal and the onset of estrus may both be regulated by the amount of fat in the body.

On the basis of Kennedy's findings—as well as those of his associates Elsie Widdowson and Peter A. McCance and work done on the composition of the human body by Francis D. Moore and his co-workers at the Harvard Medical School—I began to suspect that some aspect of body composition, particularly stored fat, in the human female might also influence the timing of the first menstrual cycle in girls. My colleagues and I therefore investigated the changes in body composition in girls during the adolescent growth spurt and at menarche. Specifically, we looked at the amounts of both fat and lean body mass, which together account for one's weight. The lean mass, or the nonfat parts of the body, includes the muscles, the fat-free skin, the viscera (the heart, kidney and other organs) and the skeleton.

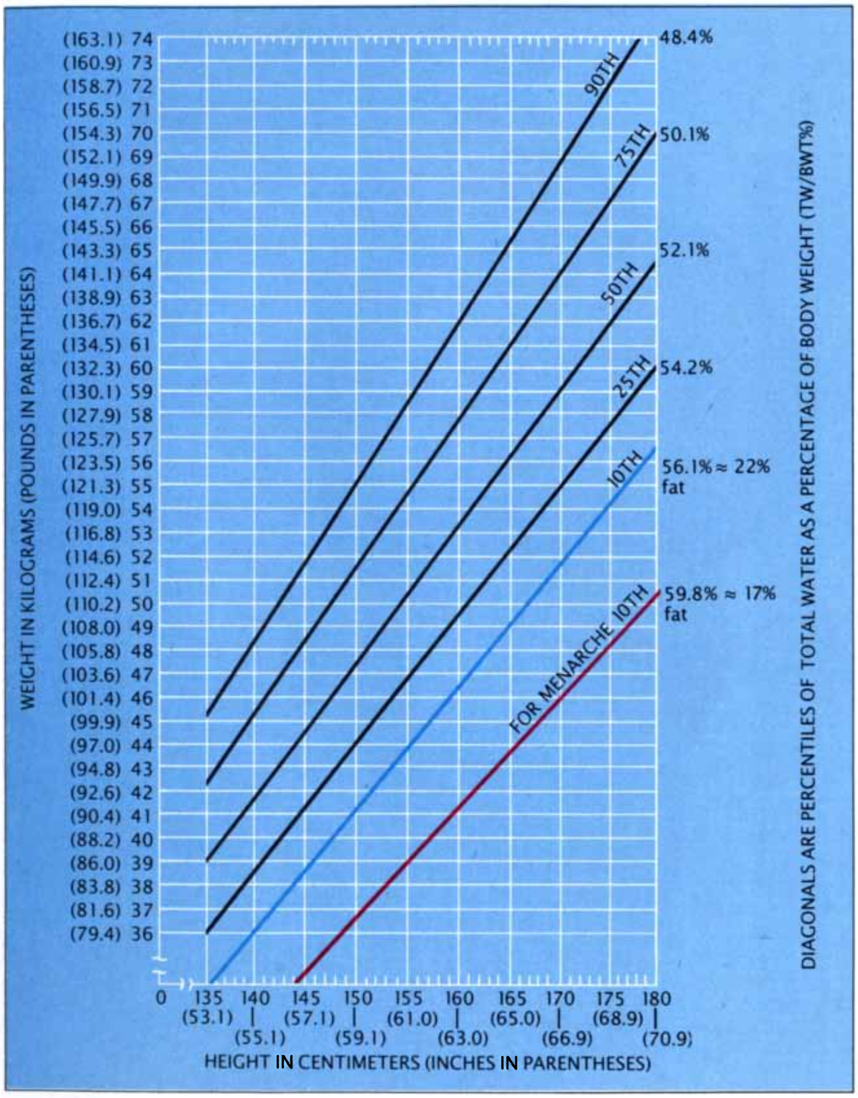
The 181 girls in the longitudinal studies had long since grown up, but we had their heights and weights at menarche and throughout the adolescent growth spurt. We could there-

STONE AGE FERTILITY FIGURE, the Venus of Willendorf, is an exaggerated representation of a recently established fact: a woman must have an adequate amount of body fat in order to reproduce successfully. The critical amount can usually be acquired by gaining a prescribed amount of weight, since most weight gain in nonathletic adults is actually a gain of fat. The figure, from Austria, is approximately four inches in height.

fore calculate their body composition by inserting this information into standard equations that estimate the total amount of water in the body; this value makes it possible to determine the relative contributions of lean mass and of fat to a person's total weight. The equations are based on direct measurements of body water that had been made on a normal group of girls and women.

One might ask: How could measurements of body water provide information about the composition of the entire body? The answer lies in the fact that after childhood 72 percent of the lean mass of the body is accounted for by water. Knowing the total amount of water in the body, one can determine the weight of the lean mass, which is equal to the amount of water divided by .72. Subtracting the weight of the lean mass from the total body weight yields the weight of the fat. The percentage of water in the entire body is a fatness index. Since fat has little water (from 5 to 10 percent), the more fat one has, the lower will be the percentage of water in the entire body.

We found that the greatest change in estimated body composition during the premenarcheal weight spurt is a large increase in body fat: on the average, both early- and late-maturing girls had a 120 percent increase in body fat (from five to 11 kilograms), whereas they had only a 44 percent increase in the weight of lean mass. Therefore the ratio of lean mass to fat declined from about 5 : 1 at the initiation of the growth spurt to about 3 : 1 at menarche. What all the girls had in common at menarche, even though individual weights varied, was a similar ratio of lean mass to fat and a similar amount of water as a percentage of body weight. On the average at menarche 55 percent of the body weight of the girls was water (a "body-water percentage" of 55), which indicated that an average of 24 percent of the weight was accounted for by body fat.

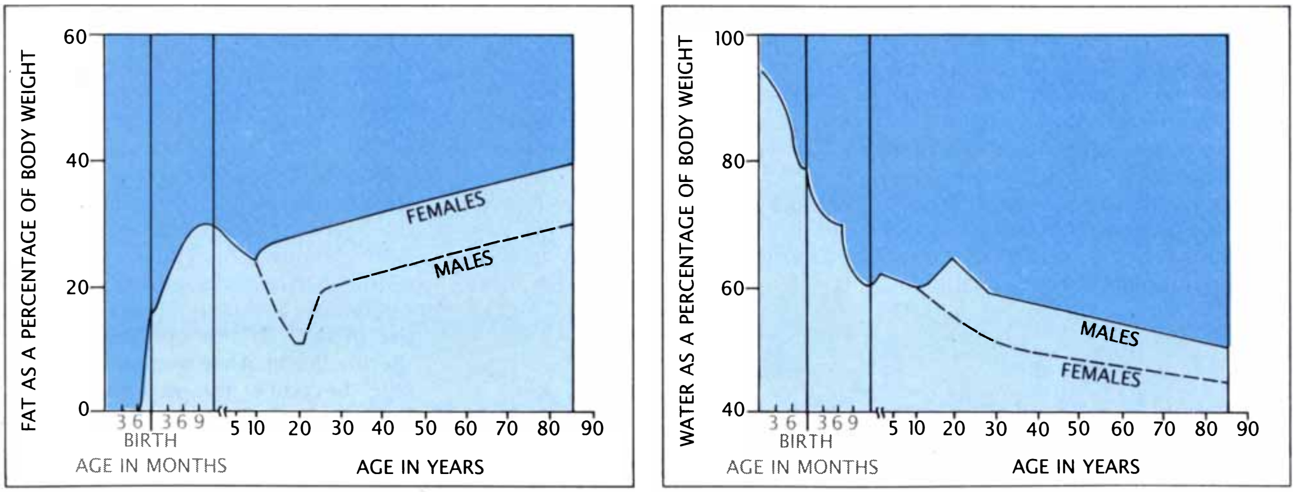


NOMOGRAM in clinical use indicates the threshold, or minimum, weight a female of a given completed height must attain in order to have normal menstrual cycles. How much weight must be gained above that threshold varies with the individual and cannot be predicted. The red line shows the minimum weight necessary for menarche (the initiation of menstrual cycles). The blue line shows the minimum weight for correcting secondary amenorrhea (the cessation of menstruation) in a mature woman. For example, according to the nomogram an amenorrheic woman 165 centimeters tall (five feet five inches) would have to reach at least 49 kilograms (108 pounds) before menstruation would resume. The top five diagonal lines indicate percentiles of total water as a percentage of body weight—which is an index of fatness—for fully grown, mature women in a normal sample; the blue line is the 10th percentile. The red line represents the 10th percentile for the same sample at menarche. (The 50th-percentile line indicates the normal weight for height of mature women who are from 18 to 25 years old.)

The findings strongly suggested that girls must reach a threshold lean/fat ratio, and hence a certain degree of fatness, in order to menstruate. Still, the average ratio for the population did not predict the threshold for any given individual. Janet W. McArthur of the Massachusetts General Hospital and I tackled this problem by comparing the height, weight and estimated body composition of young girls and women who had menstrual disorders with the same measures for our normal subjects at menarche and at age 18.

We compared the groups by means of charts called nomograms, which indicated the weights for heights that corresponded to the levels of relative fatness observed in the 181 normal girls at menarche and at the completion of growth [*see illustration at left*]. The relative-fatness levels were indicated on the chart by percentiles of our fatness index (total water as a percentage of body weight).

Somewhat to our surprise, we discovered that cases of delayed menarche and amenorrhea were each associated with being below a threshold value of relative fatness. Specifically, the comparison showed that in order to have menarche, young girls whose height growth was nearly or fully completed had to gain enough weight for their height to decrease their body-water percentage to 59.8



PERCENTAGE OF BODY WEIGHT represented by fat (*left*) increases with age in females after age 10 and in men considerably later. This trend is accompanied by a decrease in the water percentage (*right*), because fat has little water compared with the muscles and other parts of the lean body mass. The "body-water percentage" is therefore an index of fatness. By age 18 women are typically about 28 percent fat; men are about 13 percent fat. The large fat store in women is presumably required to supply energy for pregnancy and lactation. The graphs are based on data from Bent Friis-Hansen of the University of Copenhagen.

percent. This would ensure that they had at least 17 percent of their weight in fat (the lower the body-water percentage, the fatter the body). The threshold body-water percentage was the 10th-percentile value of normal girls at menarche. That is, 90 percent of the sample at menarche had a lower value and so were fatter. For a 15-year-old girl whose completed height was 165 centimeters (five feet five inches), the numbers translated into the prediction that she had to weigh at least 44 kilograms (97 pounds) before menarche would be expected.

The comparisons also showed that women who were amenorrheic because of simple weight loss needed to reach a weight that was about 10 percent heavier than the menarche threshold in order to restore and maintain normal ovulatory cycles. They had to have a total body-water percentage of no more than 56.1, indicating that fat accounted for about 22 percent of the body weight. This value was the 10th percentile of the body-water percentage at completion of growth for the normal girls at the age of 18. A woman this age or older whose height was 165 centimeters had to weigh at least 49 kilograms (108 pounds) before her cycles could be expected to resume. (Also, cycles would stop if her weight fell below this threshold.)

We were somewhat puzzled to find that heavier weights and higher fat percentages were necessary for regular, ovulatory cycles in the mature group. Why should one need more fat at age 18 than at menarche? An analysis of the data from the normal sample provided the answer: normal women gain fat during the interval between their first menstrual cycle and age 18. (By 18 they typically have completed the phase of adolescent subfertility, when the ovary, uterus and oviducts are still growing, and there are many anovulatory cycles.) We found that both early- and late-maturing girls gain an average of 4.5 kilograms (10 pounds) of fat between menarche and age 18, when our normal subjects had an average of 28 percent of their body weight in fat. At first we could hardly believe the human female stores such a large amount of fat when she is ready to reproduce, but data in the literature based on postmortem dissections confirmed this value.

Many physicians now rely on the nomograms as a guide to recommending weight increases for underweight women with amenorrhea or anovulatory cycles. (Of course, all pathology must be ruled out.) In many cases a weight increase of merely three to five pounds above the threshold weight for height results in the resumption of menstruation; it is not yet possible, however, to predict the precise amount any given individual has to gain above the threshold. At body weights close to the required minimum, women may still have anovulatory cycles even if they are menstruating. (Ultrasound images of the ovary or measurements of hormone levels would reveal the disorder.) A further weight gain—up to the range indicated by the 25th percentile for body-water percentage—could be necessary to ensure regular, ovulatory cycles. The narrowness of the weight threshold is actually quite surprising: some athletes we studied turned their menstrual cycles on and off at will with just a three-pound weight change.

I must note a few caveats here. Because factors other than weight, such as emotional stress, affect the onset or maintenance of menstruation, cycles that are disrupted by weight loss may not resume in some women even though they reach the minimum weight for their height. Also, the time it takes before the cycles resume with weight gain varies with the length of time a woman was amenorrheic. I should also point out that our nomograms apply as yet only to U.S. or European females and do not apply to extremely muscular women. Because muscles contain a lot of water (80 percent), which is heavy, they are heavy themselves; hence a muscular woman with a normal weight for her height may actually have little fat in relation to her lean mass.

By what mechanism might a modification in the amount of body fat affect fertility? The critical role of the hypothalamus in controlling reproduction was established by measurements of a hypothalamic hormone called gonadotropin-releasing hormone (GnRH). This substance, which is secreted in pulses, controls the chain of events leading to ovulation. In underweight or excessively lean women the pattern of secretion of gonadotropin-releasing hormone is abnormal in amount and timing and is similar to that of prepubertal girls. As a result the cascade of hormonal events that normally leads to ovulation and prepares the uterus to support a pregnancy is disrupted.

In the mature female body GnRH pulses stimulate the pituitary gland to release two other hormones: follicle-stimulating hormone, which controls the growth of an ovarian follicle (specialized cells that encase an egg), and luteinizing hormone, which controls the cyclical release of the egg from the follicle. When these hormone levels are low, ovulation cannot occur.

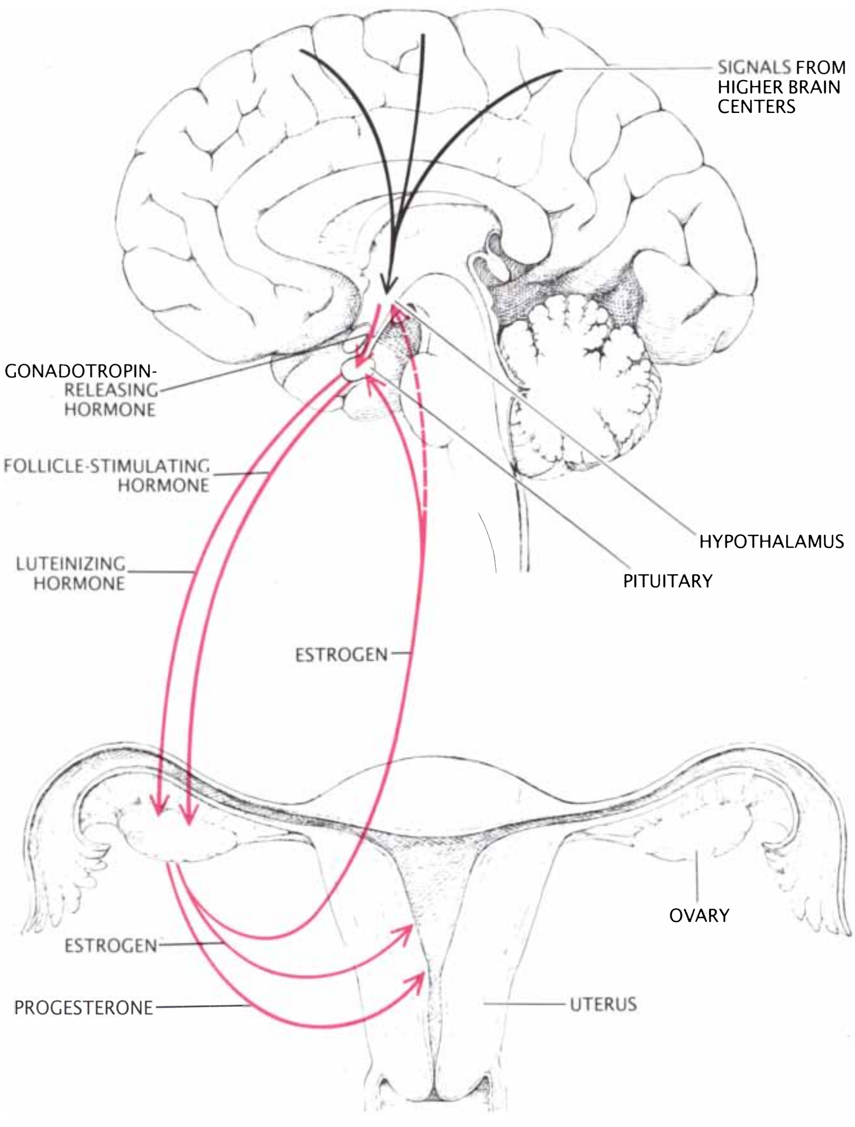
Estrogen and progesterone secreted by the ovary are also necessary for ovulation and menstruation. These hormones are affected by a decline in the levels of gonadotropin-releasing hormone and the consequent changes in follicle-stimulating hormone and luteinizing hormone. The growing follicle normally releases estrogen, which modulates the hormonal secretions of the pituitary, leading to a midcycle surge of luteinizing hormone and hence to ovulation. Estrogen also stimulates the growth of the breasts and the uterine lining. After ovulation the follicle becomes the corpus luteum ("yellow body") and secretes progesterone, which increases the vascularity of the uterine lining in preparation for implantation of the fertilized egg. If no egg is implanted, the levels of progesterone and estrogen fall and the monthly flow of blood ensues.

Just what signal the hypothalamus responds to when menstruation is disrupted is still not certain. It may receive signals from the abnormal temperature control or the changes in metabolism that are associated with loss of body fat due to undernutrition and a high expenditure of energy. It is also possible that the hypothalamus may receive signals from changes in estrogen that result from fat loss.

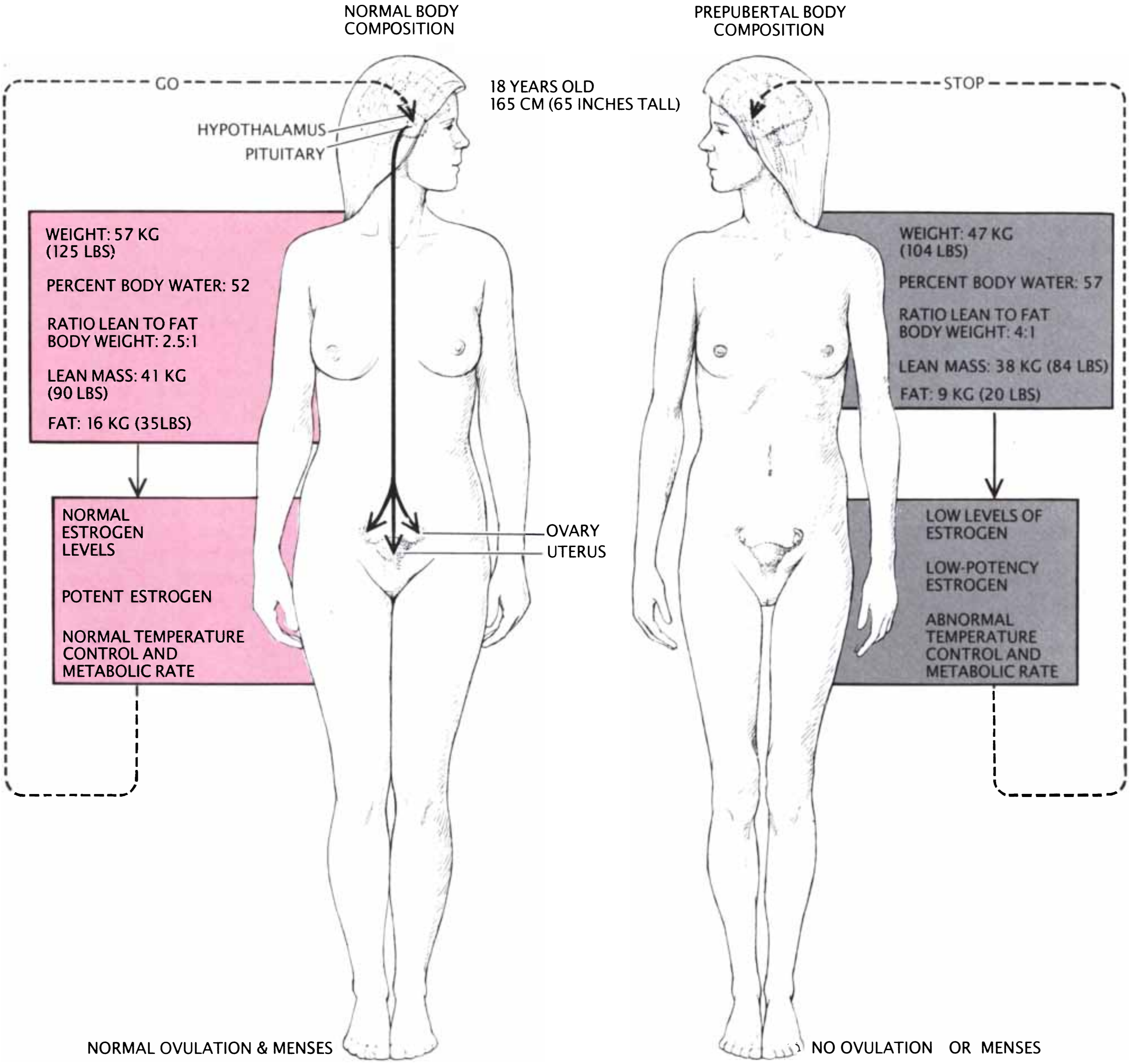
How can fat make a difference to estrogen? Adipose tissue was once thought to be inert and to merely insulate and cushion the body. It is now known to be quite active in the turnover of fuels in the body. It also stores steroids (sex hormones) and influences the amount and potency of estrogen circulating in the blood.

Several research groups have delineated the interrelation between body fat and estrogen. Pentti K. Siiteri of the University of California at San Francisco School of Medicine and Paul C. Macdonald of the Cecil H. and Ida Green Center for Reproductive Biology Sciences in Dallas have found that fat tissue converts androgen (male hormones) into estrogen. This conversion accounts for roughly a third of the estrogen that circulates in the blood of premenopausal women and is the main source of estrogen in postmenopausal women. (Men also convert androgen into estrogen in body fat.) The fat of the breast, abdomen, omentum (the apron of fat in the abdomen) and, as my colleagues and I have found, the fatty marrow of the long bones all convert androgen into estrogen.

