SCIENTIFIC AMERICAN

MEGALITHIC MONUMENTS

$2.00

July 1980

Consumer Orientation

No. 6 in a Series
of Technical Papers

6

Subject:
The Blank Paper Principle.
Genesis of Porsche's
Newest and Best.

The Porsche 928 began as a blank piece of paper. There were no rules stating that what had been done before must be done again. Instead, there was an objective: To build from scratch the very best Porsche to meet the dreams and realities of the balance of this century. What emerged is a sports car that combines the aggregates of power, handling, comfort, and luxury into one—as no Porsche ever has. Priced at more than $35,000, the 928 is the newest and finest Porsche ever built.

The 928's engine is the first Porsche production V-8. It is made of lightweight but highly durable aluminum alloy. Front-mounted, liquid-cooled, fuel-injected, displacing 4.5 liters, it produces 220 bhp at 5500 rpm and generates 265 ft-lbs of torque at 4000 rpm.

On the track at Weissach, the 928 accelerates from 0 to 60 mph in 7.5 seconds. It covers the quarter mile from a standing start in 15.5 seconds. Its speed at the ¼-mile mark is 93 mph. Its maximum speed: 140 mph.

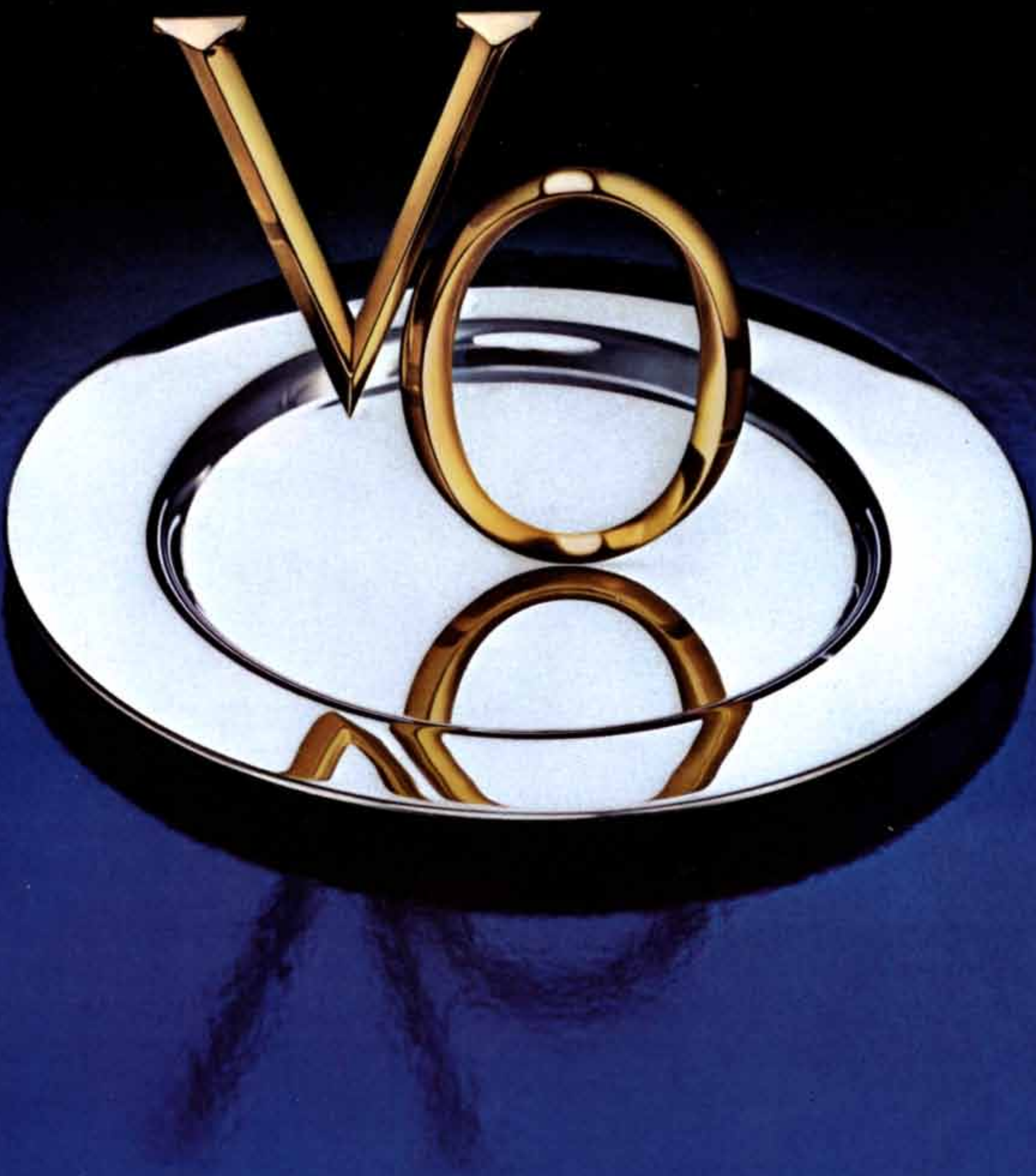
The 928 has the unique Porsche transaxle design which places the engine in front and transmission in back. By mounting these two main inertial masses at opposite ends of the car, the transaxle produces an almost perfectly-equal 50-50 weight distribution between the front and rear wheels for balanced braking and improved cornering. And it results in a high polar moment of inertia that reduces pitching and increases directional control.