Jack Fishman, H. Leon Bradlow and their associates at Rockefeller University have shown that whether an-



HORMONAL SECRETIONS (*colored arrows*) that lead to ovulation and to regularly timed menstrual cycles are controlled by the hypothalamus. Pulses of gonadotropin-releasing hormone secreted by the hypothalamus cause the pituitary gland to release follicle-stimulating hormone, which controls the growth of an ovarian follicle (specialized cells that encase an egg), and luteinizing hormone, which controls ovulation (the midcycle release of the egg). In the first half of the cycle the growing follicle secretes estrogen. This hormone modulates the activity of the pituitary and perhaps the hypothalamus (*broken arrow*), and it stimulates the growth of the breasts and the lining of the uterus. The ruptured follicle becomes the corpus luteum, which secretes progesterone to promote the vascularization of the uterine lining. If the egg is not fertilized, estrogen and progesterone levels fall and the lining is shed (menstruation). Hypothalamic function can be affected by signals (*black arrows*) from higher centers of the brain.



WOMAN of normal weight (*left*) who loses about 15 percent of her body weight (*right*) may not be shockingly thin but may stop menstruating anyway. Such a weight loss, which is mainly a loss of fat, interferes with the ability to menstruate because hypothalamic functioning is disturbed. How the effects of fat loss are perceived by the hypothalamus is not known. The author proposes that the hypothalamic control of reproduction may be disrupted by decreases in the concentration and potency of circulating estrogen and by disturbances in the body's ability to control its temperature and to maintain a normal metabolic rate.

drogen is converted into a potent or a nonpotent form of estrogen is related to how fat one is. For instance, anorectic lean girls have elevated levels of a relatively inactive form of estrogen, whereas fatter women produce less of this form and have an elevated level of a highly potent type. Siiteri and his colleagues also found that obese women have a relatively low level of a substance known as sex-hormone-binding globulin; as its name suggests, the hormone binds estrogen. Low levels of this binding protein result in a high concentration of free estrogen in the circulation. Leaner girls at menarche have higher amounts of the binding protein and therefore less free estrogen.

Several recent studies, including our own, have shown that dieting is not the only way women become lean enough to impair their hypothalamic function and disrupt menstruation. Well-trained athletes of all kinds, such as runners, swimmers and ballet dancers, have a high incidence of delayed menarche, irregular cycles and amenorrhea. This pattern implies that exercise could be the cause—presumably by building muscles and reducing fat, thus raising the ratio of lean mass to fat.

My colleagues and I have been able to establish that regular intensive exercise is indeed the explanation. We found that collegiate swimmers and runners whose training began before menarche first menstruated at an average age of about 15 years; women whose training began later had an average menarcheal age of about 12.7 years—similar to that of both our nonathletic control group and the general population. In our sample each year of premenarcheal training delayed menarche by five months.

Our data also showed that training that begins before menarche is associated with a high incidence of menstrual irregularities. Of our college athletes who had begun training before menarche, only 17 percent had regular cycles; 61 percent had irregularly timed cycles and 22 percent did not menstruate at all. On the other hand, 60 percent of the women who began training after menarche had regular cycles, 40 percent had irregular cycles and none were amenorrheic. During intensive training the incidence of irregular cycles and amenorrhea increased in both groups, in association with loss of body weight and increased leanness.

Measurements of hormones added support for the suggestion that fat loss contributes to menstrual disorders in athletes. The findings indicated that lean athletes with disrupted menstrual cycles or late menarche had low levels of estrogen and low levels of luteinizing hormone. Cessation of exercise because of an injury restored the hormones to normal levels, and the cycles resumed. Recent measurements made by other workers have further shown that well-trained athletes with irregular cycles and amenorrhea had hypothalamic dysfunction: the levels of gonadotropin-releasing hormone were abnormal and were similar to those of underweight women, as would be expected if a loss of fat and an increase of muscle mass were the cause of menstrual disorders.

We have also studied a female body builder who does not run, jog or dance but who concentrates on building her muscles with exercise machines. She becomes amenorrheic when she is in shape for competition. Measurements of her hormones have shown that her levels of estrogen, follicle-stimulating hormone and luteinizing hormone are as low as those of dieting and athletic women.

New findings from former college athletes add a different kind of confirmation that relative fatness has an important effect on the human reproductive system. A group of us in Boston studied data on 5,398 college graduates between 20 and 80 years old, half of them former athletes and half of them not. The former athletes had a significantly lower lifetime occurrence of breast cancer and cancers of the reproductive system than the nonathletes. The analysis took into account possible confounding factors, such as age, family history of cancer and smoking. The likeliest explanation is that the former athletes (who were leaner than the nonathletes in every age group) had lower estrogen levels and a higher proportion of low-potency estrogen. Apparently the same factors that, when extreme, can cause infertility in underweight or athletic women also exert a protective effect against cancers that are sensitive to sex hormones.

For direct evidence of a connection between body composition and fertility one has to turn to animals, which can be dissected to determine the weight of the lean body mass and fat. At present all methods for measuring body fat in the human being are indirect, including underwater weighing. D. Mark Hegsted, Koji Yoshinaga, both then at Harvard, and I have shown that rats given a high-fat diet had first estrus significantly earlier than rats fed a low-fat diet of equal calories. Direct carcass analysis of the tissues showed that both groups had a similar body-water percentage at first estrus and hence a similar lean/fat ratio even though their absolute body weights differed.

Other animals provide more circumstantial but nonetheless interesting data. Consider the case of the so-called double-muscled Charolais cattle, whose particularly lean meat makes them a desirable commodity. Unfortunately for the farmers who breed them, these animals have delayed puberty and difficulty reproducing. The Charolais bulls similarly are relatively infertile.

Another set of findings relates to an unusual phenomenon known as flushing: an increase in the rate of twin births in sheep when they are fed a high-calorie diet for a short time—perhaps a week—before mating. Sheep farmers routinely capitalize on this phenomenon to increase the number of twin births. The effect is so strong that one can calculate the number of twin lambs that will be born to a healthy ewe on the basis of the number of calories she eats.

The well-nourished human female fortunately does not normally superovulate in response to a high-calorie meal (at least as far as we know). Nevertheless, there is evidence for a residual flushing effect. The rate at which women gave birth to fraternal twins (from two eggs) but not to identical twins (from one egg that divides) declined in Holland during World War II when there were food restrictions; the rate of birth of fraternal twins returned to normal after the food supply became more plentiful.

Data from both the animal and the human studies suggest an explanation for the variation in the natural fertility of human populations, both historically and today. The total number of live children born to couples who do not use contraception can vary from as few as four among the hard-living nomadic Bush people of the Kalahari Desert in Africa to an average of 11 children among the well-nourished Hutterites, a prosperous, noncontracepting religious sect in the U.S. I explain these differences in terms of a direct pathway from food intake to fertility.

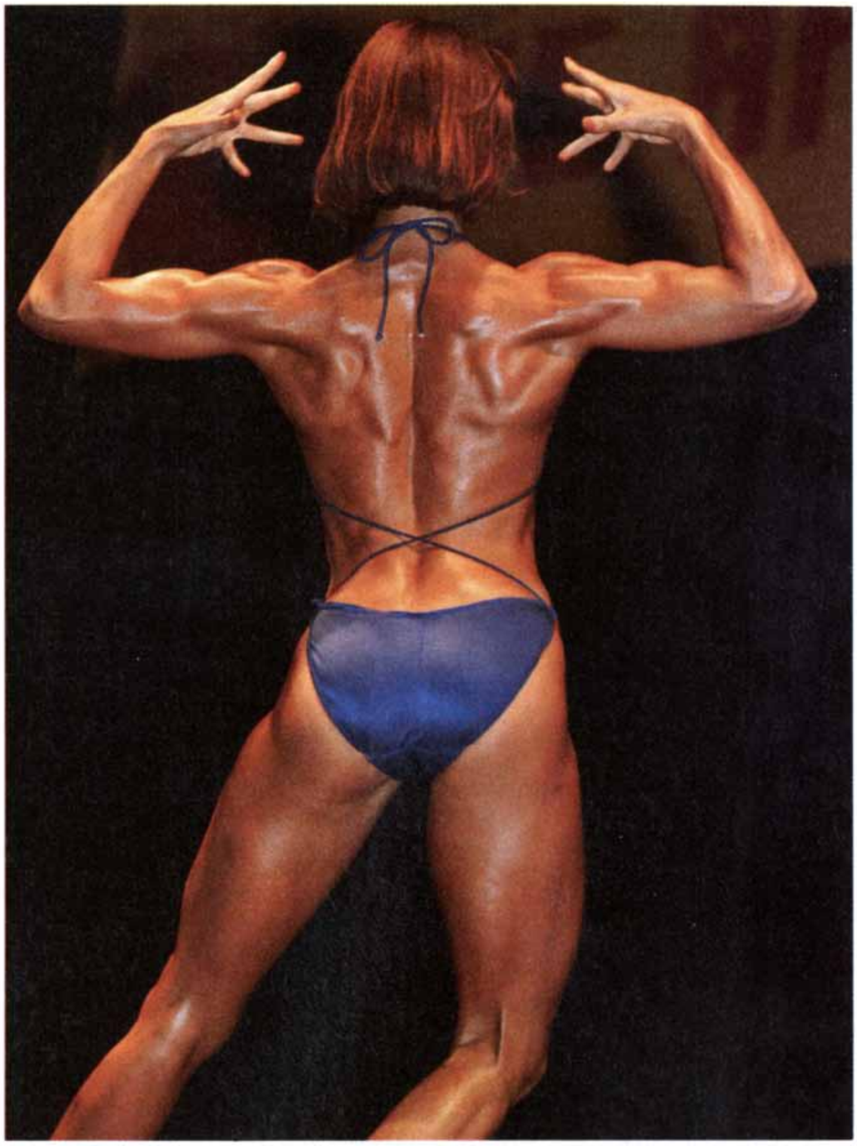
The idea is not completely new. Charles Darwin described this commonsense relation years ago with several observations: domestic animals (which have a regular and plentiful food supply) are more fertile than their wild relatives; "hard living retards the period at which animals conceive"; the amount of available food affects the fertility of a given animal, and it is difficult to fatten a lactating cow. I have shown that Darwin's dicta apply to human beings and that food supplies can affect fertility throughout the life span.

For instance, data on growth and reproduction of women in Great Britain in the mid-1800's show that poorly nourished females, who grew to maturity relatively slowly (a pattern that results in less fat per unit of lean mass), differed from well-nourished females in each event of the reproductive span. Their menarche was later, adolescent subfecundity lasted longer, their age of peak fertility came later, the number of live births in a given age group was smaller and the number of unsuccessful pregnancies was larger. Moreover, the duration of amenorrhea during breast-feeding was longer and the interval between births was therefore longer, and the age of menopause (when menstrual cycles cease) was earlier. In that British population, as in many other populations of the past, poor couples still living together at the end of their reproductive lives had had only six or seven living births.

Most poor couples in developing countries today also have about that number. This may seem like many offspring, but the number is actually well below the average human capacity. The relative infertility of the women can be explained by the fact that they tend to be undernourished and also to perform hard physical labor. If undernourished women have fewer children than well-nourished ones who do not practice contraception, why is there such rapid population growth in the developing countries? The paradox is explained by the fact that the death rate in these countries has been reduced by modern public-health measures while the birthrate has remained the same.

Whether they live in developing or developed countries, women need to be aware that they can become pregnant even though they are breastfeeding. In developing nations nursing mothers who are ill nourished generally do not resume ovulating and menstruating until a year or more after they give birth. Women who are well nourished, however, can resume ovulatory cycles as soon as three months after delivery, even if they are nursing full time. Having too short an interval between births is potentially dangerous because pregnancy is an energy drain on the mother and can cause the baby she is carrying to have a low birth weight; this would threaten the child's survival and put it at risk for neurological and other problems. Moreover, hormonal changes that accompany a pregnancy reduce the amount of breast milk the mother produces. Particularly in poor countries the health of the already-suckling child would then be jeopardized; in such nations it is important that mothers nurse for as long as possible because infants are often weaned onto diets that are low in protein.

Some investigators have argued that the pattern of suckling is the only factor that affects the length of "lactational amenorrhea." Yet studies in Africa by Peter G. Lunn and Roger G. Whitehead of the University of Cambridge show that nutrition and physical work clearly affect the time to resumption of regular, ovulatory menstrual cycles.



BODY BUILDER has less fat and a great deal more muscle than nonmuscular women of equal weight have. This young woman typically stops menstruating when she is in shape for competition, as in this picture. Measurements have shown that her levels of estrogen, follicle-stimulating hormone and luteinizing hormone are also low at those times, as would be expected from her very lean body composition. Although athletic women who train intensively are not as muscular as body builders, they too may become amenorrheic because of an increase in their muscle and a decrease in body fat.

My work has focused on the female, but undernutrition and weight loss also affect the male's ability to procreate. In a classic study of starvation in men Ancel Keys, Josef M. Brozek and their associates at the University of Minnesota found that a decrease in calorie intake and subsequent weight loss first cause a loss of libido. Continued weight loss results in a reduction of prostate fluid and then in lessened motility and longevity of sperm; the production of sperm is reduced when men weigh approximately 25 percent less than the normal weight for their height. Weight gain restores function in the reverse order of the loss.

Recent studies of male marathon runners and other top-ranking lean athletes have shown that many male athletes, like female ones, have hypothalamic dysfunction: the hypothalamus of many subjects secreted too little gonadotropin-releasing hormone, or released it in an abnormal pattern, causing the levels of testosterone (a male hormone) to be lower than normal. Whether such athletes may also have decreased fertility is not yet known.

There is as yet little discussion of the significance of fertility changes related to fatness in men. In contrast, the value of temporary infertility in women who are too thin was recognized a century ago. In 1884, when there were many cases of nutritional amenorrhea, the physician J. Matthews Duncan advised the Royal College of Physicians to treat a so-called sterile woman with an ample diet, such as roast beef and a French wine. "If a seriously undernourished woman could get pregnant," he said, "the chance of her giving birth to a viable infant, or herself surviving the pregnancy, is infinitesimally small."

The Quantum-Effect Device: Tomorrow's Transistor?

The components of ordinary integrated circuits can be made only so small before disruptive effects impair their function. Beyond that size limit a new species of semiconductor device could take over

by Robert T. Bate

The electronics industry and integrated circuits share an inverse destiny. The industry grows as circuits shrink, and growth will continue as long as more and more circuits can be crammed on a single chip. But common sense and careful analyses indicate that perhaps within a decade downscaling will run up against the limits of circuit technology. Even if practical limits are overcome, the physical laws that govern the behavior of circuit components set fundamental limits on the size of the components' features. In order to keep expanding, the electronics industry needs another technological revolution.

As a physicist with Texas Instruments, Incorporated, I have for many years been aware of the urgency of developing a new frontier for semiconductor devices. In 1982 my colleague Pallab K. Chatterjee published a study that heightened my concern by stressing how close the downscaling endpoint was. There is still some disagreement over that figure, with estimates of minimum feature sizes ranging between 100 and 500 billionths of a meter. While disputing the problem, many of us arrived at the same solution: that some of the very phenomena that impose size limits on ordinary circuits could be exploited in a new generation of vastly more efficient devices. The functional bases for these devices are quantum-mechanical effects that carry semiconductor technology into a realm of physics where subatomic particles behave like waves and pass through formerly impenetrable barriers. With the so-called quantum semiconductor device, I believe it will be possible to put the circuitry of a supercomputer on a single chip.

The structures for quantum devices have already been made using the same materials as today's chips: doped silicon, doped and undoped gallium arsenide, and aluminum gallium arsenide. Because they can be about 100 times smaller than the devices in present-day integrated circuits, however, designing and fabricating a viable device presents a formidable challenge. Manufacturing processes will have to become considerably more sophisticated, and new strategies for interconnection and architecture will have to be devised to cope with the special problems of size reduction.

As daunting as they are, these adjustments are worth making in order to realize the ten-thousandfold reduction in cost per function that quantum devices could bring about. They are also minor compared with the difficulty of introducing new materials for which no relevant process technology exists. And the progress that has been made at Texas Instruments as well as at other industry, government and academic laboratories around the world suggests that quantum devices just might embody the revolution the electronics industry awaits.

The motive for shrinking the components of integrated circuits is minimizing the cost and time needed to perform each circuit function. Most functions are carried out by transistors, which act essentially as switches. In a transistor the speed and precision with which switching can be controlled, as well as the power needed to produce the switching, has everything to do with the time and cost per function attained by the device. Because of its size, a transistor switch that operates on the principles of quantum mechanics would be faster and would consume less power than a conventional transistor; because of effects peculiar to quantum phenomena, it could also afford a greater degree of control.

These attributes can best be appreciated in comparison with the performance of conventional transistors. The most commonly used transistors today are field-effect transistors, or FET's. They are made from semiconducting materials doped with elements that provide carriers for electric charge. The charge carriers can be either electrons, which bear a negative charge, or positive "holes"; a semiconductor that has electrons as charge carriers is said to be negatively doped (*n*-doped) and a semiconductor that conveys charge by the movement of holes is said to be positively doped (*p*-doped). Silicon has been the traditional stuff of integrated circuits, but gallium arsenide (GaAs) transistors have been constructed that are faster.

The two types of transistor have slightly different configurations [*see illustration on page 98*]. In a typical silicon FET a region of *n*-doped silicon called the source is separated from another *n*-doped region, the drain, by a *p*-doped channel. On top of the channel there is a metal electrode called the gate, which is kept from coming in direct contact with the *p*-doped silicon by a layer of insulating silicon oxide. (This metal-oxide-semiconductor arrangement is the derivation for the common acronyms *n*-MOS, *p*-MOS and MOSFET.) A positive voltage is applied to the drain; when a weaker positive potential is also applied to the gate, electrons cluster in the silicon channel under the gate and create a bridge of negative charge carriers between the two *n*-doped regions. This bridge, called the inversion layer, enables electrons in the source to flow toward the positive voltage on the drain. The current flow can be interrupted by removing the potential on the gate, thereby dispersing the electrons in the inversion layer.

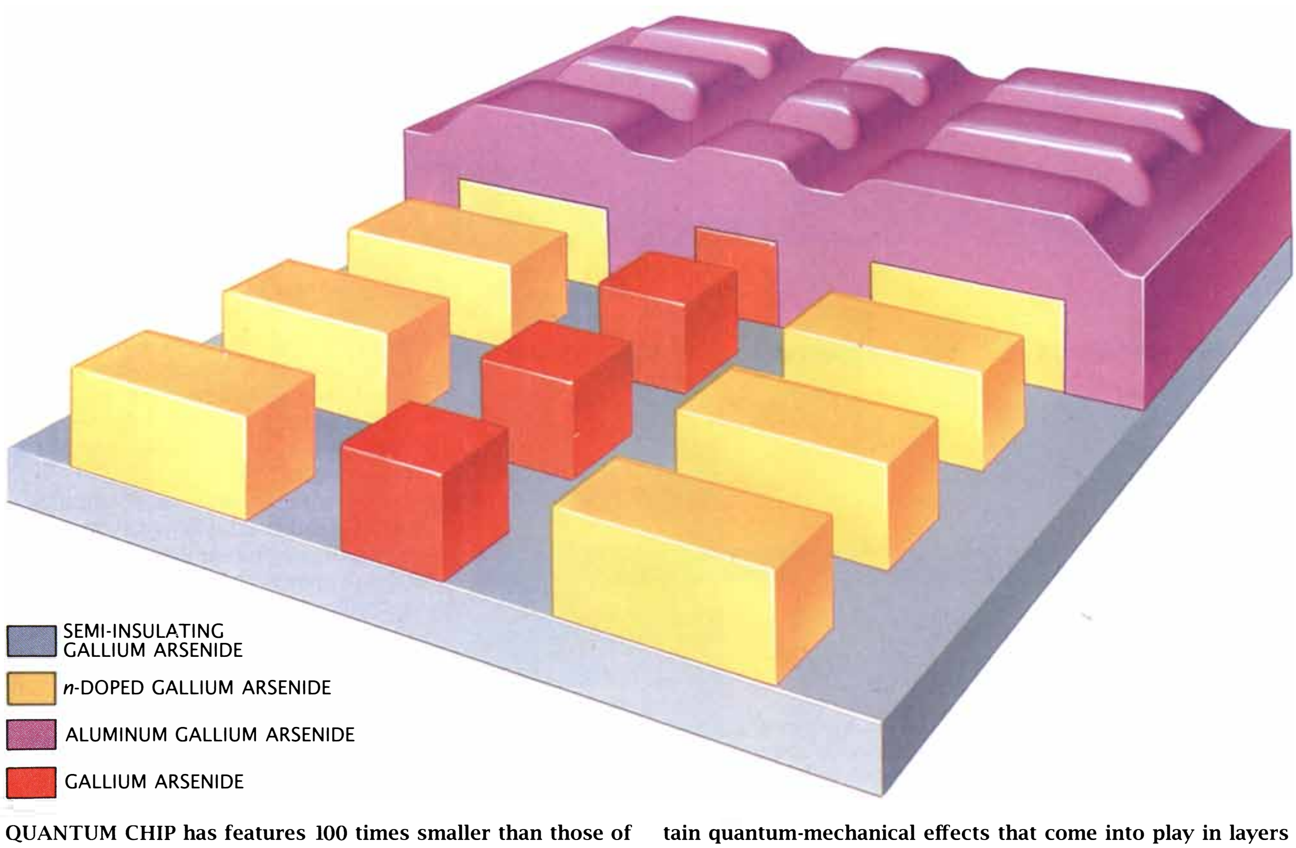
A gallium arsenide transistor also has a gate electrode and terminals that serve as source and drain, but the *n*-doped part of the substrate is not localized [see "Gallium Arsenide Transistors," by William R. Frensley; SCIENTIFIC AMERICAN, August, 1987]. When a positive potential is applied to the gate and the drain, current flows freely from the source; if the gate is given a negative voltage, it repels electrons from the area under it, blocking the path of conduction.

Both transistors are three-terminal devices, and in both of them adjusting the voltage on the gate is the most sensitive means of switching the device. Hence the transistors can be switched "on" and "off" by changing the voltage on the gate. These devices work well at present scales, but with downscaling the distinction between switching states becomes blurred. At smaller scales current leakage prevents a transistor from being truly "off"; it also causes unnecessary consumption of power. Impurities or defects in the semiconductor crystal can scatter electrons, slowing both conduction and switching. For all its usefulness, the modern FET has a problem: the smaller it gets, the worse it switches.

Because the way in which quantum semiconductor devices would function is qualitatively different, quantum devices promise more precise and efficient control of switching in a size regime ordinary transistors could never approach. This difference is manifested in the current-voltage characteristics. In particular, some quantum semiconductor devices exhibit negative differential resistance: that is, there is a voltage range in which the current *decreases* as the applied voltage is increased. On a graph of current versus voltage, this property translates into a current peak and a current valley [*see top illustration on page 100*]. The presence of negative differential resistance is often the only indication a physicist has that quantum effects are operative in an experimental device.

The elusive phenomenon at the heart of quantum effects is the wave nature of electrons. Quantum theory predicts that an electron will exhibit wavelike behavior whenever the region within which it is confined, or the barriers erected to contain it, has dimensions approaching the electron's wavelength. Hence at least one dimension of the features in a quantum device is comparable to the wavelength of an electron. In gallium arsenide at room temperature that wavelength measures just 200 angstrom units (20 billionths of a meter).

The barriers that can contain electrons are barriers of energy rather than physical barriers. All electrons possess a finite amount of energy and are said to occupy energy levels; the levels available are characteristic of a given material. A group of closely spaced levels is called a band. In most solids the energy levels in each band are so closely spaced that they are essentially continuous, and so an electron can change levels with only



QUANTUM CHIP has features 100 times smaller than those of standard chip components. Current flows from one negatively doped (*n*-doped) gallium arsenide block to another through a layer of aluminum gallium arsenide, a gallium arsenide cube and then another aluminum gallium arsenide layer. Because of certain quantum-mechanical effects that come into play in layers of this size, the current a quantum device conducts is extremely sensitive to differences in applied voltage and can therefore be closely controlled. This is an idealized model; a functioning device of such sophistication has not yet been fabricated.

an infinitesimal boost of energy.

The relative positions of energy bands determine whether electricity can be conducted across two different materials. For an electron to pass from one material to another with no change of energy, the bands of the two materials must overlap. Specifically, in the first material the average level occupied by electrons—called the Fermi level—must coincide with an energy band of the second material. If the energy band of the second material occurs at a much higher energy level than the Fermi level of the first, the second material acts as a barrier to electron movement.

For example, under ordinary circumstances aluminum gallium arsenide (AlGaAs) presents a barrier to the electrons in *n*-doped gallium arsenide. An electron cannot pass from the doped GaAs to AlGaAs because the conduction band of AlGaAs is at a much higher energy level than the Fermi level of the GaAs. Yet if the physical dimensions of the barrier are altered in such a way that the wave nature of electrons comes into play, an electron will "tunnel" through the AlGaAs that was once an obstacle. Hence when a layer of AlGaAs thinner than 200 angstroms is sandwiched between two pieces of doped GaAs, the electrons tunnel through it to the GaAs on the other side. This tunneling is one kind of quantum effect.

When barriers confine electrons within a space comparable to an electron wavelength, the electrons are subject to two other, interrelated quantum effects: size quantization and resonance. Size quantization causes the continuum of energy levels that usually exists in the conduction band of a solid to become articulated into discrete energy quanta, or states. It is most aptly described by a density-of-states graph, which shows the number of allowed discrete states of an electron within a fixed energy range [*see illustration on opposite page*].

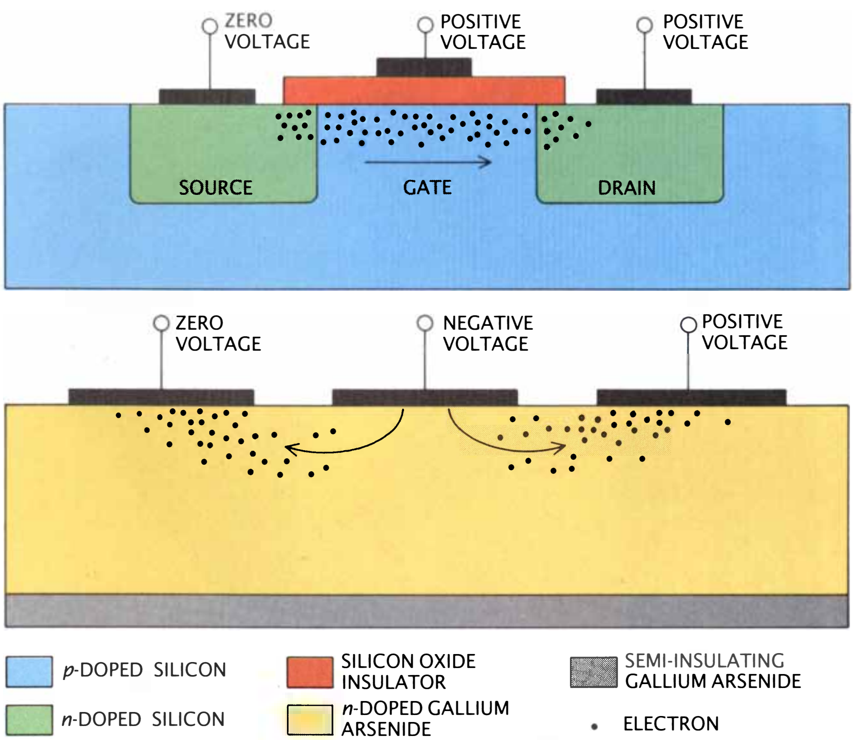
When, for example, a sliver of undoped gallium arsenide is enclosed within AlGaAs barriers, the density-of-states graph for the GaAs looks more like a ladder than a hill. The degree of quantization depends on the degree of confinement. When the electrons in GaAs are restricted in all three dimensions (a "quantum dot"), their energy levels are completely discontinuous; in one-dimensional restriction (a "quantum well") and two-dimensional restriction (a "quantum wire") the levels are still somewhat continuous.

Resonance, the other consequence of quantum confinement, occurs only when some degree of size quantization has been achieved. Electron waves that enter, say, a quantum well are reflected off the far wall of the well; the waves essentially bounce back and forth within the quantum chamber [*see bottom illustration on page 100*]. In doing so they increase the tunneling current substantially—they resonate. Both size quantization and resonance result from the constructive interference of the forward and backward waves. It is difficult to separate the current enhancement that can be attributed to resonance from the enhancement that results from the increased density of states at a given energy level.

As it happens, that distinction is not crucial for transistor operation. What does matter is that in a quantum-effect device two slightly different voltages can evoke profoundly different responses. The differences should be most pronounced in the most confined structure, the quantum dot, because it exhibits the highest degree of quantization. At voltages where tunneling occurs, current is enhanced by the high density of states and by resonance effects to create a peak; at other voltages, the total absence of states at energies intermediate between quantum levels ensures that very little tunneling occurs, and a valley in the current is thus created.

To visualize how these quantum effects could come in handy in a transistor, imagine two slabs of *n*-doped GaAs separated by an AlGaAs-GaAs quantum dot. Electrons trying to pass from one slab of doped GaAs to the other must tunnel through a layer of AlGaAs into the quantum dot and then through another stretch of AlGaAs. They cannot enter the quantum dot, however, unless one of the energy levels in the dot is on a par with the Fermi level of the doped gallium arsenide from which the electrons are emitted.

The Fermi level of the GaAs "emit-



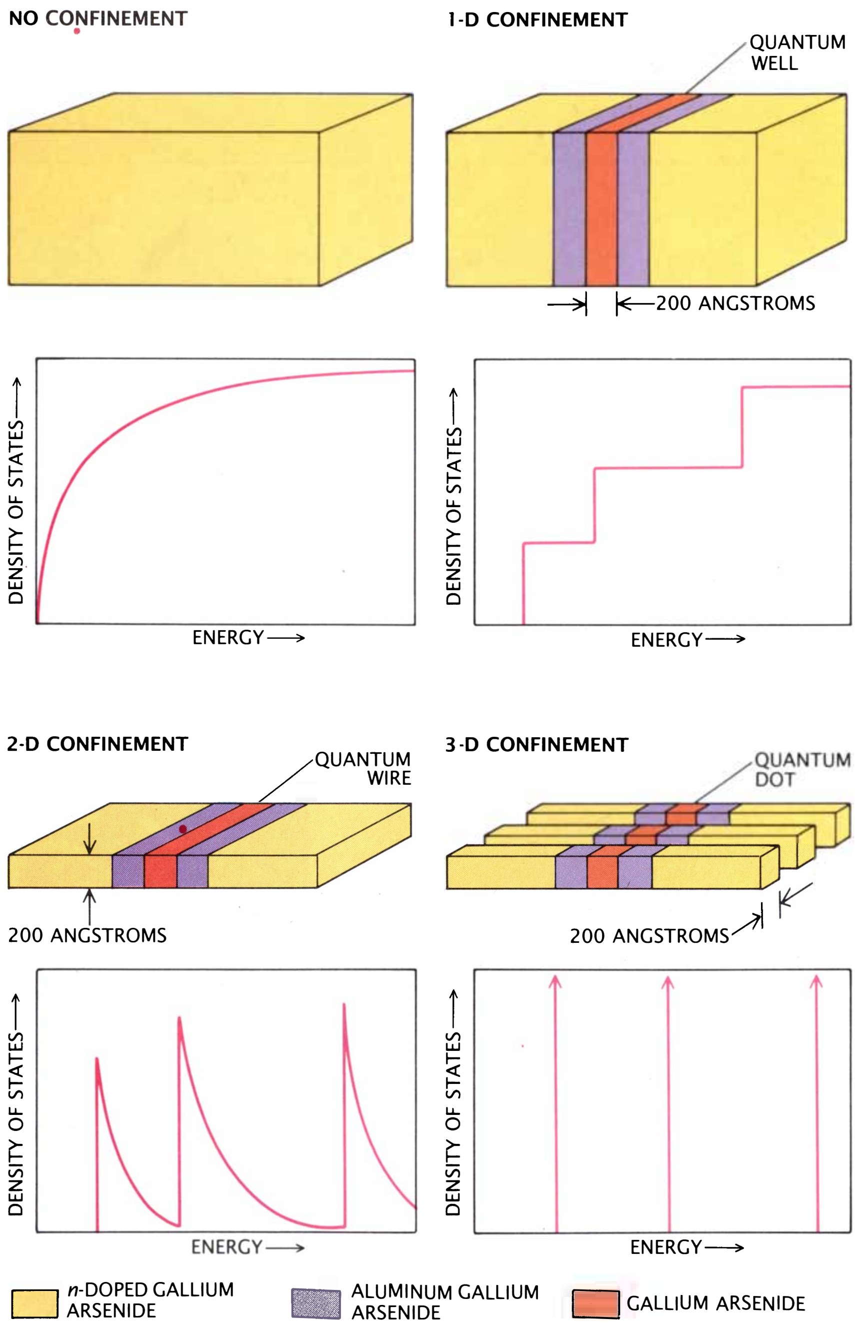
FIELD-EFFECT TRANSISTORS make up the majority of integrated-circuit components today and operate according to the laws of classical physics. In the silicon transistor (*top*) electrons flow between the source and the positively biased drain when a positive voltage is applied to the gate. The gate potential creates a kind of electron bridge between two *n*-doped regions; without it the electrons in the positively doped (*p*-doped) silicon channel disperse and the channel becomes impassable. In contrast, the gallium arsenide transistor (*bottom*) conducts when there is no potential on the gate, but the application of a negative voltage disrupts the flow of electrons from source to drain.

ter" can be raised with respect to the rest of the structure by applying a positive voltage to the doped GaAs on the opposite side of the dot—the "collector." At some voltage the Fermi level of the emitter will attain the same energy as one of the energy levels in the dot, and electrons will move into and resonate within the dot. There is a single voltage at which this occurs; the conduction that takes place at other voltages owing to thermal excitation and to leakage and scattering is negligible. Here, then, is a way to control precisely the switching of a semiconductor device.

Although from this description the incorporation of a quantum-dot structure in a so-called quantum coupled device may seem like a remote possibility, actually the realization of such a device may not be too many years away. Indeed, the exploitation of quantum effects in semiconductor devices dates from the 1950's. The Esaki tunnel diode, named for its inventor, Leo Esaki, now at the IBM Corporation's Thomas J. Watson Research Center in Yorktown Heights, N.Y., was the first quantum semiconductor device. In this diode *n*- and *p*-doped semiconductors were juxtaposed to create a layer having no charge carriers at all. When the doping was extremely high, the so-called depletion layer became thin enough for electrons to tunnel through. The diode never had widespread appeal, however, because the three-terminal devices that were coming of age at the time proved to be more efficient and convenient.

In the 1960's workers at the Watson Research Center verified that quantum confinement in one dimension takes place in the inversion layer of silicon MOSFET's. Because the influence of quantum effects on device characteristics was so small, that discovery had little impact on transistor development. Subsequent work by Nick Holonyak, Jr., of the University of Illinois at Urbana-Champaign made quantum wells standard ingredients in lasers. In the 1970's Esaki, along with Leroy L. Chang of the Watson Research Center and Raphael Tsu, now at North Carolina Agricultural and Technical State University, carried out the earliest experiments on resonant tunneling through wells. Quantum effects were not deliberately induced in transistors until recently, in the so-called modulation-doped FET's. The quantum wells in these devices, however, serve only to improve the mobility of electrons that otherwise act as they do in conventional transistors.

While seemingly tangential, these developments helped to advance the techniques required to make quantum semiconductor devices, so that the technology for constructing experimental structures was at hand when interest in the field finally blossomed. For the past four years the realization of zero-dimensional quantum structures has been the focus of attention for workers around the



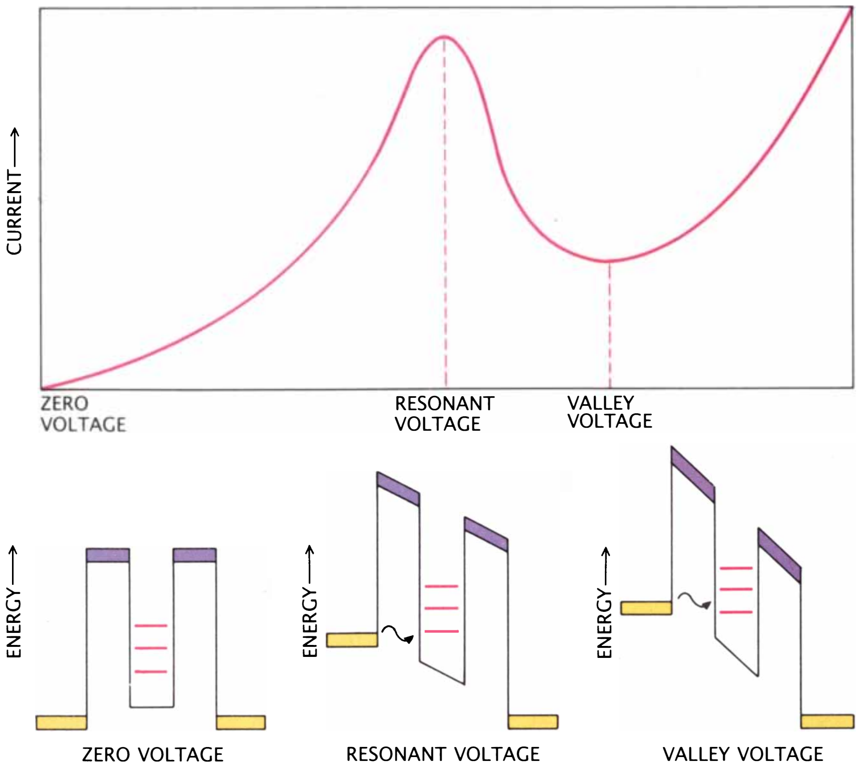
QUANTUM CONFINEMENT alters the energy states an electron can occupy in a conducting material. For example, in an ordinary piece of *n*-doped gallium arsenide (*top left*) electrons move freely among a continuum of states, but when barriers of aluminum gallium arsenide are erected in one dimension around a gallium arsenide quantum well the width of an electron wavelength (200 angstrom units), the density of energy states in the well becomes quantized, or discontinuous (*top right*). Restricting the height of the well gives rise to a quantum wire (*bottom left*). The degree of quantization depends on the degree of confinement; true quantization is realized only when gallium arsenide is confined in three dimensions in the quantum-dot structure (*bottom right*).

world. At the AT&T Bell Laboratories, IBM, the Massachusetts Institute of Technology, the University of Cambridge and the Philips Research Laboratories, size quantization in quantum wires has been demonstrated in silicon and gallium arsenide devices alike; quantum dots have been fabricated at AT&T, Bell Communications Research, the Hughes Research Laboratories and the University of Glasgow as well as at Texas Instruments, where the clearest indication of size quantization in dots has been found.

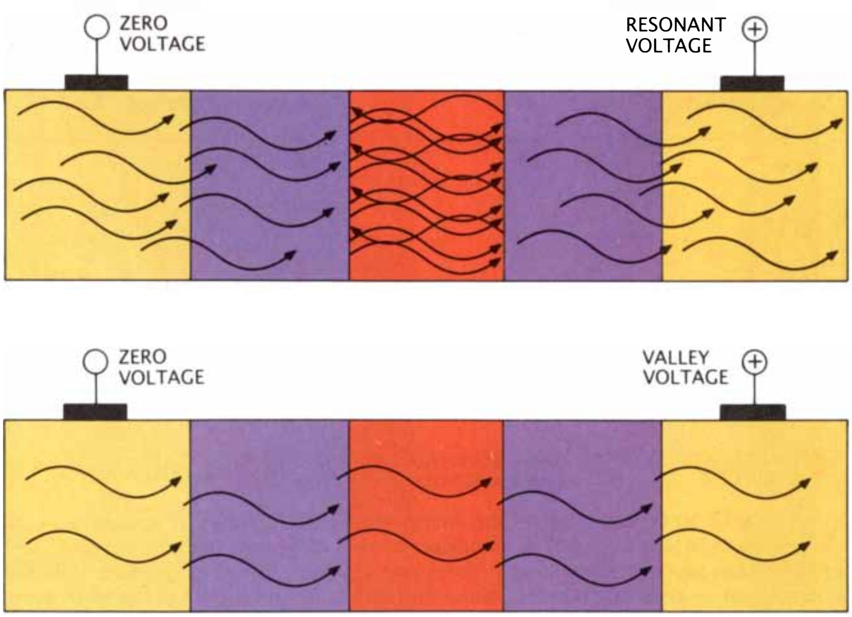
An operational semiconductor device has yet to be constructed from a quantum-dot structure, but a prototype should be available within one or two years. One of the objectives of current research is the conversion of quantum devices, which are most readily constructed as diodes, to three-terminal devices with a third contact directly modulating the potential of the quantum structure. Such a connection would yield the most compact device, and one that would most closely approach the maximum switching speed afforded by tunneling. Devising a technology to manufacture reliable and nondestructive contacts for such thin layers, however, will require a great deal of ingenuity.

By placing quantum dots in close proximity, electrons might also be enabled to tunnel from one dot to another—from one quantized state to another. This arrangement would provide the ultimate in circuit control because the energy states the electrons could assume at both the point of departure and the point of arrival would be strictly dictated. Again, the challenge lies in the formidable task of fabricating structures hundreds of times smaller than any of the features in current semiconductor products. And that degree of downscaling will in turn bring about problems with interconnections and architecture that industry will have to solve before quantum semiconductor devices can be regarded as marketable entities.

The commitment of so many research teams to a problematic technology attests to the tremendous potential of these devices and to the faith that they will take the lead in the next semiconductor revolution. The costs and risks involved must be borne in order to revitalize a rapidly maturing electronics industry; the results can only benefit a society that has learned to depend on integrated circuits in many ways.



CURRENT-VOLTAGE CHARACTERISTICS of a quantum-well device reflect the quantization of energy states in the gallium arsenide well. Such devices show a range of voltage in which the current conducted by the device decreases as the voltage applied to one of the *n*-doped gallium arsenide contacts increases. This happens because at one voltage (the resonant voltage) the average energy of electrons in the *n*-doped substance (*top of yellow band*) shifts to a level that coincides with one of the quantum states (*red*) in the well, but beyond that voltage the energy band of the doped gallium arsenide occurs between quantum states. Hence at the resonant voltage an electron (*arrow*) can tunnel through the aluminum gallium arsenide energy barrier (*purple*) into the well, whereas at the valley voltage there are no states for the electron to tunnel into.



TUNNELING ELECTRONS (*arrows*) resonate in a gallium arsenide quantum well (*red*) when a positive bias called the resonant voltage is applied to one of the contacts (*top*). The electron waves bounce back and forth inside the well, enhancing the current to give rise to the peak on the graph at the top of this page. At the valley voltage (*bottom*) little tunneling or resonance takes place, consequently the current dips dramatically.

No other system of keeping up can compare with ours.

YOUR SYSTEM: a time-consuming, futile struggle to keep up with the information explosion.

The classic texts are convenient references—but the information they contain is obsolete before publication.
Like many physicians, you probably rely on the texts you first used in medical school. But even using the most recent editions, you find material that no longer reflects current clinical thinking—and they lack the latest information on such topics as herpes, oncogenes, AIDS, and photon imaging.

Reading stacks of journals alerts you to recent developments—but can't give you quick answers on patient management.
Struggling through the hundreds of journal pages published each month—even on only the really significant advances in the field—is arduous and memory-taxing. And it's a task that costs physicians valuable time—their most precious resource.

Review courses cover clinical advances—but, months later, do you recall the details of a new procedure or unfamiliar drug?
Seminars can also be costly and make you lose valuable time away from your practice—expenses that may amount to several thousand dollars. And, the speaker's skill often determines how much you learn.

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

A comprehensive, 2,300-page text in two loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.
This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.

Each month, six to nine replacement chapters to update your text *plus* new references, a four- to five-page news bulletin, and a completely new index.
You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.
Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the floppy-disk version for your IBM® PC/PC Jr., Macintosh,™ or Apple® IIe or II+ (or compatible).