The 928 also has independent suspension with coil springs on all 4 wheels. And it introduces the unique Weissach rear axle.

With conventional cars, the rear wheels toe-out during braking or deceleration. But with the 928's Weissach axle, a kinematic effect changes toe-out to toe-in in no more than 0.2 seconds during braking or deceleration. It eliminates rear-wheel steering and improves directional control.

Porsche 928

Porsche+Audi
Nothing
Even Comes
Close



Established 1845 **SCIENTIFIC AMERICAN** July 1980 Volume 243 Number 1

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (USPS 509-530), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1980 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. CONTROLLED CIRCULATION POSTAGE PAID AT NEW YORK, N.Y. (USPS 509-530), AND AT CHICAGO, ILL. (USPS 508-030). AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR, U.S., ITS POSSESSIONS AND CANADA; $27 PER YEAR, ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017

THE COVER

The photograph on the cover shows three of the 12 original standing stones of the megalithic monument at Stenness on the island of Orkney. This megalithic array is 102 feet in diameter and is surrounded by a ditch more than 197 feet in diameter, produced by the quarrying of an estimated 1,300 tons of sandstone. Unlike the stones at Stonehenge, the most famous megalithic ring monument, the stones at Stenness were not shaped uniformly after they had been quarried. The four that remain upright, however, are all about the same height: 16 feet. Carbon-14 analysis of wood and bone samples associated with the ring monument indicates it was put up late in the third millennium B.C., or somewhat earlier than the second phase of construction at Stonehenge in late Neolithic times (see "Megalithic Monuments," by Glyn Daniel, page 78).