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

Name ______ Specialty ______
Address ______
City ______ State ______ Zip Code ______

Or, call toll-free: 1-800-345-8112 (in Penn. 1-800-662-2444). 7R

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

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

Your subscription includes:

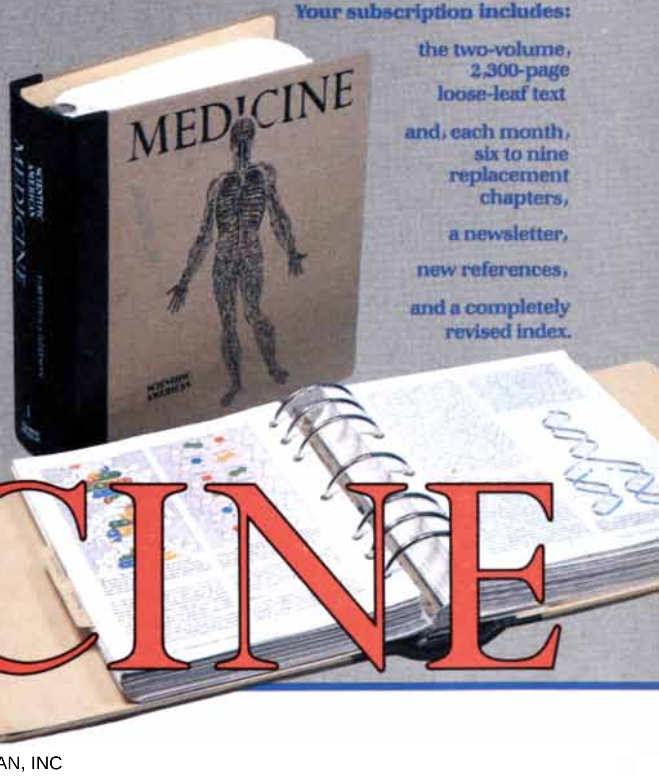
the two-volume, 2,300-page loose-leaf text

and, each month, six to nine replacement chapters,

a newsletter,

new references,

and a completely revised index.



MEDICINE

The Roman Port of Cosa

At its height in 100 B.C. it included an amphora factory, a winery, a fishery and a water-distribution system as well as a harbor with concrete piers—all controlled by one noble, entrepreneurial family

by Anna Marguerite McCann

In ancient times—when the sea was the chief medium for international trade and communication—ports were even more important places than they are today. They were centers for the importation of goods, ideas and people as well as for the accumulation of wealth through exports. Yet in spite of their significance, archaeologists have paid little attention to ports. Since the invention of the self-contained underwater breathing apparatus (SCUBA) in 1942, shipwrecks have received the lion's share of attention from marine archaeologists. Hence it is significant when an ancient port is studied with the full complement of methods of modern archaeology. That is exactly what has happened at the Roman port of Cosa, where a team of archaeologists, geologists, engineers and others has been working for 22 years under my direction.

During late Republican times, in the second and first centuries B.C., Cosa was a port of first importance for the western Mediterranean. The foundation of its wealth was a remarkable commercial complex, including a harbor with concrete piers, a lighthouse, a fishery, factories for making amphoras, wine and fish products along with machinery for lifting fresh water from a spring. Much of this array was probably controlled by a single family known as the Sestii, who seem to have served as patrons of technology much as the Medici served as patrons of the arts during the Renaissance. Although the Sestii had been known as important Roman politicians, their role in international trade was a surprise.

Other surprises too have come from the work at the port of Cosa. The origins of Rome's international trade have been pushed back two centuries. The earliest dated example of hydraulic concrete made with a type of mortar called pozzolana has been found, as well as the remains of the earliest Roman lighthouse. In the lagoon behind the port the first Roman commercial fishery was found. We also uncovered the earliest archaeological evidence for water-lifting machinery of a type still employed in the Near East. As the result of our excavations something rare has taken place: a major new ancient port site has been added to the map.

The Romans founded the Latin colony of Cosa in 273 B.C., soon after conquering the nearby Etruscan cities of Vulci and Volsinii. About 140 kilometers (an easy day's sail) north of Rome, Cosa offered the only high promontory with a protected anchorage for many hundreds of kilometers. In the days of coastal navigation, when lights to steer by and coastal lookouts were essential, the promontory of Cosa would have been a key point of reference for ships moving westward around the Argentario peninsula to the northwest of Cosa and up the coast of Gaul (modern France) and Spain. Control of Cosa's headland was vital to any system of naval defense or trade not only in the immediate region but also in the entire western Mediterranean.

The first half of the third century B.C., which spanned the founding of the colony of Cosa, was a crucial time for Rome. The conquest of the Etruscans in the north and the Greeks in the south enabled the Romans to control the entire Italian peninsula. Having done so, they pushed outward and engaged the great maritime power of Carthage in the three Punic Wars, from which Rome emerged in the middle of the second century B.C. as the chief naval and commercial force in the Mediterranean. The success of this expansion was due largely to the systematic establishment of coastal colonies along the Tyrrhenian coast, and the defensive role of those colonies has long been recognized. Our excavations at the port of Cosa, however, suggest that commercial rivalry with Carthage for western Mediterranean trade was also a driving force in the establishment of the coastal colonies.

The importance of commerce yields a new perspective on the colony of Cosa, which stretched for more than 30 kilometers along a part of the Tuscan coastline known today as the Maremma. It should be noted that the port was not the only component of the colony. Above and behind the port lay the fortified hill town of Cosa, partially excavated by the American Academy of Rome in a long dig that began in 1948. The town was clearly the religious and civic center of the colony, and it was initially thought to be the key to the colonial life of the region. No one suspected that a major Mediterranean port lay below, but it is now evident that the Portus Cosanus was the economic center of the colony and the primary source of its wealth.

One reason for the port's success was that it possessed the only source of fresh water in the area. Whereas the town of Cosa and the nearby port of Hercules (modern Port'Ercole) on the Argentario peninsula had to rely on rainwater, the port had numerous springs that provided abundant potable water for ships, sailors and several industries: an amphora factory, a winery, a fishery and a saltery (an installation for salting fish). In the course of our excavations the largest of the springs was released, gushing forth at a rate of some 1,500 liters per minute. Judging from the number of broken drinking vessels found near it, this spring must have served the inhabitants of the surrounding area in ancient times.

What was the port of Cosa like at the peak of its commercial activity in about 100 B.C.? The harbor, which was the central element of the complex, had been modified considerably from its natural state by Roman engineers. To augment the protection offered by the promontory, a breakwater of limestone rocks quarried from the cliffs overlooking the harbor had been thrown out. Some 110 meters long and 70 meters wide, the rocky platform has sunk into the sandy harbor floor, but in Roman times it must have protruded from the water. It was designed to break the incoming waves while allowing free circulation of offshore currents through its loosely packed rocks, thereby preventing silting, which is a major threat to harbors. John D. Lewis of the U.S. Army, the excavating engineer, has concluded that this design was the work of experienced engineers who were quite familiar with local wave and current conditions.

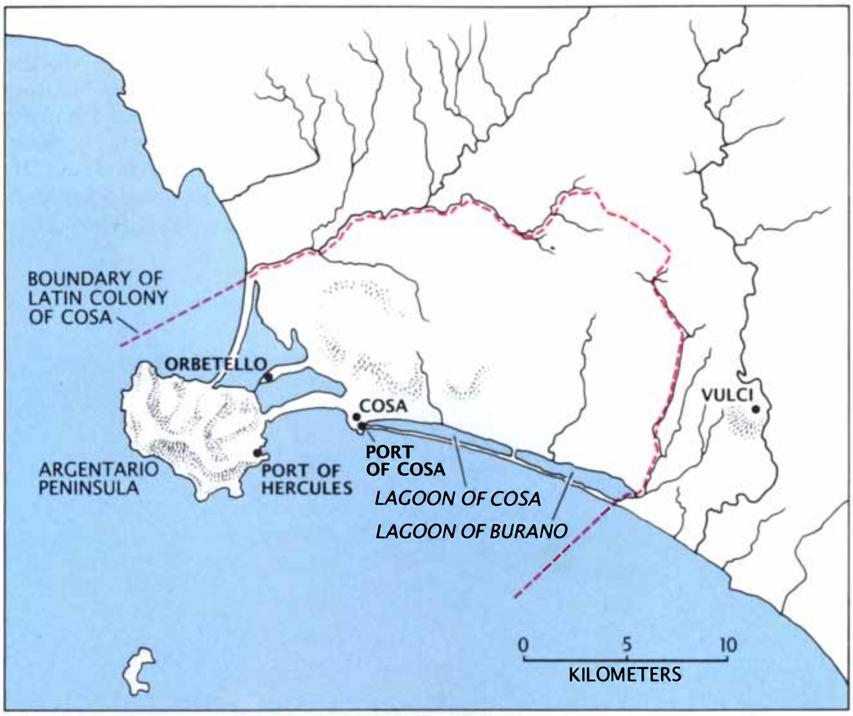
Extending from the shore was a row of three concrete piers that are still visible above the water at Cosa; another two piers, now submerged, rested on top of the breakwater. The piers provided facilities for mooring and unloading as well as protection for a smaller, inner basin. They were made of pozzolana mortar combined with pieces of limestone, tuff (a volcanic rock) and amphora fragments. (Their composition, along with that of Cosa's other concrete structures, was analyzed by Elaine K. Gazda of the University of Michigan.) Pozzolana is a powdery volcanic ash; when mixed with lime and water, it forms a tenacious binding material that sets and endures in salt water or fresh. The discovery of pozzolana by the Romans at ancient Puteoli, near Naples, in the third or early second century B.C. enabled them to become the greatest harbor builders of the ancient Mediterranean. The piers at Cosa are the first known example of this technology.

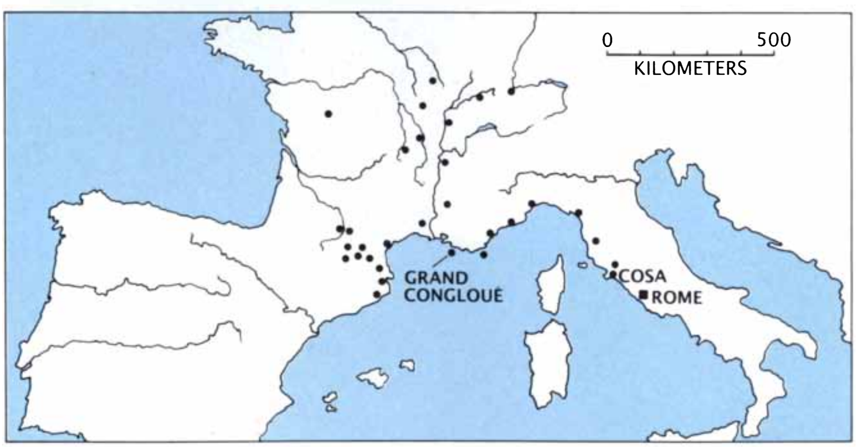
The piers were reached by a single entrance channel just off the eastern tip of the main breakwater. The channel was about six meters deep, ample clearance for a Roman merchantman 20 meters long. Such ships could have anchored in the protected basin, whereas smaller craft were moored to the breakwater or the piers or hauled up on shore. Floating docks may have extended from the breakwater or the beach. Loading and unloading was carried out by barges and small boats working from shore, and there may have been

HARBOR OF COSA as it appears today is dominated by a limestone promontory. The promontory, which sheltered the harbor and provided a navigational landmark, was one reason the Romans established a colony at Cosa in 273 B.C. Extending in a row from the shore are the remains of three concrete piers built by the Roman port engineers as they modified the natural harbor.

cranes to unload ships tied up at the piers. Small boats anchored in shallow water were loaded or unloaded by men wading out from the beach; the men would have carried amphoras on their shoulders.

An element often taken for granted in a harbor is the lighthouse, but when Cosa was at its peak, the lighthouse was a relatively recent innovation. Invented in Alexandria in the third century B.C., it was made a standard part of harbor works by the Romans. There is no evidence for a lighthouse during Cosa's earliest years, but we have found significant evidence for one during the peak commercial period. A model lighthouse made of terra-cotta, found at Vulci and dated to the late second century or the first half of the first century B.C., has been identified by Paola Zancani Montuoro of Naples as the Cosa light. She calculates that the full-size light may have been 30 meters tall with a square base five meters on a side. It seems likely that the lighthouse sat on Pier 5, which was built on the eastern tip of the breakwater and is now submerged.





PORT OF COSA was the economic centerpiece of a Roman colony extending for 30 kilometers along the sandy Tyrrhenian (west) coast of Italy (*top*). The colony included the small fortified hill town of Cosa, which may have had 1,000 inhabitants, and the nearby Port of Hercules. Amphoras bearing the marks of the Sestius family, which dominated Cosa for more than two centuries, have been found at some 30 sites in the western Mediterranean (*circles on bottom map*). Most of those sites are in Gaul (modern France).

This harbor complex, too extensive for private use, was doubtless intended for commercial purposes. No inscriptions have been found to tell us who built it, controlled it and grew rich from it. The profusion of amphora fragments uncovered in our excavations, however, provides some information to fill the gap. Roman amphoras were frequently stamped with the mark of the manufacturer, and so the study of amphora fragments offers much evidence about who dominated trade at Cosa. Some 1,000 amphora fragments were catalogued, and Elizabeth Lyding Will of the University of Massachusetts at Amherst has carefully studied them. Her work suggests how tight the hold of the Sestii was on Cosa during its commercial zenith: 86 percent of the amphora stamps found at the port are Sestius stamps.

The vessels from which these fragments came were the chief shipping containers of classical antiquity. The amphora shape evolved from the need to store and ship bulk quantities of liquids, fruits and other foods in a vessel small enough to be handled by one man—a necessity for loading or unloading without mechanical equipment. It seems probable that the Sestii had a factory for making these critical containers at Cosa. Mineralogical analysis by Raffaello Trigila of the University of Rome and Jelle Z. de Boer of Wesleyan University of the coarse red clay of the Sestius amphoras shows that the clay came from Cosa. Although the remains of the amphora factory have not been found, it probably stood on the sand behind the harbor.

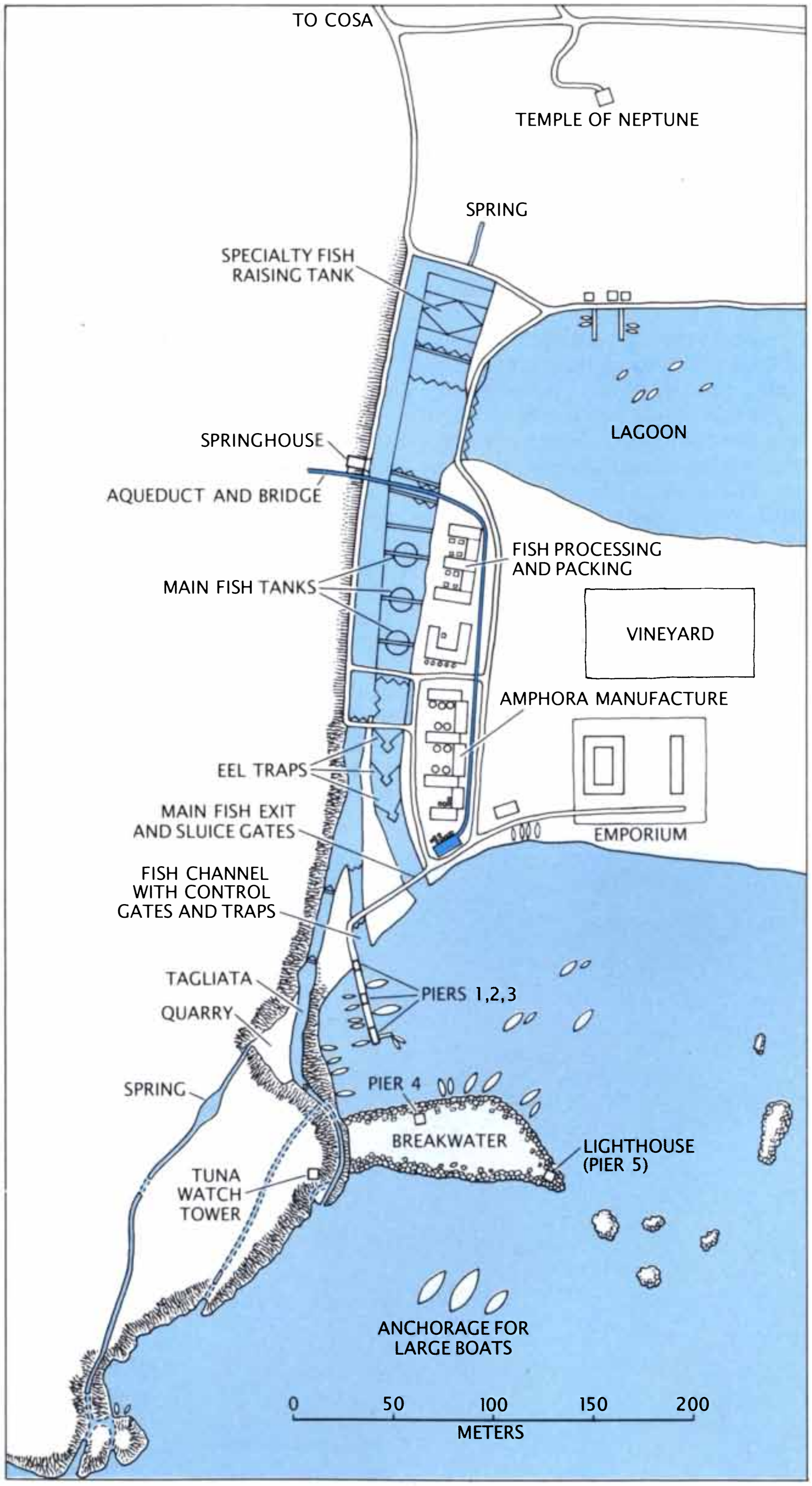
Cosa, with its amphora factory, was undoubtedly the home port of the Sestii. From there Sestius products were widely shipped: amphora fragments bearing Sestius trademarks have been found at more than 30 other sites in the western Mediterranean as well as in the Agora at Athens. Initially the main export was wine. Painted inscriptions on amphoras indicate that the primary Sestius product was wine. The Roman statesman Cicero writes that, in addition to ships, the Sestii owned a villa at Cosa, which may well have included a vineyard. It was probably there that they made the table wine that was exported throughout the western

Mediterranean, particularly to Gaul. On the return journey the merchant ships may have brought back metals and slaves to work in the Sestius industries.

Among the best-known of the sites where Sestius amphora stamps have been found is Grand Congloué, a rocky island off the French coast near Marseilles where two Roman merchant ships were wrecked. The uppermost wreck, excavated by Jacques-Yves Cousteau, went down in about 100 B.C.—just at the peak of Cosa's commercial eminence. Its cargo consisted of 1,200 Sestius wine amphoras, a haul that could only have come from a factory such as the one we think existed at Cosa. Although the ill-fated ship has been known since 1952, the port from which it put to sea could only be guessed at; now it seems clear that the merchantman put to sea at the Sestius home port.

The amphoras from the Grand Congloué wreck show that to capture a mass market the Sestii developed techniques of mass production. Most of the 1,200 amphoras that were found on the wreck are of a single type, which has been designated Will Type 4a. About a meter tall, it held 26 liters of wine. The Type 4a shape, with its tall neck, long vertical handles and narrow belly, was developed to provide increased capacity and greater ease of handling. By the late second century B.C. it had become dominant: along with the closely related Type 4b it makes up 70 percent of the material found at the port. Indeed, the Will Type 4 amphora was the most popular of all the Roman wine jars of the Republican period.

If mass-production methods and monopolies have a modern ring, so does another aspect of the Sestius commerce at Cosa: diversification. With the profits from their export trade in wine, these shrewd merchants apparently diversified into the manufacture of fish products on an industrial scale. The ancient geographer Strabo gives evidence of a fishing industry on Cosa's shores in the late first century B.C. Strabo writes that a watch for tuna was kept on the promontory, one of only three such watches he identifies on the long Tyrrhenian coastline. The annual eastward migration of the big fish, which lasts from May to October, was eagerly awaited, and where there was a tuna watch there must also have been a means of processing the fish once they had been dragged up on the beach.



PORT AT ITS HEIGHT was an industrial complex for the manufacture, packaging and shipping of two chief products: wine and processed fish. The fish, from a lagoon fishery behind the port, was dried, pickled or salted; wine was made at a winery near the harbor. Both commodities were shipped in amphoras manufactured at Cosa. The harbor included five piers and a breakwater surmounted by a lighthouse 30 meters tall. The harbor was linked to the lagoon by artificial and natural channels. This illustration and the one on page 108 are based on drawings by J. F. Warren of I. M. Pei Partners.

Exactly how the catch was processed is not yet known; the remains of the fish factory or saltery have not been found. Yet our discovery of a fish farm in the adjacent lagoon makes it clear that Cosa was indeed an industrial fishery in the first century B.C. We found concrete fish tanks, more than 100 meters long, covering about a hectare (2.47 acres) of the west end of the adjacent lagoon. The lagoon has largely silted up, but in antiquity it extended 25 kilometers or more to the east. Joanne Bourgeois of the University of Washington estimates that in Roman times the lagoon was about 800 meters wide at its widest point and about five meters deep. Such coastal barrier lagoons are found elsewhere in Italy only near Venice, and even today they are one of Italy's richest natural resources; a modern lagoon fishery operates today at the town of Orbetello, near Cosa.

The shallow, brackish lagoon was rich in vegetation that supported a wide variety of marine life. With relatively little modification it could be turned into an industrial complex for raising and catching fish. The one modification that was essential was to provide a continuous circulation of water to control temperature, oxygen levels and salinity. In early colonial days the Cosans must have exploited natural channels cutting through the long sand barrier from the sea to the lagoon. Even more favorable for the development of a fishery, however, was the limestone promontory, which was split by natural fractures and dotted with freshwater springs. The fractures could be widened and improved, creating a permanent link between the sea and the lagoon.

One fracture, called the Tagliata (cut), was modified particularly for the fishery. This channel, cut in part from bedrock, winds 150 meters to join the south side of the ancient lagoon. Its seaward section was fitted with sluice gates to control the flow of water and to serve as a tank for catching fish feeding in its rocky passages. In the drier seasons, or when the tides were rising, the Tagliata carried cool seawater (and fish) into the lagoon. In the rainy fall and winter, when the lagoon was high, the Tagliata carried the overflow back to the sea. In this way the Roman engineers made it possible to raise and keep fish in the large concrete tanks, which could have yielded about 150,000 kilograms of fish per year.

What those fish may have been is suggested by the species raised and taken at the modern fishery at Orbetello. Various members of the eel family make up about 50 percent of the catch. Eels ascend in the spring and descend in the fall and early winter to spawn only once in deep water. They are caught today—as they were in antiquity—in special traps set along the boundary between the lagoon and the sea. Other species adaptable to brackish water are the gray mullet, the sea bass, the gilthead and the sole. Mullet were particularly prized by the Romans; large ones brought exorbitant prices, and a special type of fish sauce was made from their liver.

Now, some of this catch would have been eaten by the local population: the people of the port, the town and the nearby farms. Some of it would have been shipped in special boats equipped with deep wells for holding live fish. In the days before refrigeration, however, the largest part of the catch must have been

CONCRETE PIER has endured for almost 2,000 years in the harbor at Cosa. The binding element of the concrete was pozzolana, a volcanic ash. When mixed with lime and water, concrete with pozzolana mortar sets and endures in salt water or fresh water. In the lighter upper layer pozzolana was mixed with limestone and amphora fragments. In the darker lower layer the mortar was mixed with tuff, a volcanic rock that lasts in water. The holes accommodated formwork in which the concrete was molded.

pickled, salted or dried (fish processed in these ways was a staple of the ancient Roman diet). The most famous and profitable of all processed fish products, however, was the sauce known as garum. Made from the guts of fish left to ferment in the sun, garum was usually produced at a saltery as a by-product. Although sometimes blended and drunk as a liquor, garum was usually consumed as a pungent, salty condiment and in that form it was on every table: the Roman culinary authority Apicius includes more than 350 recipes calling for garum in his famous cookbook.

Not only was garum in great demand but also it was highly profitable to make. The historian Pliny the Elder writes that an amphora of a special variety of garum fetched 10 times as much as an amphora of even the choicest wine. Under those conditions it is no surprise to find that the Sestii seem to have diversified into the manufacture of garum. Will has identified two of our amphora shapes from the late second and first centuries B.C. as having been specifically for garum; it is also possible that some other jars, formerly thought to have been for wine, were actually for the expensive fish sauce. In any event, the evidence from Cosa is startling, because it had been thought garum was not exported from Italy until the first century A.D., and even then not in quantity.

Although it is logical to assume that the Sestii were indeed behind the elaborate fishery complex, it must be kept in mind that we have no explicitly written sources to go by. The archaeological material, however, is overwhelmingly in favor of the Sestius influence, particularly in the period of the port's prime. The Sestii employed a symbol of the Cosa lighthouse in one of their amphora stamps. Most of the other stamps include symbols of Neptune such as the trident, a ship's wheel, an anchor, a fishhook, fish spines and a palm branch. Neptune, the god of the sea and of fresh water, was also the deity of the port: his temple stood on a hill behind the lagoon. Perhaps when Cosa became a municipium (a self-governing town of Roman citizens) in 90 B.C., the Sestii obtained from the town a franchise for the fishing and water rights in the lagoon.

Whatever the legal arrangements for the control of the fishery may have been, one of its essential practical requirements was a supply of

SESTIUS AMPHORA is of the kind called Will Type 4a, the variety most commonly found at Cosa. About a meter tall, it held 26 liters and was generally filled with wine. Like other Sestius amphoras, Type 4a was stamped with one of the Sestius marks (*right*): "SES" and a small device—here the trident, Neptune's symbol. Neptune was not only the god of the sea and of fresh water but also the port deity. This amphora was found in the walled hill town of Cosa, above and behind the port.

fresh water. As I have mentioned, the port had several springs, but to be of use for industry, water must be collected and brought to where it is needed in daily operations. To fulfill those functions, the Cosans built a springhouse directly over the spring we released during our excavations. This unique structure, which contained equipment for lifting water from the spring to an aqueduct, served the port in its prime and was later refurbished and used again by the large villa that was then occupying the site. The most important single find of our entire excavation came from the springhouse: pieces of the wood machinery that lifted the water from the spring.

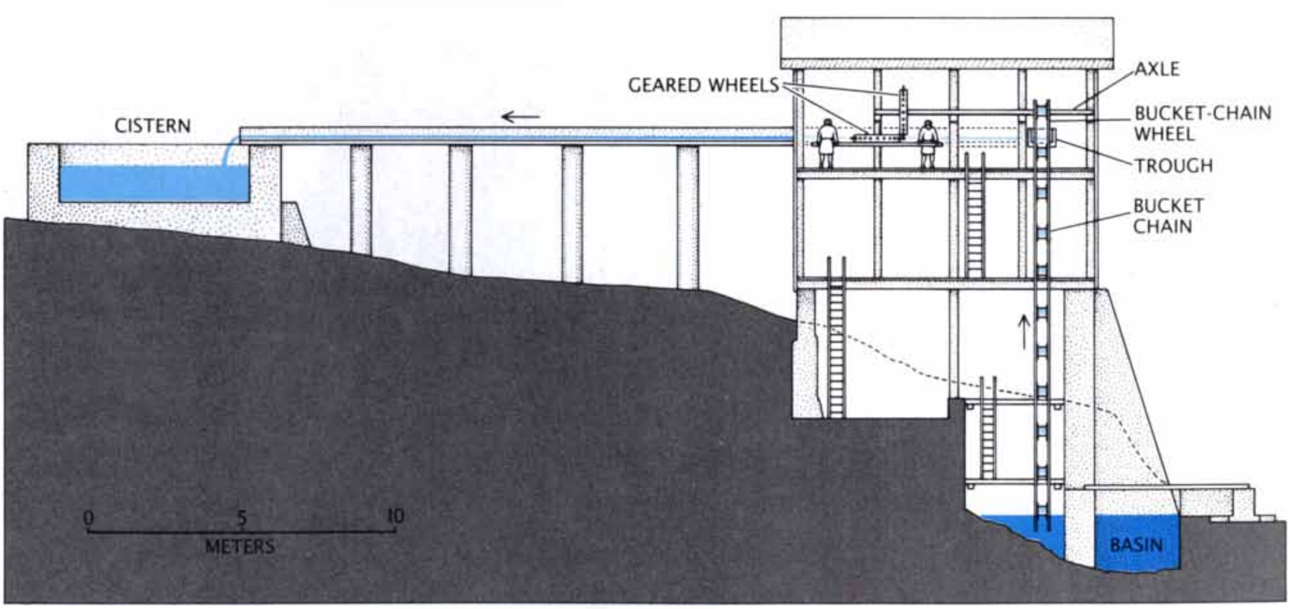
The first phase of operation of the springhouse began early in the first century B.C., when the port was at its commercial zenith. The initial design called for the mechanism to lift water five meters from the ground-level spring to a collecting tank, from which it flowed east on an aqueduct across the lagoon to the commercial area and the port, making fresh water available to industries, ships and sailors. At the end of the first century B.C. the springhouse was abandoned, to be rebuilt only after about a century, this time lifting water 13 meters up the sloping hillside to the west across an aqueduct that emptied into a large cistern.

The second springhouse complex was destroyed by fire in about A.D. 150. Fortunately, pieces of its machinery fell into the room over the spring, where they were preserved until we found them almost 2,000 years later. Because the design of the water-lifting mechanism does not seem to have changed fundamentally during the 250-year life of the springhouse, the remains help us to understand the principles of the machinery employed in the first phase of operation as well as the second one. The job of analyzing the machine was undertaken by John P. Oleson of the University of Victoria in British Columbia, who excavated the springhouse and restored parts of the water-lifting machinery.

The overall design was based on a pair of meshing wood gear wheels, one horizontal and one vertical. Turning the horizontal wheel (by means of a bar extending from its axle) drove the vertical wheel. The vertical wheel was linked by a shaft to a third wheel, also vertical. As the third wheel turned it propelled a continuous chain of buckets that lifted water from the spring and emptied it into a trough immediately under the bucket-chain wheel. Machinery of this type, which was probably developed at Alexandria in Egypt in the third century B.C., is still employed in parts of the Near East, where it is known by its Arabic name of *saqiya*. The remains from Cosa—including parts of seven buckets and a spoke from the bucket-chain wheel—are the earliest archaeological evidence for a *saqiya* mechanism.

The springhouse containing this apparatus climbed the steep hillside above the lagoon in three sections. A concrete platform extended into the lagoon itself. Attached to the platform was a basin that held 17,000 liters of water with additional overflow capacity. Joined to the basin was a tall structure that supported and protected the lifting apparatus. The bucket chain, Oleson concludes, consisted of a pair of rope loops, each 27 meters long, to which 24 wood buckets were attached at intervals. The buckets, covered with pitch inside and out, held about seven liters apiece. Once they had emptied their contents into the collecting trough, the water flowed to the cistern, from which it was carried to the nearby villa through a system of lead pipes.

In most modern *saqiya* systems draft animals provide the motive force. The dimensions of the springhouse and the difficult access to the top of the structure that held the drive mechanism at Cosa indicate that animals were not used there; human power moved the chain. Slaves, criminals or prisoners probably did the work, one on each end of the bar that turned the drive wheel. Oleson calculates that it would have taken



SPRINGHOUSE at Cosa held machinery for lifting water from a natural spring. The machinery was probably worked by a pair of slaves, who pushed on the ends of a long bar. The bar turned a horizontal gear wheel that drove a vertical one. The vertical wheel was linked by an axle to a third wheel, which lifted a chain of 24 buckets. As they passed over the bucket-chain wheel, the buckets emptied their contents into a trough. From there the water flowed out to a cistern and was then distributed in lead pipes.

two slaves about 45 hours to fill the cistern, which held 127,000 liters.

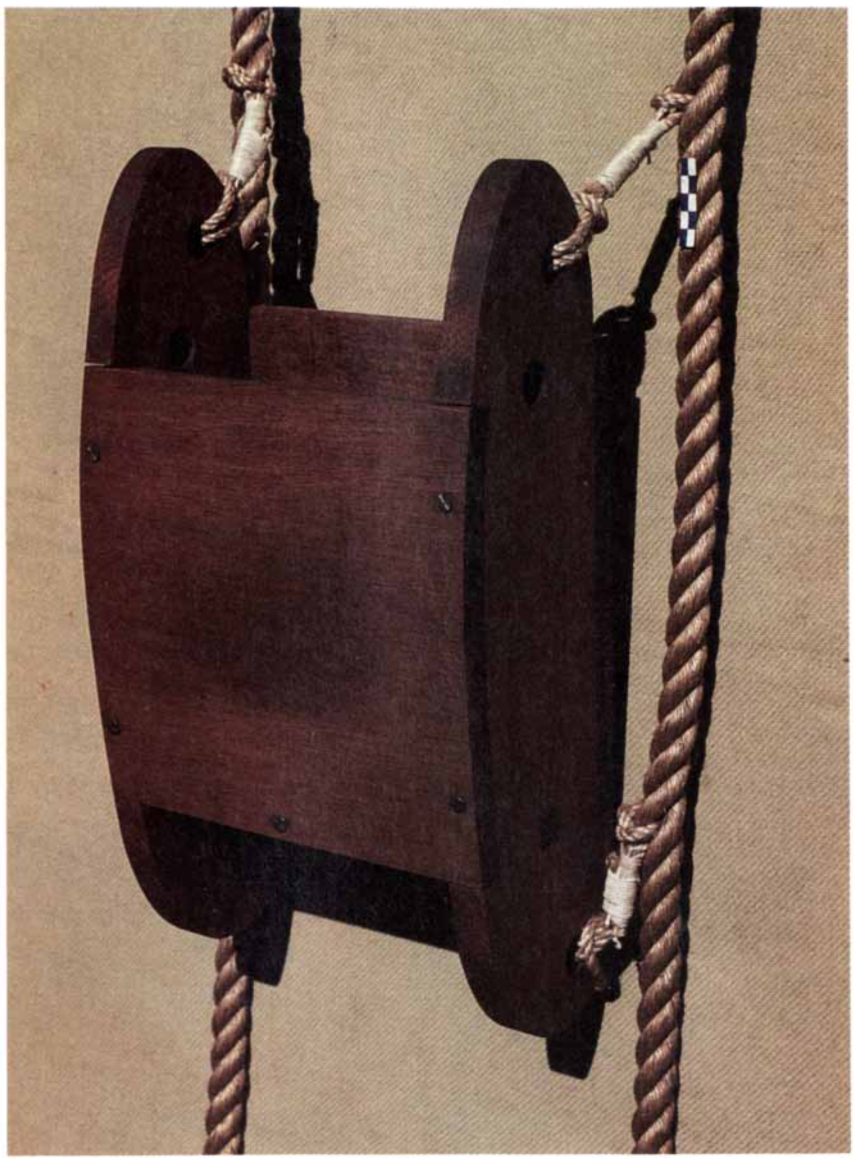
The remains of the water-lifting system at Cosa are unique. Not only are they the earliest dated archaeological evidence for such a machine but also the mechanism at Cosa lifted the water from ground level to a higher elevation. All other examples known from antiquity lifted water from subterranean springs to ground level. What is more, the buckets are the only examples made of wood that are known from ancient times; all the others that have been found are made of terra-cotta and date to the early third century A.D. or later.

By the time the second springhouse was built, sometime in the second half of the first century A.D., many changes had taken place at the port. No longer a great center for exports and industry, it had become a facility for the importation of the luxury goods demanded by the wealthy owners of the villas in the area. The merchant ships, sailors and laborers had vanished, along with the bustle of harbor life. In their place on Cosa's shores was the large villa mentioned above, which may have been the property of an emperor. Whether the villa was imperial or not, it was certainly a luxurious enterprise, and it was probably to supply its large bath building (a regular feature of any Roman villa) that the springhouse machinery was rebuilt.

Why had the port changed so dramatically? Several factors seem to have been at work. One of them was natural. As I have described, silting is among the greatest threats to any harbor, and this is particularly true along the sandy Tyrrhenian coast. The excavations at Cosa turned up a layer of gray sea sand filling the fish channels and spreading into the lagoon itself. Bourgeois has interpreted this as possible evidence for a great storm or tidal wave that silted up the harbor and some of the channels on which the fishing lagoon depended. Such a cataclysm might have been sufficient to reduce the commercial activity at the port.

In addition, by the beginning of the imperial period in 27 B.C. the need for a port at Cosa had diminished considerably. The emphasis of Roman trade had shifted from exports to imports: grain for the growing population of Rome and luxuries demanded by the imperial court and the aristocracy. Such imports were more readily handled by ports to the south, such as Puteoli and Ostia, the imperial port of Rome. The port of Hercules, which remained free of silt, was adequate for the remaining local traffic. For these reasons and others, by the end of the first century B.C. the port of Cosa had become a quiet backwater, reemerging only after about 75 years as the site of the elaborate seaside villa. That is what it remained until its final decline, which (along with the decline of the entire Roman world) came in the third century A.D.

In looking back on the history of the port of Cosa one is most impressed by the site at its commercial peak. The scale and vertical integration are reminiscent of much more modern corporate structures: goods manufactured at the port were packaged there and shipped by the same people who probably owned the ships and controlled the harbor. To provide the infrastructure for this enterprise, some of the most advanced technology of the day (in particular the water-lifting apparatus) had been imported. The overwhelming evidence is that the Sestius family, if not directly responsible for these developments, at least had a major role in them. Although the history of ancient ports remains to be written, it can safely be said that today's international trade has its origins in the energy of those such as the Sestii, who brought forth from a sandy beach and a freshwater spring the port of Cosa.



WOOD BUCKET from the Cosa springhouse was reconstructed by John P. Oleson of the University of Victoria in British Columbia. In the excavations one complete pine bucket and pieces of six others were found. Each was 54 centimeters tall and held seven liters.

THE AMATEUR SCIENTIST

Why sidespin helps the bowler—and how to keep scoring strikes

by Jearl Walker

In tenpin bowling you throw a heavy ball down a long, narrow lane to topple pins standing in a triangular array at the far end. The ball hits some of the 10 pins, which then crash into other pins in a chain reaction, either directly or after rebounding from the side walls or the rear of the lane. You score one point for each downed pin. If all the pins fall, the play is said to be a strike, and you not only score 10 points but also have a chance to win more points after the pins are reset; in one frame, or turn of play, you can win as many as 30 points. If your first throw is not a strike, on the other hand, you have only one chance to down the remaining pins (getting what is called a spare) before your turn is over. Clearly the goal is to make a strike with the first throw in each frame.

How should you throw the ball to increase the odds of making a strike? Many novice bowlers release the ball near the center of the lane, sending it along a straight line toward the foremost pin, the headpin. The shot is haphazard because the pin is so far away that aiming is difficult and because the action of the pins is seemingly unpredictable. Still, the play occasionally produces a strike. A seasoned bowler often adopts a more reliable strategy, in which the release is made along the side of the lane. The aim is not toward the pins but rather toward painted markers 15 feet down the lane. In addition, the ball is given a sidespin at the moment of release. The ball appears to travel parallel to the side of the lane for a short time and then to hook suddenly toward the headpin, approaching it at what seems to be a large angle. The target, which is called the pocket, is the left or the right side of the headpin.

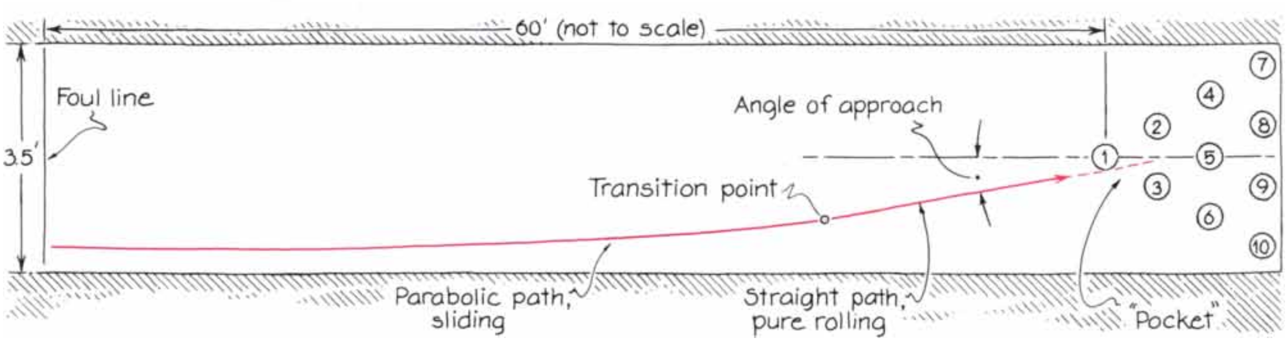
Experienced bowlers continue to claim that putting sidespin on the ball greatly increases the chance of a strike. They may be right. The game is subject to so many variables, however, that it is hard to verify the claim experimentally. Instead I set out to test its plausibility by answering several questions theoretically. Why does a sidespin yield a curved path? Does the ball hook at some particular point along its path? Why would an angled approach to the pins increase the chance of a strike? Is the angle of approach as large as some bowlers say it is?

The bowling lane is marked by a foul line beyond which the player must not step when releasing the ball. The headpin is 60 feet from the line. The wood pins, each of which is 15 inches tall and no more than 3.6 pounds in weight, are numbered according to their position; the headpin is the 1 pin. The distance between the centers of any two adjacent pins is one foot. The horizontal cross sections of the pins are circular and their maximum diameter is less than five inches.

The ball, which may weigh up to 16 pounds and be no more than about 8.6 inches in diameter, normally has three holes into which you insert a thumb and two fingers. (Weights inside the ball compensate for the weight loss from the holes.) The surface of the ball is plastic or a hard rubber composition. To throw the ball you first swing it toward the rear and then, stepping briskly toward the foul line, swing it forward in a pendulum motion, keeping your palm below and just to the rear of the ball. As you reach the foul line you crouch, sliding on the floor with one leg extended behind you so that the ball is low. When the ball reaches the lowest part of the swing, or slightly after, you release it.

The lane, made out of narrow wood boards, is $3^1/_2$ feet wide, bordered on both sides by a channel. Part of the lane is treated with an oily material so that the ball initially slides along the lane rather than rolling. How much of the lane is oiled varies from one bowling establishment to the next. At my favorite place, Tuxedo Lanes on the south side of Cleveland, the first third of each lane's length is oiled, except for narrow strips along the channels.

The curved path of the ball was first investigated mathematically in 1977 by Don C. Hopkins and James D. Patterson of the South Dakota School of Mines and Technology. I shall sim-



The alley and the ball's trajectory

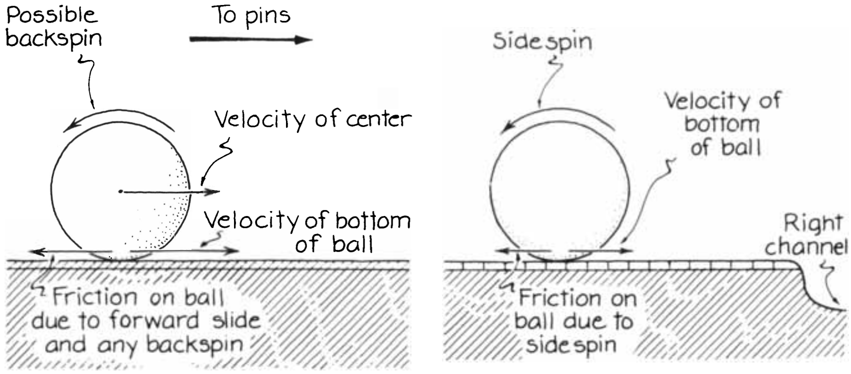
plify their treatment and also limit it to the case of a right-handed bowler who throws the ball directly forward on the right side of the lane. At the moment of release the fingers are brought smartly upward on the right side of the ball, giving it a counterclockwise spin. While the ball is sliding down the lane it is subject to two frictional forces. One force is rearward, opposing the forward motion. The other force, which is toward the left, opposes the spin. The rearward acceleration diminishes the ball's forward progress; the acceleration toward the left moves the ball away from the channel along which it initially slides.

Consider the ball from the perspective of the channel. If the bowler has put backspin on the ball (as some might do), the bottom of the ball has a forward speed that is greater than the speed of the center; otherwise it has the same forward speed as the center. For the ball to roll, the bottom of the ball must have a rearward speed that equals the forward speed of the center. As the ball slides, the rearward friction on the ball slows the center and also slows and then reverses the speed of the bottom of the ball. When the speeds are properly matched, rolling begins.

A similar interplay of friction and speed alters the sidespin. From a rear perspective the bottom of the ball moves rightward, whereas initially the center of the ball has no motion toward the left or the right. The friction resisting the spin slows it while also propelling the center of the ball to the left. When the speed of the bottom toward the right matches the speed of the center toward the left, rolling begins. The change is simultaneous with the transition to rolling in the forward direction.

During the sliding the combined friction on the ball sends it leftward along a parabolic path whose curvature depends on the initial values of the spin rate and the forward speed. For example, a greater value for the speed or a smaller value for the spin rate decreases the curvature. At the instant rolling begins, the ball leaves the parabola along a tangent to the curve, and thereafter it travels in a straight line.

To determine the angle of this straight-line roll, which is also the ball's angle of approach to the pins, Hopkins and Patterson estimated the forward speed and spin rate given to the ball. They also chose a representative value for the coefficient of friction, which is a measure of the surface roughness and degree of lubrication between the ball and the lane. In all their calculations the angle of approach was never more than three degrees. Such a measly angle hardly seems to justify the trouble of putting sidespin on the ball.



The friction on a sliding ball

I wondered if the angle was small only because Hopkins and Patterson assumed that the coefficient of friction was uniformly small throughout the ball's travel to the pins. I toyed with their equations, expecting to find that the disappearance of the oily preparation of the lane about a third of the way to the pins might create a larger angle. Perhaps when the ball reached the "dry" part of the lane, the sudden increase in friction whipped the ball into a larger angle of approach.

What I found surprised me: the angle of approach is independent of the coefficient of friction. Instead it is set by the initial ratio of sidespin to forward speed. (Backspin plays a minor role.) If the sidespin is small or the forward speed is large, the angle is tiny. If the spin is large and the forward speed is moderate, the angle can be 10 degrees or even somewhat more than that.

Although the coefficient of friction does not influence the ball's angle of approach, it does determine where along the lane rolling begins. If the coefficient is large, the ball diverges from a sharply curved parabola early in its travel and so may end up to the left of the pocket. If instead the coefficient is small, the ball diverges from a mildly curved parabola late in its travel and may end up to the right of the pocket. (If the lane is fully oiled, the ball might even reach the pins before it begins to roll.) Part of the skill in bowling is knowing how to throw the ball for your lane's particular coefficient of friction. That knowledge comes from practicing on the lane before a match, adjusting your throws until the ball hits the pocket properly. The task is not easy, because with each play the ball carries oil into the dry region, altering the coefficient of friction there.

In retrospect these results are not very surprising; indeed, they are exactly in line with advice from professional bowlers. If the ball hits the pin array "low" (too far to the right) or "high" (too far to the left), you should adjust the speed and spin in the launch to change the angle of approach. Alternatively you can adjust your position along the width of the lane to move the ball's path left or right. Or you can throw the ball toward the left or the right instead of straight ahead in order to rotate the path around the launch point.

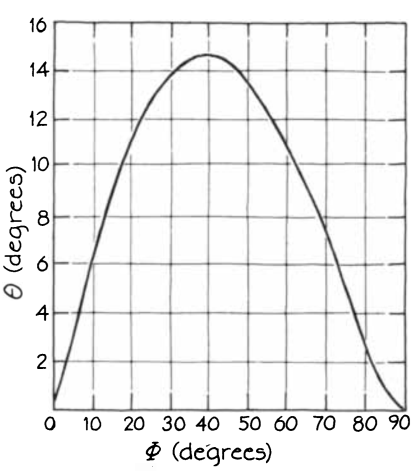
What accounts for the claims that the ball hooks suddenly? The ball can hook if it slides from an oily region of the lane to a dry one. The sudden increase in the coefficient of friction enhances the curvature of the ball's parabolic path, and the sharp increase in the ball's direction of travel toward the left is the hook. The angle of approach is not altered by the hook, however; a hook simply causes the ball to reach that angle sooner than it would otherwise.

If the angle of approach of a ball thrown with sidespin is only three degrees or so, does it actually increase the chance of a strike? To answer the question I decided to calculate what the ball does after it hits the 1 pin. I first drew an overhead view of the pin array with the aid of a template having circular cutouts. Because the center of each pin is one foot from the center of each adjacent pin, the array forms an equilateral triangle and so has 60-degree corners. Within the array the pins form smaller equilateral triangles.

I simplified the collision by assuming that it is too brief for friction between the ball and the pin to be important. (Certainly friction between the cue ball and a numbered ball is rarely important in pool.) I assumed too that the collision is elastic (no kinetic energy is lost) and that the total momentum of the objects does not change. In short, I imagined that the collision is like the collision of a large hockey puck (the ball) and a smaller puck (a pin), both pucks being on ideally slippery ice. The ball was represented by a circle having twice the diameter of a circle representing a pin; otherwise the array was drawn to scale.

Two angles are important in my analysis [*see illustration at upper left on opposite page*]. One is the angle of approach by the ball. The second angle is what I call the touch angle, which is the angle from the front of the 1 pin to where the ball touches the pin, as measured around the center of the pin. At the instant of the collision the pin is propelled by a force directed along a line connecting the centers of the pin and the ball. Hence the pin travels along a path that is angled from the forward direction by the touch angle.

The collision deflects the ball from its original direction of travel, away from the direction taken by the pin. The deflection angle (which I call theta, θ) depends on the angle (called phi, ϕ) between the ball's original velocity vector and the velocity vector imparted to the pin. A curve relating θ and ϕ is shown below. To construct the curve I assumed that the ball is four times as massive as the pin. Note that the maximum deflection of the ball is slightly greater than 14 degrees. (I shall gladly supply a copy of the calculations for the graph.)



How the ball's deflection varies

The collision between two pins is easier to follow. At the instant of a collision between pins the second (initially stationary) pin is propelled along a line connecting the centers of the pins. The first pin is deflected in a direction perpendicular to that line. (The only exception to such perpendicular deflection is when the collision is head on, in which case the first pin stops.)

Armed with my graph and template and a stack of photocopies of the pin array, I explored what happens when the ball strikes the 1 pin at various angles of approach and touch. If you do the same kind of analysis (either on paper, as I did, or with a home computer), keep in mind that the results are only suggestive, because they ignore many practical aspects. If a pin skips along the lane after being struck, friction can redirect it. If it lies down while moving, it can sweep a large path clear of pins. It may also fall to the lane and then spin about a vertical axis, knocking pins down. One pin might even collide with another moving pin. In addition, unavoidable drawing errors make the calculated path of a pin uncertain after it has struck more than one or two others.

To map the ball's travel through the pin array, first find its deflection after it hits the 1 pin. Extending the line of travel until the ball makes contact with the 3 pin, draw the ball at its point of collision. The pin's direction of travel is set by the line connecting the center of the pin to the center of the ball. The ball is deflected away from the pin, with its direction of travel changing by the deflection angle, θ. To find θ from the graph, measure ϕ, the angle between the path taken by the pin and the path of the ball just before the collision. Then draw the new direction of travel for the ball. The procedure is repeated each time the ball hits a pin.

Bowling textbooks describe a perfect strike as one in which the ball hits only pins 1, 3, 5 and 9. The 1 pin initiates a sequence of head-on collisions in which pins 2, 4 and 7 fall, while the 3 pin knocks down the 6 pin, which in turn knocks down the 10 pin. Once it is hit by the ball, the 5 pin drops the 8 pin. The fall of pins along the left side of the array suggests that the touch angle should be between about 20 and 40 degrees, with an ideal angle being 30 degrees, because a line along the left side lies at an angle of 30 degrees from the forward direction. With the touch angle between 20 and 40 degrees, I calculate that the 1 pin leaves the collision with as much as three times the speed at which the ball leaves it.

Concentrating on touch angles in that range, I investigated how variations of the angle of approach might alter what the ball does after it hits the 1 pin. If a nonzero angle of approach has any advantage, it should show up in such an analysis. If it does not show up, I would have to conclude either that putting sidespin on the ball is irrelevant or that (contrary to my experience in pool) the slight friction between the objects in a hard collision is somehow important.

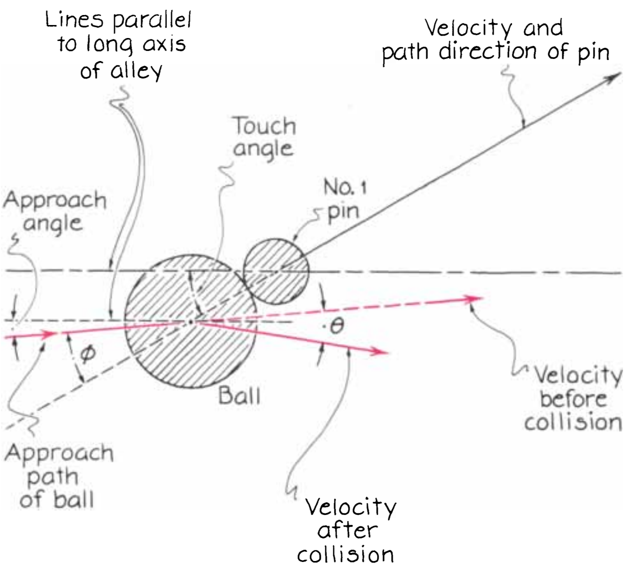
If the angle of approach is zero, the ball deflects into the right side of the array when the touch angle is 20 or 25 degrees. If the touch angle is between 30 and 40 degrees, the deflection of the ball by the 1 and 3 pins is so severe that the ball never penetrates the array. When the touch angle is between 35 and 40 degrees, the ball slants off rather sharply—in my calculations actually leaving the array before reaching the 10 pin, which remains standing. Bowlers describe the ball as taking a strange rightward bounce off the array, as if the array were a wall. The large deflection provides a distinct disadvantage; a strike is more probable if the ball, big and massive, bullies its way right through the array.

Suppose the angle of approach is three degrees. Then at all touch angles between 20 and 40 degrees the ball does penetrate the array on the right side. Although the angle of approach is only modestly different from the preceding example, the results can be dramatically different. The advantage of sidespin leading to an angled approach is that penetration is guaranteed for any touch angle that downs the left side of the array in textbook fashion. Penetration is even more pronounced for larger angles of approach. If the angle is 10 degrees, the strike is precisely the textbook example. To throw a textbook strike, then, you should launch the ball with moderate forward speed and high sidespin, adjusting your throw to lane conditions until the ball repeatedly touches the 1 pin at an angle of about 30 degrees. Even if you err somewhat, the strong penetration of the ball into the array and the complicated pin movements I have neglected will very likely still give you a strike.

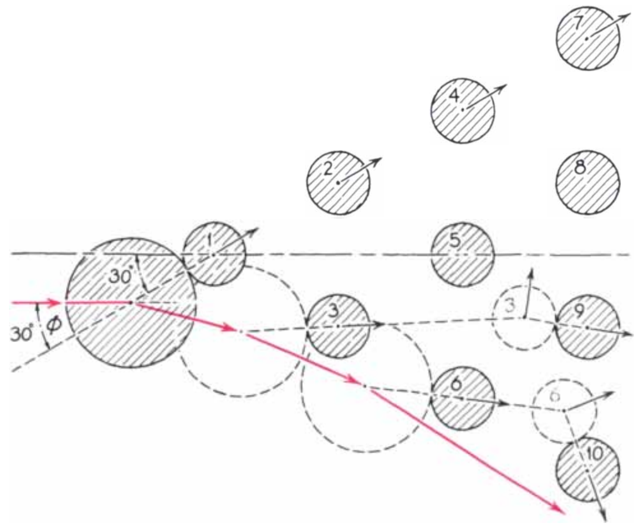
Following the ball on paper reveals another reason for not throwing the ball straight down the lane. If the ball hits the 1 pin at a touch angle of zero, the 7 and 10 pins will probably remain standing as the ball plows directly through the center of the array. To knock down the 7–10 "split" with a second throw is almost impossible. Some other difficult spares are at least feasible. Consider the spare with pins 6, 7 and 10. Bowling guides state that the ball should hit on the right side of the 6 pin, sending the pin into the 7 pin; the ball goes on to take out the 10 pin. Indeed, I find that with a touch angle of about 70 degrees on the 6 pin, the 7 pin is neatly eliminated. The collision on the 6 pin actually need not be that accurate if the pin falls and sweeps a wide path as it travels across the lane.

A wealth of material remains to be explored with a pencil-and-paper approach to bowling. You might investigate the "backup ball" throw, in which the ball approaches the 1 pin high and then curves back into the left pocket between the 1 and 2 pins. There are also hundreds of spares to consider. How should each of them be played in order to knock down all the pins?

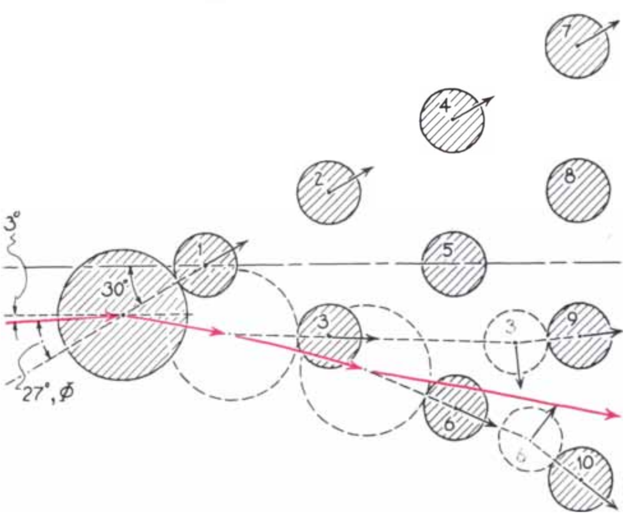
Details about the ball also raise perplexing questions. Having hit the 1 pin, the ball must slide until it begins rolling again. How does the friction on the ball during the slide change its direction of travel? (I suspect that the friction reduces the angle of deflection.) Does the nonuniform distribution of mass in the ball influence its motion? Does its large mass make it a gyroscope that resists redirection? Granted that a theoretical approach to bowling is limited because of all the variables involved in a real game, the analysis is still rewarding if it helps you to understand even approximately why the ball and the pins do what they do.



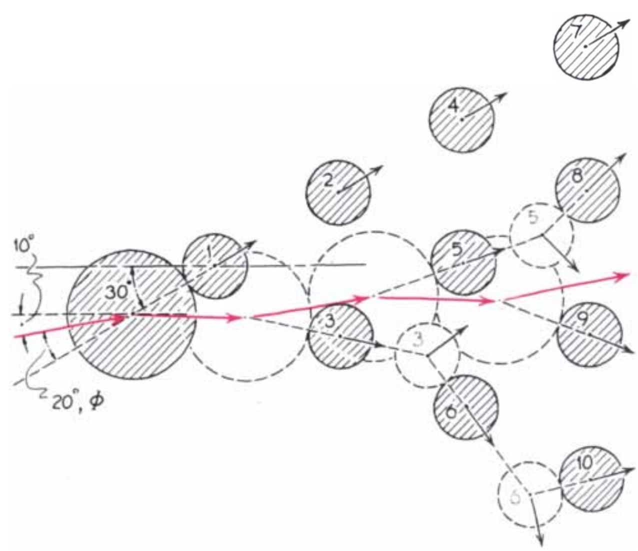
The ball's collision with the 1 pin



The ball's path when the angle of approach is zero degrees



The path when the angle of approach is three degrees



The path for a 10-degree angle of approach

COMPUTER RECREATIONS

A home computer laboratory in which balls become gases, liquids and critical masses

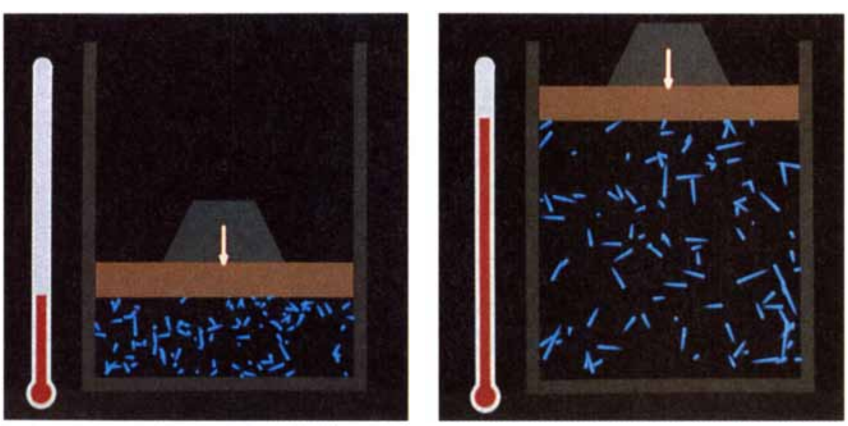
by A. K. Dewdney

A box full of balls flying in all directions, hitherto only a mental model for physical matter, can now be viewed on the screen of a home computer. With a modicum of programming effort readers can witness the molecular collisions that produce pressure within a cylinder or cheer on the random jostling of differently colored molecules as one liquid diffuses through another. Those undaunted by the prospect of a nuclear explosion in their computer can even play with critical masses of the unstable solid I call gridium.

The inspiration for this month's excursion comes from James F. Blinn, an affable researcher at the Jet Propulsion Laboratory in Pasadena, Calif. Blinn has developed simulations of various physical systems. The simulations have been gathered into a rather funny, laconically narrated videotape titled "The Mechanical Universe." Two of Blinn's simulations that caught my eye as potential subjects for this department were charming animations of a gas-filled cylinder and a diffusion chamber. In the first case a few dozen balls bounded and rebounded within a two-dimensional cylinder, driving a weighted piston into a state of jittery equilibrium. In the second case balls of two different colors bounced back and forth within a fixed enclosure; at first the two colors were separated, but as the balls moved about in Brownian fashion, the two colors became thoroughly mixed. The third simulation to be described here, the nuclear explosion, is my own entry. The idea springs from the primitive wish to watch a matrix explode.

The few hours it may take to create a working model of the program I call BOUNCE will be well repaid by the sight of a weighted piston coming into equilibrium with a cloud of bouncing balls. In Blinn's simulation the balls bounced off one another, but a much simpler and equally effective program can be written if the balls ignore one another and bounce only from the walls of the cylinder.

The user of BOUNCE may specify both the speed of the balls and their number. The position assumed by the piston follows the so-called ideal-gas law: $PV \propto T$, which is to say that the product of the pressure P and the volume V of a confined gas is proportional to its temperature T. The molecular speed specified by the user amounts to selecting a certain temperature, and the weight of the piston determines the pressure. What volume will result? The answer can be seen clearly on the screen [*see illustration on this page*]. The higher the temperature of the gas rises, the higher the piston goes up.



As the temperature rises the volume increases in James F. Blinn's simulation

There is some value in following the path of a single bouncing ball before tackling a multitude. Consider, then, a box drawn on the display screen of one's computer. The box, 50 units on a side, contains a ball somewhere within it. Because the demonstration lies in the plane, the ball has two coordinates of position, say x and y. It also has two coordinates of velocity, v_x and v_y. In a short period of t time units the ball will move from the position (x,y) to the position $(x+t\cdot v_x, y+t\cdot v_y)$.

The piston cylinder that is drawn by BOUNCE, also two-dimensional, has three fixed walls defined by the lines $x=0$, $x=50$ and $y=0$. A fourth wall represents the piston. Its line has the equation $y=h$, h being the current height of the piston. No ball-animation program can escape the need to discover if and when a given ball will collide with one of the walls. BOUNCE does this by testing the current direction of a ball against each of the four equations: where does the direction line intersect each of the four wall equations? The left half of the bottom illustration on page 116 provides the answers.

In each case the t value listed gives the time (from the present) at which the ball will strike the wall in question. The actual point of impact is computed by substituting the t value into simple position formulas. If the t value happens to be negative, it means that the ball, continuing in its present direction, will never strike the wall because it is currently moving away from it. If the t value is infinite, one of the ball's velocity coordinates must be zero. Here too the ball will never strike the wall. Naturally BOUNCE skips the computation of a t value if that is the case.

The wall that has associated with it the smallest nonnegative t value is the one the ball will strike first. That value of t can be used to calculate the actual point at which the ball will strike the wall. Again a small table comes in handy [*see right half of bottom illustration on page 116*].

Most simulation programs exploit

one of two possible techniques for advancing the imaginary timepiece known as the simulation clock. The time-slice technique advances the clock by a small fixed increment, updating appearances in a uniform manner. The critical-event technique advances the clock to the next event of interest. Readers may remember the star-cluster simulation program featured here in January, 1986. The cluster's entire appearance was updated by using a time slice of just a few years. BOUNCE more naturally employs the critical-event method, where an event is said to take place when a ball bounces off a wall.

Crucial to the critical-event approach is an array called *event*. If BOUNCE is juggling *n* balls, it must know which ball will strike a wall next. The array is actually a priority queue. Here is how BOUNCE operates:

repeat
  1. get time, ball and wall from *event*(1)
  2. update system by time
  3. determine next event for ball
  4. insert time, ball and wall in *event*
  5. display the system in new state
until key

The cycle of five steps will be repeated until the home computer scientist presses a key; the December 1987 column on special effects outlines the use of keys for this purpose.

Step 1 is fairly simple. The event queue actually consists of three arrays: *event* stores the times to the next events in the order of their occurrence, an array called *ball* stores the index of the balls that correspond to these events and another array called *wall* stores the corresponding walls. For example, *event*(2) holds the time until the second event and *ball*(2) contains the number of the ball *b* involved; *wall*(2) stores the number of the wall *w*, which takes on one of the four values, 1 ($x = 0$), 2 ($x = 50$), 3 ($y = 0$) or 4 ($y = h$). Step 1 thus includes three instructions to place the three items of information in variables *t*, *b* and *w*.

Step 2, updating the system, means changing the position of all the balls in the box. This can be done within a loop of the following form:

for $i \leftarrow 1$ to n
  $x(i) \leftarrow x(i) + t \cdot v_x(i)$
  $y(i) \leftarrow y(i) + t \cdot v_y(i)$
if $w = 1$ or $w = 2$ then $v_x(b) \leftarrow -v_x(b)$
if $w = 3$ or $w = 4$ then $v_y(b) \leftarrow -v_y(b)$

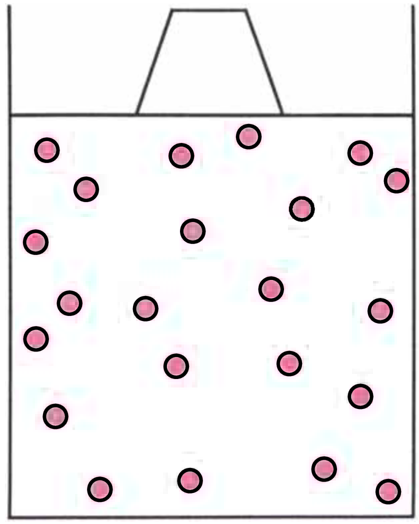
Readers will recognize the position formulas introduced above. There is no danger that any of the balls will end up outside the box because the one ball that comes closest is only at a wall, not through it, by time *t*. The two if-then statements handle the bounce of ball *b*. If the ball strikes either of the two vertical walls, its *x* velocity is reversed; a bounce on a horizontal wall means that the velocity coordinate in the *y* direction must be reversed. Updating the system also means decrementing all the times stored in the array event. Here BOUNCE merely subtracts *t* from each array entry.

I have already described step 3, the determination of the next event for ball *b*, implicitly; now I shall describe it in more detail:

if $v_x(b) = 0$ then $t_1 \leftarrow 5{,}000$
  $t_2 \leftarrow 5{,}000$
 else $t_1 \leftarrow -x(b)/v_x(b)$
  $t_2 \leftarrow (50 - x(b))/v_x(b)$
if $v_y(b) = 0$ then $t_3 \leftarrow 5{,}000$
  $t_4 \leftarrow 5{,}000$
 else $t_3 \leftarrow -y(b)/v_y(b)$
  $t_4 \leftarrow (h - y(b))/v_y(b)$

The time-to-impact for ball *b* and each of the four walls is encoded in four variables, t_1 through t_4. The unreasonable value of 5,000 is assigned to any of the four variables if ball *b*, continuing in its present direction, would never strike the corresponding wall. A computational sieve then serves to isolate the next wall to be struck by ball *b*. First BOUNCE sets *t* to 5,000, then it tests each of t_1 through t_4 in turn. The test has three parts: if t_i is less than *t*, if t_i is greater than 0 and if *wall*(*b*) is not equal to *i*, then the program replaces the current value of *t* by t_i and replaces *w* by *i*.

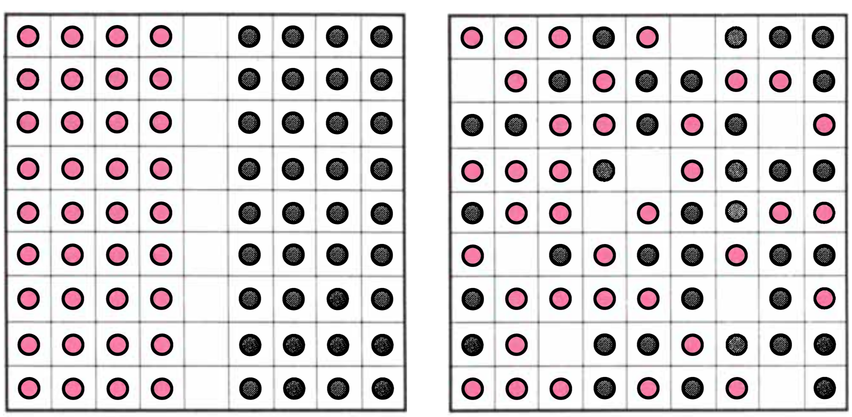
In step 4, BOUNCE refurbishes the event queues. The main task is to find where in the queues the next bounce of ball *b* belongs. That is easily accomplished:



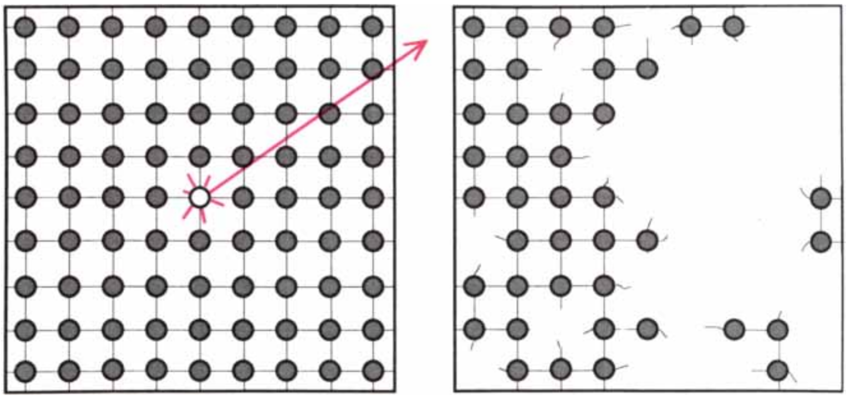
A schematic of a pressure cylinder

$k \leftarrow 2$
while $event(k) < t$
  $event(k-1) \leftarrow event(k)$
  $ball(k-1) \leftarrow ball(k)$
  $wall(k-1) \leftarrow wall(k)$
  $k \leftarrow k + 1$
$event(k-1) \leftarrow t$
$ball(k-1) \leftarrow b$
$wall(k-1) \leftarrow w$

Here BOUNCE searches through the event queue to find where ball *b* now properly belongs. As it searches it shifts items in all three arrays one cell to the left. As soon as it finds a cell *k* in which *event*(*k*) is not less than *t*, it is obvious that the time to the next impact of ball *b* belongs in the preceding cell, *k*. (When the balls in the cylinder are given their initial positions and velocities, the event queue must also be set up. To avoid programming effort, the positions



Two liquids diffuse in a 9-by-9 matrix



A sample of gridium before (left) *and after* (right) *a chain reaction*

and velocities can be selected by hand and the event times worked out. The initial event-queue entries can then be typed in.)

The final step requires BOUNCE to display the box and all the balls in their current positions. This straightforward operation must be supplemented by erasing the balls in their previous positions, of course.

To get started, BOUNCE asks the user to type in a speed at which the balls are to travel. If the variable is called v, the program then selects n random numbers between $-v$ and v, placing these in the x-velocity coordinates $v_x(i)$. The y velocities are then calculated by means of the Pythagorean theorem, namely the square root of $v^2 - v_x^2(i)$. Readers may want to try various values of v in the range from 1 through 10.

Those who write BOUNCE will naturally want to watch the weighted piston move up and down in response to intermittent pressure from within the cylinder. For this purpose, step 2 must include a calculation that subtracts the downward movement of the piston due to gravity from the upward movement of the piston due to the transfer of momentum from one ball. The result is a new h.

For the sake of simplicity the interaction of the balls with the piston has been distorted to the point of chicanery. In time t the piston falls $t^2/100$ units. When a ball with y velocity v_y strikes the piston, the piston moves up $v_y/100m$ units. My apologies to physicists, but it works! Users of BOUNCE must, of course, input the mass m of the piston, as well as the number of balls, n, and their speed, v.

The diffusion program I call BLEND treats the molecules of a liquid as balls that are able to move only one step horizontally or vertically at a time. A 31-by-31 array called *tank* holds 930 balls, which are set up initially somewhat like a checkers game. BLEND is even less realistic than BOUNCE but it too works. The balls in the left half of the box are colored red and those in the right half are colored gray [*see bottom illustration on preceding page*].

BLEND proceeds by selecting an array position at random, then selecting a random adjacent position. If a ball occupies the first position and no ball occupies the second, the ball is erased at its current position and redrawn at the new one. If a ball is absent at the first position or present at the second, BLEND could not care less. With microseconds of time on its hands, a movable ball will sooner or later turn up. How long will it take for the two stylized liquids to diffuse completely? No detailed algorithmic description accompanies this recreation. Readers are left to their own devices: hardware and mindware.

With unholy haste I have rushed to the matter of gridium, a violently unstable element. One atom of the stuff spontaneously decays, shooting off a neutron in the process. I am reminded of a famous demonstration by Walt Disney; the great man once filled a high school gymnasium with cocked mousetraps. Each trap held a Ping-Pong ball resting on the snapping arm normally meant for crushing rodents. With a mischievous grin, Disney threw a single Ping-Pong ball into the gymnasium. Within less than a second the air was literally filled with thousands of balls and there was a strange roar. This was probably the safest chain reaction ever.

Gridium happens to be a crystal. Its atoms are arranged like balls in a square matrix measuring $n+1$ balls on a side. On a signal from the reader the central ball decays, producing a single neutron that zips along a random line through the matrix. Another ball will decay if it is struck by the neutron. Will a chain reaction ensue? If the matrix is small, the reaction will most likely be a dud. But larger matrices are increasingly likely to go off with a silent roar [*see top illustration on this page*]. How much larger? That is for the home experimenter to discover. There can be little doubt that even a microcomputer would have vastly relieved the work load at Los Alamos in 1944.

This department has no military secrets. I shall be pleased to report values for the critical mass of gridium as discovered by readers brave enough to try the experiment. The program I call BOOM uses an array called *grid*. Because most readers do not have variable array dimensions available to them, I have decided to make *grid* 50 by 50. When the user types in a value for n, BOOM will automatically limit its considerations to the first n rows and columns of *grid*. These will contain the atoms, one per cell in the region. For purposes of both display and calculation, atoms of gridium have one-unit radii and six-unit spacings. It is convenient to regard the atom at array position (i,j) as having coordinates $(6i,6j)$.

When an atom of gridium decays, BOOM selects a random angle a and calculates the sine and cosine of that angle, storing the values in the variables s and c. If the decaying atom has coordinates (x,y), the point and the angle together determine a line. The key question is, "How close does the line come to another atom having coordinates (u,v)?" The answer is given by a formula, $(u-x)^2+(v-y)^2-[(u-x)\cdot c+(v-y)\cdot s]^2$. The formula represents the square of the distance between the point (u,v) and the closest point to it on the line in question.

WALL	t VALUE
1: $x=0$	t_1: $-x/v_x$
2: $x=50$	t_2: $(50-x)/v_x$
3: $y=0$	t_3: $-y/v_y$
4: $y=h$	t_4: $(h-y)/v_y$

WALL	POINT OF BOUNCE
1: $x=0$	$(0, y+t_1 \cdot v_y)$
2: $x=50$	$(50, y+t_2 \cdot v_y)$
3: $y=0$	$(x+t_3 \cdot v_x, 0)$
4: $y=h$	$(x+t_4 \cdot v_x, h)$

Tables for computing the time (left) *and place* (right) *of collisions*

Rather than bothering with taking the square root of the resulting number, BOOM merely checks whether the number is less than or equal to 1. If it is, the atom at (u, v) must undergo fission.

Every time an atom in *grid* thus explodes, BOOM enters a loop that systematically plugs all n^2 coordinates into the variables u and v, omitting the exploding atom. All atoms that must undergo fission next are added to a list in the form of an array called *fire*. An index called *fired* separates those atoms already fired from those yet to fire. A second index called *end* marks the end of the list. Every time a new atom explodes, all the atoms intersected by the neutron path are added to *fire*, the index *end* being incremented each time. When all such atoms have been determined, *fired* is incremented and the cycle is reentered. When *fired* equals *end*, the common value of these indices tells the number of atoms that took part in the chain reaction. For what value of n will the entire matrix blow up? The statistical fluctuation in the answer will not be much.

Because of space limitations, this description of BOOM must suffice. It is possible to install a display routine within BOOM, but it is not necessary to undertake such an effort to find the answer. I shall endeavor to publish the most interesting picture of a partially blown-up matrix sent to me, however. It will represent a value of n very close to the critical mass.

Copies of the videotape "The Mechanical Universe" can be obtained by calling the toll-free number 1-800-LEARNER. For information about a newsletter by the same name, readers should write to "The Mechanical Universe," California Institute of Technology, Pasadena, Calif. 91125.

Last November this department revisited the Mandelbrot set and introduced a myriad of its cousins, all called Julia after the French mathematician Gaston Julia, a leading investigator in dynamical systems. Programs that generate Julia sets are as simple as those that produce the Mandelbrot set. Consequently hundreds of readers, perhaps even thousands, have now written programs—or would like to. Those who would like to pursue the subject further may want to subscribe to the fractal newsletter (in subject, not appearance) called *AMYGDALA*. It is filled with numerous programming hints for producing Mandelbrot and other fractal sets. The newsletter also contains addresses of programmers willing to share or sell software. Interested readers may write to the publisher, Rollo Silver, a Mandelbrot aficionado who can be reached at Box 219, San Cristobal, N.M. 87564. Silver also has slides for sale.

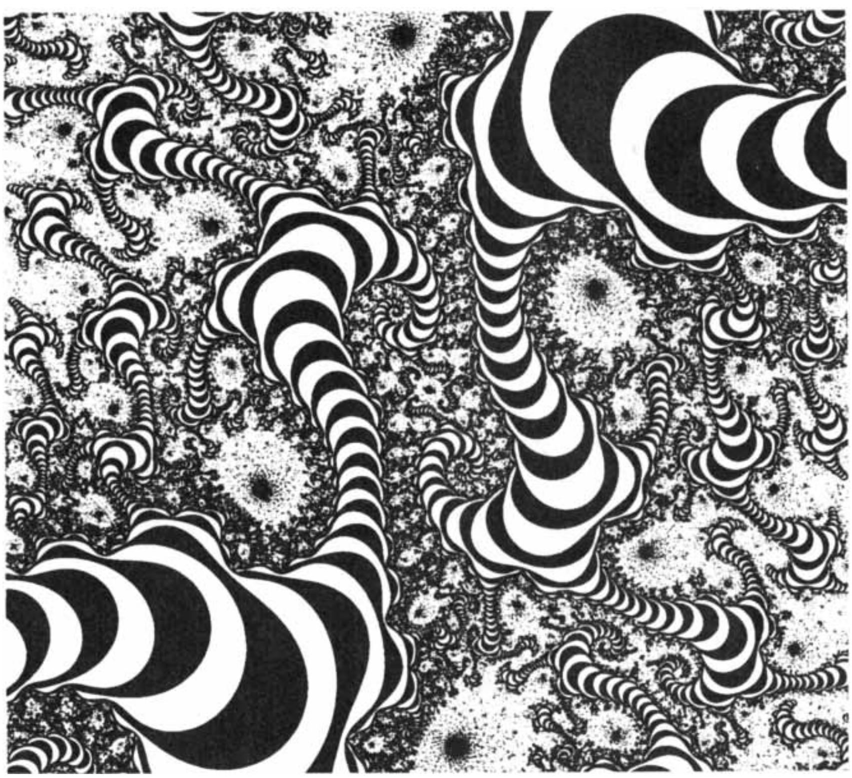
At the end of the November column I displayed a three-dimensional monster that was actually a cross section of a four-dimensional Mandelbrot set. Heinz-Otto Peitgen, the West German mathematician and artist, has revealed the prescription for the four-dimensional set. Iterate the equation $z \leftarrow z^3 - 3a^2z + b$, where a and b are both complex parameters. What points play the special role of $z = 0$ in the four-dimensional set? Actually two candidates appear, namely $z = +a$ and $z = -a$. There are consequently not one but two four-dimensional Mandelbrot sets.

Gary Teachout of Everett, Wash., writes that he views the Mandelbrot and the Julia sets as cross sections of a four-dimensional set, although his four-dimensional set does not correspond to either of Peitgen's. The basic iteration formula, $z \leftarrow z^2 + c$, produces a Mandelbrot set if z remains fixed and c is varied. It produces a Julia set if c is fixed and z is varied. But z and c together define four dimensions, two for each number. By "stacking" together all the Julia sets, not in a one-dimensional pile but in a two-dimensional pile, the superset emerges. M. G. Harman of Camberley, England, speculated on the existence of the same object.

The four-dimensional prize for initiative goes to Winston D. Jenks of Stamford, Conn. Not content with ordinary complex numbers, Jenks has invented his own, four-part numbers. His numbers are similar to quaternions, a sophisticated set of numbers invented by mathematicians in the 19th century but used today by physicists. Jenks has sent me a three-dimensional representation of his solid as a sequence of 29 slices through it.

Of course, readers who want to generate images of the Mandelbrot set also want to save time in doing so; besides images, the iteration process has produced many chewed fingernails. Dick Holt of Silver Spring, Md., has found that he can increase speed 8 percent by avoiding the squaring routine of his programming system. Multiplication (x times x) works faster than squaring (x squared).

Finally, the illustration below is another addition to the ongoing Mandelbrot gallery. The image was sent by James E. Loyless of Lilburn, Ga. His equipment sounds like an Indianapolis 500 car: an "XT turbo clone with an 8087 math coprocessor and a Toshiba 321 24-dot printer."



James E. Loyless' addition to the Mandelbrot gallery

BOOKS

Tigers at bay, super-rice, incompetent cells, dancers and bullets, shatter cones

by Philip Morrison

Tigers of the World: The Biology, Biopolitics, Management, and Conservation of an Endangered Species, edited by Ronald L. Tilson and Ulysses S. Seal. Noyes Publications ($64).

The tiger in the forests of the night and the human afoot by day are alike recent creatures, both framed from ruder forms during the past two million years. The fossil record, the biochemical studies of DNA hybridization and electrophoretic analysis of soluble-protein mobility leave no doubt concerning the date. The five extant tiger subspecies (three have fallen to extinction in this century) emerged more recently still; like the races of humankind, they have been distinct for only some 10,000 years.

The oldest form and the most distinctive in cranial character is the South China tiger: it is quite surely the stem population. Although 1986 was the Year of the Tiger, there were only 30 or 40 South China individuals left in the wild; a somewhat larger number are in the zoos of China. Only two China tigers are known to be alive outside the country; they live in the Moscow zoo. The China tiger's enemy is the armed and expert poacher, anxious for the body parts that symbolize strength and courage: tiger bones for medicinal wine fetch a high price.

The China tiger is the smallest of the subspecies. The Amur tiger of Siberia and northern Korea is the largest. Some 400 of them live in the wild; about the same number are kept in zoos around the world. So the count goes; only small, fragmented populations of the big cats remain.

India is to a degree an exception. The Bengal tiger is in some ways the symbol of the entire Indian subcontinent, older and deeper in meaning than even the Sanskrit myths themselves. During the 1970's India saw the establishment of a dozen protected tiger reserves; on the subcontinent what they call Project Tiger is off to a hopeful start. The census claims that about 1,000 animals live on those reserves; about 3,000 more inhabit other wild areas. That count is tenfold lower than the pre-World War II population. Or is it? A sharply critical chapter presents persuasive empirical evidence that the historic count, which depends on the claim that managers can identify individual animals from their pugmarks, is seriously optimistic. Still, it seems unlikely that the population of 2,000 interbreeding animals per subspecies can persist. The number is critically important: a smaller population cannot maintain the genetic diversity that long-term survival requires. The Bengals seem too fragmented to make it on their own, although there are stocks in India, Nepal, Bangladesh and Burma.

A tiger needs a few square miles of thriving forest to support its big hoofed prey; protected reserves as big as Connecticut are hardly to be found. Tiger and human together are in crisis; the next century will tax them both, until perhaps one day the pressures on wild lands will lessen. Meanwhile wild stocks must be scrupulously managed, mixed and even bolstered with animals bred in captivity; to do all that human beings need to know much more than they now do about tigers.

Long ago the big cats took primates, including our early hominid forebears, as prey. The existing subspecies of hominid and feline have learned to coexist; human beings are not good game for tigers. Yet there smolders still a kind of low-intensity hunting; the best accounts of the conflict come from the big delta of the Ganges, the multitude of low stream-braided islands that form the shore between Calcutta and Dhaka along the Bay of Bengal. That wild, watery terrain is covered by the largest mangrove forest in the world. Home to several hundred tigers and many deer and pigs, it is visited by men and women who gather firewood, fruit and wild honey.

Each year both tigers and humans die, scores of them, usually more humans than tigers. Most humans die in the bite (usually applied to the right nape) of big males that have made a specialty of man-killing, perhaps one tiger among 100. One animal in 10 is an opportunistic predator of humans, led to the hunt both by increasing familiarity with the prey and by the availability of victims, most of whom are taken during the honey season. Understandably the delta's people claim a prior right to protection. Project Tiger in the Indian sector of the Sundarbans has undertaken a "resolute and imaginative" campaign to reduce man-killing indirectly by the use of electrified fences and even of clay dummies costumed in old clothes, necks wrapped with electric-shock wires (fused against fatal shock). Eleven dummies have been attacked, all in the buffer zone outside the core of the preserve. Human deaths there have fallen since 1981 from about 40 a year to only 20.

The 46 papers in this volume are the yield of a 1986 symposium. Big-cat specialists from laboratory, museum, zoo, field and even lobby wrote them. The documents are diverse in every way, but the focus is so clear, the organizers and editors are so involved in the task and the common concern is so strong that the book is a model among such reports. Here is tiger zoo diet, white-tiger mutants, tiger reproduction, an antibiotic cure for the "tiger disease" and many plans and proposals.

There is even some good writing. John Seidensticker is a widely traveled zoologist from the National Zoo in Washington, D.C. He evokes no merely numerable, genetic tiger but the wild, fearful Tyger itself. "When the water drains off the higher ground with the ebb and steel-gray mud banks are exposed and shimmering under a scorching mid-day sun, the place to look for a tiger is in the shade at the mouth of small, side khals. And if you are truly fortunate you see it: head raised, lying half-submerged, intently watching as you slip by in the launch—a classic Sundarbans tiger." One is glad for the light of day.

Global Aspects of Food Production, edited by M. S. Swaminathan and S. K. Sinha. International Rice Research Institute and Tycooly International, Oxford, Riverton, N.J., and Dehra Dun (paperbound, $24.50).

Three curves wiggle across the graph field. What they trace is per capita food production in Asia, Latin America and sub-Saharan Africa (FAO data) from the early 1960's to 1983. Malthus finds no support here; two of the graphs climb steadily, Asia up 20 percent. Only the African line amidst drought and civil wars dips sadly and steadily downward. On the same page real rice prices are also plotted; Indian and Indonesian consumers have had the satisfaction of seeing the price of their staple fall by one-third or more in the past 15 years.

This volume, consisting of 18 expert chapters, examines the task of feeding the world in the decades ahead—the tag end of this century. Its editors are a crop geneticist and a plant physiologist. The senior editor, M. S. Swaminathan, is the current director general of the International Rice Research Institute near Manila; Professor S. K. Sinha works in New Delhi. The staff of the IRRI stands among those who have caused two blades—better, two heads—of rice to grow where one grew before. Twenty years ago when dwarf rice strains from Japan and Taiwan were successfully crossed with a tall Indonesian variety, the first semidwarf lodging-resistant offspring showed high yields. That richer paddy suffered more than its ancestors did from weeds and pests. The newer strains are widely resistant. They repay added inputs of fertilizer, pesticides and water generously; they also give more grain and less stem even to a poor farmer who cannot afford modern methods.

Director Swaminathan submits a concise balance sheet for the Green Revolution in rice; he grasps the system as a whole, eschewing mere tabulations of yield per hectare. In fact, he directs attention—this alliteration must be well known around the IRRI these days—to five domains: economics, equity, employment, energy, ecology. It is engaging to read that the first important benefit of all was the generation of self-confidence among farmers and among the people who worked with them; "this is a great gain," one easily ignored among the tables of data.

A global overview is helpful to a reader far from green fields. Although the problem is not uniform, we must depend on understandable if sometimes misleading aggregates. About four-fifths of what people eat is cereal and the other produce of farm fields; waters and rangelands contribute the remainder, each in equal proportion. On the global average, two of us now share one ton of grain equivalent per year; that covers a range from about one ton for each person in the developed world to one ton per five or six people in the least developed countries. Only a fraction of that difference is in human calorie intake; most of the spread represents the costly feeding of animals for indirect nutrition here in the affluent and indelicate meat-eating world.

Our youthful global population, now five billion strong, will certainly continue to grow in size; maximum primary production should be able within 50 years' time to feed the anticipated complement of eight billion diners a more than ample diet of one ton per head per year. To do so will require a 50 percent increase in arable land and a doubling of yield. That seems plausible, presuming modest land cost and an improved increase in yield. The facts support the yield side of the picture. A table of best performances shows grain production tripling in the Punjab within 10 years, to be sure the most remarkable record of all time. (India now holds a major reserve grain stock.) Average yields expected under high-input but affordable conditions today are more than three times the present value worldwide. More likely, annual take will remain distinctly smaller than one ton per head, but the much improved world diet need not imitate the beefy luxuries of land-rich North Americans.

The chapters of this collection take up question after question. A few of the details are illuminating. Increased market demand for food before 2000 will come more from the increasing income of poorly fed consumers in many countries of the developing world than from population growth. The norms that define human nutritional needs are themselves crude aggregates. A table lists the energy intake of nearly 30 young adult women, each of whom regularly prepared 500 to 600 chapati during an intense day's work in a community kitchen near Poona. The woman with the highest measured intake needed about 2,500 calories a day; two others managed with less than 1,600 calories, the lowest intake measured. There was little correlation between intake and body weight, nor did members of the group gain weight on the job. They had been among "the poorest of the poor" and were now guaranteed enough to eat. Yet we know from Japan that people there grew taller as over decades their diet improved.

More commercial energy input is inescapable if yield is to increase. The Indian experience documented here demonstrates its power. Over a period of 20 years the commercial energy input rose from very little to a sizable amount. By the 1980's the investment earned three times as much sunfed crop energy as the energy spent each year, all that hydropower and crude oil amplified and rendered good to eat. The prospect does raise real questions for nations worried about paying for imported oil; yet the amount of energy devoted to farming globally would hardly amount to as much as 5 percent of total energy use, not much more than is spent for cooking (worth it for our daily bread).

The industrial fixation of nitrogen will continue to increase rapidly, even if in the long run engineered nitrogen-fixing bacteria might be induced to dwell at the cereal roots. No decentralization of efficient large-scale nitrogen fixation seems to lie ahead. One remarkable little truck-borne plant, however, has been developed to fix nitrogen from the air at large-farm scale, using an electric arc fed by a solar panel.

The epidemiology of crop diseases is being approached at a newly sophisticated level. Control need not be obtained by the brittle scheme of evading pest strains until one or another selected enemy bursts out anew. Rather, a wide set of infecting strains are kept under control simultaneously by a host plant that supports only a lower rate of disease spread, say by allowing fewer fungus colonies on any leaf. Many genes, not one, are usually involved.

The special circumstances of some farmers are not ignored: chapters address the saltiness of irrigated farms, desertification and control of insects, from locust swarms to aphids. We are getting steadily better at meeting most of these challenges.

Global ecology may move toward center stage in a few decades. One unsolved issue is urgent: although a world grain stock of 300 million tons is on hand, 500 million people and more go hungry every day (as the streets of your own city may exhibit). Equity and compassion say there is much to be done beyond production for the market. All the same, give the optimistic director general the final word. He writes from developing lands with modest reserves and very many at table: "We live in this world as guests of green plants and of the farmers who cultivate them. If farmers are helped to produce more, agriculture will not go wrong. If agriculture goes right, everything else will have a chance for success."

The publishers have made this international book inexpensive, attractive and prompt; inadequate copy editing leaves the reader with quite a few small puzzles.

MICROCOSMOS, by Jeremy Burgess, Michael Marten and Rosemary Taylor. Cambridge University Press ($29.95). **STOPPING TIME: THE PHOTOGRAPHS OF HAROLD EDGERTON,** text by Estelle Jussim, edited by Gus Kayafas. Harry N. Abrams, Inc., Publishers ($35).

A generous new visual anthology of the microcosm spreads before us some 300 large photomicrographs illuminated by clear captions. This is the treasure trove of all the microscopies, images made through the light microscope, under the depth-revealing scanning electron probe or by the transmission electron microscope's perusal—with its coherent beam—of those airily thin specimens the device requires. All three instruments are here engaged with tiny samples of the world at large: the cells and tissues of the human body and all the kingdoms, animal, vegetable, mineral and metallic, as well as many everyday materials and industrial samples. The expert compilers include an electron microscopist and two London photo librarians; these authors have both combed the rich archives and made new images. Two specialists, Mike McNamee, the materials engineer and microscopist, and Rob Stepney, a medical writer and photographer, contributed entire chapters.

There are other good collections of the kind; this volume is the widest and most attractive compilation made so far. It offers wonderful images to admire and to pore over. One big page shows an image made by a scanning electron microscope (SEM) of the hundreds of tiny elongated bumps pressed into a plastic CD that store a split-second part of Mozart's 40th Symphony. Those bumps enlarged a thousandfold are rigorously aligned along a tight spiral: a gap one micron wide separates successive turns. The variable lengths and spacing of the bumps along that precise spiral path encode the delicious music. The bumps are read out by IR reflection from the aluminized mirror surface as almost a million of them spin past the scanner every second. The disc is an artifact of wonderfully controlled intricacy, even though it is only a static structure and not metabolically alive as is the typical silicon chip. On the opposite page a scratch that mars an LP record looks when magnified sixtyfold as gross as it sounds.

In another chapter there is a grim but exemplary image in black and white. It is an SEM scan that evokes a fully three-dimensional tissuescape, magnified about 1,000 times. We see some 50 plump human cells at the surface of a tiny cancer of the human bronchus, the disease fatal to so many smokers. Disorganized, these thriving malignant cells, their primitive surfaces smooth and bare, seem to be shouldering their heedless, incompetent way among the normal cells, whose furry surfaces bristle with functional hairlike cilia that move rhythmically to cleanse dust and bacteria out of the innermost cavities of the lung. One can all but hear a rasping cough.

Another photograph, made by a high-resolution, high-voltage transmission electron microscope, shows six atoms of uranium arrayed in a neat hexagon around another central uranium atom, the entire assembly ensconced within a minute microcrystal of the compound uranyl acetate. The 100-odd light atoms that complete the crystal structure do not scatter the electrons of the beam enough to show. It is a fine image at the uncommon magnification of 120 million, although its bright colors are merely computer-made, a showy mapping of photographic density, no more meaningful than was the red on the map of that empire on which the sun never set. The free use of such false color is catchy as long as it does not mislead; the captions here are explicit.

Two of the authors and most of the images in this book come from the Science Photo Library of London. The credit for these archived pictures is laconic: it consists only of the name of the photographer. The reader learns neither place nor date and must surmise purpose; a listing intended to inform, and not merely to meet the standards of the copyright law, would add much to the value of such a scientific album, at the cost of only a few pages of type. The occasional full reference to a book or a paper appearing among all those repeated names is as welcome as spring rain. A readable appendix, helpful and not very technical, reviews the instruments, their history and how they work, and teases our viewing with a few pictures through some of the new specialized windows into the microcosm, such as the scanning tunneling microscope. A needed and heartfelt essay on the high importance of sample preparation in microscopy is included.

Time stops in the photographs of Harold Edgerton for dancers and bullets and bands and glassy plankton and splashes that look like tiny coronets. The fast flash and the swift optical shutter—flicking at speeds as high as hundredths of a microsecond—show us the microstructure of events. This volume, presenting some fine new images according to a familiar plan, reviews the photographic techniques and work of Edgerton, the M.I.T. electrical engineer who developed the first electronic gas flash tubes in about 1928 and has kept at it ever since. The present beautiful volume presents about 125 pictures, a couple of dozen of them in color; selection and commentary seek to place the man and his work as part of the history of the photographic art. Doc Edgerton is a celebrated artist-teacher, as wise a man as you can find, steadily inventive over six decades, free of pretense and invariably warm, generous and helpful.

It is he who more than anyone else has visibly sliced William James's "thickened *Now*" to nourish readers far outside the limits of science and technology. This good-humored engineer from Nebraska, long and hard at work among the sober Yankees, remains enough of countryman and poet to express precisely the reason people so enjoy his pictures: "Anything that moves is my game. It's always beautiful."

First example: Indian Club Demonstration, 1939. Here are two images: a time exposure that ties a shapely but blurred knot of light, and next to it Edgerton's hundredth-of-a-second multiflash that unties that same knot into a big many-petaled flower. No wonder the painter-photographer from the Bauhaus, László Moholy-Nagy, traded one of his own prints with Doc for that one. Second example: an exquisite flying fish floats past, frail fins spread, revealed as no flier at all but a stiff-winged glider, tail-launched out of the sea off Catalina. Third, from the 1970's: a pair of interference-tinted vortexes of heated air swirl elegantly in rainbow unison as a Bunsen flame is cut by a fastspinning fan blade; all that coupled movement is stilled by flash. This one-man show (with a little help from devoted friends) is a retrospective out of which some of the most prized sensibilities of our epoch unfold.

METEORITE CRATERS, by Kathleen Mark. The University of Arizona Press ($29.95).

They fly in from the astronomical void to crash into hard-rock geology, blasting out strange landforms. In this brief and vivid narrative, Kathleen Mark, a delighted, careful geologist long resident at Los Alamos, leads us through an intriguing history. We learn how "the investigation of some holes in the ground" led to the recognition of great natural explosions underground, innocent of volcanic ash or lava, the solid crust punctured not from below but from on high.

Not far south of the Zuni Indian Reservation in New Mexico a nearly circular salt lake a mile across shines blue within its slightly upturned rim. It is called Zuni Salt Lake. A couple of black cinder cones rise from the lake, whose ashy edge is whitened by drying salt. It was beyond the understanding of the first geologist to visit, back in 1873. Now we know it as a maar, a volcanic explosion pipe, the remains of a few brief explosive bursts of gas that rose from magma below to shatter the rock surface, and build up the rim and the cinder cones out of small amounts of volcanic debris. The lava below then retreated, and a lake fed by saline springs slowly filled the crater. In 1955 such a maar was seen to form in the Chilean Andes as repeated violent half-hour discharges of gas erupted on and off through an entire season until they gradually dwindled away.

Only a couple of hours' hard driving west from Zuni Lake brings a traveler to the brink of a deep, dry, almost circular hole, almost as large as the lake. Once it was called Coon Butte. It is strangely free of volcanic materials, its rim of limestone and sandstone strata slopes away on all sides, and plenty of nickel-iron meteoritic fragments can be found nearby. The redoubtable G. K. Gilbert, chief geologist of the U.S. Geological Survey before the turn of the century, made of it an example of the method of scientific inference. On balance he concluded that it was the volcano-born result of a steam explosion, although he went on to ascribe the lunar craters to meteorite impact. He had searched for signs of the impacting body, but as he wrote to a friend, "I did not find the star because she is not there."

The formation is now known as Meteor Crater; the name has been used since 1906, when mining engineer D. M. Barringer (and others who amplified his arguments) spoke strongly for its meteoritic origin. Indeed, for the rest of his life Barringer searched the site for a buried starry messenger of nickel-iron: he sank two dozen expensive drill holes within the crater and found nothing.

Perhaps because of the gap between the astronomers and the geologists the governing energetics of impact was grasped very slowly. Any big falling star is well and truly vaporized; its shock waves pass swiftly and powerfully radially out through the rocks to excavate a telltale hole, circular in shape whatever the direction of impact. The ideal circle is somewhat vulnerable to retouching by the spew of rock that flies up; rebound can even throw up a central cone in a big crater, as it does on the moon. By 1930 or so workers recognized the distinction between explosive shock events and the smaller, slower percussive falls that dig odd-shaped pits, meteorite chunks left within.

At mid-century the literature listed only a dozen meteorite craters on five continents, all confirmed in the most direct way by the presence of meteoritic material. The chance wartime sighting of the remarkably round lake now called the New Quebec Crater led to a search among the millions of images held in the Canadian Air Photo Library in 1955. Dozens of fossil craters, big and little, show up on the ancient bared rocks of the Canadian Shield; there are plenty of examples, including the circular eastern edge of Hudson Bay. No diamond drill ever found a buried star, but deep below the crater often lay a distinctive lens of heated and broken rock. Artificial craters were made in test nuclear explosions; their profiles are drawn here for direct comparison with Meteor Crater. A fascinating set of sketches presents the sequence of events in the growth of a big impact crater, a sequence well understood by the 1960's.

It was the physics of high pressure that produced the strongest positive evidence. Step by step it became clear that the transient high pressures applied by the strong passing shock could transform the medium; the first transformation recognized was a new heavy phase of silica, coesite (made from quartz by Loring Coes in 1953). By the 1970's a variety of products of such shock metamorphism were recognized, including shatter cones. These cones are forms of fractured rock that indicate not only the presence of very high pressure but also the direction taken by the pressure wave. Now the craters could be searched for evidence of shocks from above, shocks too strong to come from the low-energy density of merely molten magma.

The laboratory work illuminated observations in the field. The unique geologic structures of the Bushveld Igneous Complex in the Transvaal were tentatively made out in 1975 as four simultaneous giant deep impacts, extremely old and much modified, stretching over a couple of hundred miles of mountainous landscape. The rich nickel-mining district at Sudbury in Ontario is another large puzzle: there are impact signs in plenty, but no ring is left. The text ends with the glassy tektites, seen as upthrown molten debris of impact, formed by hundreds of miles of flight through the atmosphere. This is a narrative of light slowly penetrating the shadows; everything is not yet clear.

The latest reference among 400 to 500 is from 1979; the chronicle ends then, surely by design. In 1980 a new period began. The Alvarezes reported meteor-borne iridium in the thin clay layer within strata formed at the time of the dying of the great lizards. By now it is at least arguable that the direction of organic evolution is marked as deeply by giant impacts as is the desert of Arizona: meteoritic biology. This text is a model of clear nontechnical exposition and apt illustration. In a couple of decades another volume as good as this will be needed to survey the impact-extinction story, however it may turn out.

Not every new wonder endures. A French official reported in about 1920 that in the Adrar desert in Mauritania there lies a "compact, unfractured" iron mass 100 meters long and 50 meters high, half-buried in the sand, its visible face a vertical cliff of metal polished here and there to a mirror surface by the sand-filled winds. That Chinguetti mass, as elusive as the fallen star Gilbert sought in Arizona, has never again been found. There is a 10-pound sample, and a good-size real crater (without a meteorite) is known from the same region; the rest is fiction—or is an all too massive iron hidden under drifting sands?

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

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 flat fee of $1.25 per copy of each article or any part thereof. 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., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/88. $1.25 + 0.00.

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

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

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 953, Farmingdale, N.Y. 11737. Telephone inquiries: 1-800-227-5782. The date of the last issue on subscriptions appears at the right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ____________________

New Address

Street ____________________

City ____________________

State and ZIP ____________________

Old Address

Street ____________________

City ____________________

State and ZIP ____________________

BIBLIOGRAPHY

Readers interested in further exploration of the subjects covered by the articles in this issue may find the following lists of publications helpful.

THE NONACOUSTIC DETECTION OF SUBMARINES

ANTI-SUBMARINE WARFARE AND STRATEGIC STABILITY. Donald C. Daniel. University of Illinois Press, 1986.

STRATEGIC ANTISUBMARINE WARFARE AND NAVAL STRATEGY. Tom A. Stefanick. Lexington Books, 1987.

COMPARISON OF A NEW RADAR OCEAN-IMAGING MODEL WITH SARSEX INTERNAL-WAVE IMAGE DATA. Dennis Holliday, Gaetan St. Cyr and Nancy E. Woods in *International Journal of Remote Sensing,* Vol. 8, No. 9, pages 1423-1430; 1987.

GRAVITY AND ANTIMATTER

GRAVITATION AND SPACETIME. Hans C. Ohanian. W. W. Norton & Co., 1976.

EXPERIMENTS TO MEASURE THE GRAVITATIONAL ACCELERATION OF ANTIMATTER. T. Goldman and Michael Martin Nieto in *Physics Letters,* Vol. 112B, No. 6, pages 437-440; May 27, 1982.

ELEMENTARY PARTICLES AND THE LAWS OF PHYSICS (THE 1986 DIRAC MEMORIAL LECTURES). Richard P. Feynman. Cambridge University Press, [1987].

GEOPHYSICS AND THE LAW OF GRAVITY. F. D. Stacey, G. J. Tuck, G. I. Moore, S. C. Holding, B. D. Goodwin and R. Zhou in *Reviews of Modern Physics,* Vol. 59, No. 1, pages 157-174; January, 1987.

CATALYTIC ANTIBODIES

ENZYMATIC CATALYSIS AND TRANSITION-STATE THEORY. Gustav E. Lienhard in *Science,* Vol. 180, No. 4082, pages 149-154; April 13, 1973.

TRANSITION STATES IN BIOCHEMICAL PROCESSES. Edited by Richard D. Gandour and Richard L. Schowen. Plenum Press, 1978.

ENZYME STRUCTURE AND MECHANISM. Alan Fersht. W. H. Freeman and Company, 1985.

CATALYTIC ANTIBODIES. Alfonso Tramontano, Kim D. Janda and Richard A. Lerner in *Science,* Vol. 234, No. 4783, pages 1566-1570; December 19, 1986.

A STEREOSPECIFIC CYCLIZATION CATALYZED BY AN ANTIBODY. Andrew D. Napper, Stephen J. Benkovic, Alfonso Tramontano and Richard A. Lerner in *Science,* Vol. 237, No. 4818, pages 1041-1043; August 28, 1987.

CONSTRUCTION CRANES

CRANES AND DERRICKS. Howard I. Shapiro. McGraw-Hill Book Company, 1980.

USA CRANE MARKET SURVEY. In *Cranes Today,* No. 144, pages 21-56; August, 1985.

WHEELS OF POWER. In *Cranes Today,* No. 150, pages 53-54; March, 1986.

TRAVELLING CRANES ON THE HIGHWAY. In *Transportation Engineer,* Vol. 34, No. 3, pages 39-45; May/June, 1987.

HOW THE LEOPARD GETS ITS SPOTS

A UNITY UNDERLYING THE DIFFERENT ZEBRA STRIPING PATTERNS. Jonathan B. L. Bard in *Journal of Zoology,* Vol. 183, Part 4, pages 527-539; 1977.

A PRE-PATTERN FORMATION MECHANISM FOR ANIMAL COAT MARKINGS. J. D. Murray in *Journal of Theoretical Biology,* Vol. 88, No. 1, pages 161-199; 1981.

ON PATTERN FORMATION MECHANISMS FOR LEPIDOPTERAN WING PATTERNS AND MAMMALIAN COAT PATTERNS. J. D. Murray in *Philosophical Transactions of the Royal Society of London, Series B,* Vol. 295, No. 1078, pages 473-496; October 7, 1981.

A NEW APPROACH TO THE GENERATION OF PATTERN AND FORM IN EMBRYOLOGY. J. D. Murray and P. K. Maini in *Science Progress,* Vol. 70, No. 280, Part 4, pages 539-553; 1986.

FATNESS AND FERTILITY

MENSTRUAL CYCLES: FATNESS AS A DETERMINANT OF MINIMUM WEIGHT FOR HEIGHT NECESSARY FOR THEIR MAINTENANCE OR ONSET. Rose E. Frisch and Janet W. McArthur in *Science,* Vol. 185, No. 4155, pages 949-951; September 13, 1974.

HYPOTHALAMIC DYSFUNCTION IN SECONDARY AMENORRHEA ASSOCIATED WITH SIMPLE WEIGHT LOSS. Robert A. Vigersky, Arnold E. Andersen, Ronald H. Thompson and D. Lynn Loriaux in *The New England Journal of Medicine,* Vol. 297, No. 21, pages 1141-1145; November 24, 1977.

POPULATION, FOOD INTAKE, AND FERTILITY. Rose E. Frisch in *Science,* Vol. 199, No. 4324, pages 22-30; January 6, 1978.

EFFECTS OF PHYSICAL ACTIVITY ON REPRODUCTIVE FUNCTION AND DEVELOPMENT IN MALES. Stephen R. Wall and David C. Cumming in *Seminars in Reproductive Endocrinology,* Vol. 3, No. 1, pages 65–80; 1985.

FATNESS, MENARCHE, AND FEMALE FERTILITY. Rose E. Frisch in *Perspectives in Biology and Medicine,* Vol. 28, No. 4, pages 611–633; Summer, 1985.

THE QUANTUM-EFFECT DEVICE: TOMORROW'S TRANSISTOR?

LIMITS TO PERFORMANCE OF VLSI CIRCUITS. R. T. Bate in *VLSI Handbook,* edited by Norman G. Einspruch. Academic Press, 1985.

REPORT ON ARTIFICIALLY STRUCTURED MATERIALS. National Research Council. National Academy Press, 1985.

PROSPECTS FOR QUANTUM INTEGRATED CIRCUITS. R. T. Bate, G. A. Frazier, W. R. Frensley, J. W. Lee and M. A. Reed in *Proceedings of the SPIE* (International Society for Optical Engineering), Vol. 792, pages 26–35; 1987.

OBSERVATION OF DISCRETE ELECTRONIC STATES IN A ZERO-DIMENSIONAL SEMICONDUCTOR NANOSTRUCTURE. M. A. Reed, J. N. Randall, R. J. Aggarwal, R. J. Matyi, T. M. Moore and A. E. Wetsel in *Physical Review Letters,* in press.

THE ROMAN PORT OF COSA

THE ANCIENT COMMERCIAL AMPHORA. Elizabeth Lyding Will in *Archaeology,* Vol. 30, No. 4, pages 264–270; July, 1977.

ANCIENT TRADE AND SOCIETY. Lionel Casson. Wayne State University Press, 1984.

GREEK AND ROMAN MECHANICAL WATER-LIFTING DEVICES: THE HISTORY OF A TECHNOLOGY. John P. Oleson. University of Toronto Press, 1984.

THE ROMAN PORT AND FISHERY OF COSA: A CENTER OF ANCIENT TRADE. Anna M. McCann et al. Princeton University Press, 1985.

THE AMATEUR SCIENTIST

BOWLING FRAMES: PATHS OF A BOWLING BALL. D. C. Hopkins and J. D. Patterson in *American Journal of Physics,* Vol. 45, No. 3, pages 263–266; March, 1977.

COMPUTER RECREATIONS

MICROCOMPUTER DISPLAYS, GRAPHICS, AND ANIMATION. Bruce A. Artwick. Prentice-Hall, Inc., 1985.


Caprice.
No luxury car is
so big for so little.
Brougham
LS


Let's get it together... buckle up.
GM 6/60
QUALITY COMMITMENT PLAN