THE ILLUSTRATIONS

Cover photograph by Evan Hadingham

Page	Source
18–31	Gabor Kiss
44–51	Andrew Tomko
53	David Koffler, Hahnemann Medical College
54–58	Bunji Tagawa
59	David Koffler, Hahnemann Medical College
60–61	Bunji Tagawa
63	Marvin Leventhal, Bell Laboratories
64–70	Ilil Arbel
79	Cliché Lapie-Photothèque (*top*), West Air Photography (*bottom*)
80	Patricia J. Wynne
81	Evan Hadingham
84	Photograph © Françoise Legrand
85	Patricia J. Wynne
86	Roger Viollet (*top*), Irish Tourist Board (*bottom*)
87–88	Evan Hadingham
90	West Air Photography
92	Clemente Simon, Spainfo Ingenieros, S.A.
94–100	Alan D. Iselin
103	H. Linde, Central Institute for Physical Chemistry, Berlin
104–105	Alan D. Iselin
111	Keith Gillett, Animals, Animals
113–116	Tom Prentiss
119	Gary Cohen, State University of New York at Stony Brook
120–132	George V. Kelvin
133	Jerome Vinograd, California Institute of Technology
135	Elsam, Mann and Cooper, Ltd. (*top*), Ralph Morse (*bottom*)
137–140	Dan Todd
141	F. M. N. Brown, University of Notre Dame (*top*); Dan Todd (*bottom*)
142	Steven Finberg
143	Don Monroe
144	Dan Todd
147	Justin E. Kerwin, Massachusetts Institute of Technology
148	Lewis Research Center, National Aeronautics and Space Administration
151–161	Michael Goodman

LETTERS

Sirs:

I was very pleased to see a reference to my study of early Sumerian and proto-Elamite metrology and number systems in the note "From Reckoning to Writing" ["Science and the Citizen," SCIENTIFIC AMERICAN, March], where the study was related to the important discoveries presented by Denise Schmandt-Besserat in her article "The Earliest Precursor of Writing" [SCIENTIFIC AMERICAN, June, 1978]. I should like to add a few completing and clarifying remarks.

It is true that the "proto-Sumerians" of what is now Iraq, as well as the "proto-Elamites" of Iran, used on their clay account tablets, written before 3000 B.C., several different number systems simultaneously, but so did the "classical" Sumerians several centuries later: one system for counting, another for capacity measures, a third for area measures and so on. (The Sumerian script and language are today relatively well known to professional Sumerologists, but the proto-Sumerian tablets can be read only in a very rudimentary way, and the proto-Elamite script has until now been almost completely undecipherable.) Due to the fact that some of the account tablets contain quite complicated computations, it is possible to decipher exclusively the mathematical and metrological parts of the inscriptions.

It turns out that the proto-Sumerians and the proto-Elamites, in spite of the fact that their scripts and languages seem to have been almost totally unrelated, made use of essentially identical number systems and metrological systems, different in several respects from the systems used later by the Sumerians. The only difference between the proto-Sumerian and proto-Elamite systems seems to have been that the proto-Sumerians made use of a "protosexagesimal" system for counting people, animals and things, whereas the proto-Elamites used the protosexagesimal system to count people and things but a decimal system to count animals. As a matter of fact, there are many examples of proto-Elamite account tablets of a special type, on which the protosexagesimal and decimal systems are used at the same time, with conversions from one to the other. So much for the controversy about which number system is older, the sexagesimal or the decimal!

Another remarkable fact is that a comparative analysis identifying certain types of proto-Sumerian and proto-Elamite texts ("bread-and-beer accounts," "seed-grain accounts" and so on) with their much later Sumerian counterparts makes it possible to determine quite accurately the "absolute values" of, for example, the capacity-measure units used in Sumer and Elam in the latter half of the fourth millennium B.C. This means that we can deduce from the account tablets of those extremely remote times such things as how big the herds of animals were, how big the food rations were and how much grain was stored in the granaries.

The determination of the absolute values of the early capacity-measure units mentioned in your account is interesting also because it leads to the surprising conclusion that modern (although premetric) sequences of capacity measures such as the Anglo-Saxon pint, peck, bushel and barrel are direct descendants of their counterparts on the earliest-known written records: the clay tablets of ancient Sumer and Elam.

JÖRAN FRIBERG

Chalmers University of Technology
and University of Göteborg
Göteborg, Sweden

Sirs:

In his interesting article [SCIENTIFIC AMERICAN, December, 1979] on tephra, Laurence R. Kittleman, in discussing the ecological effects of tephra, notes that "not all the effects of tephra are harmful," and that "repeated tephra falls in the Tropics probably renew the fertility of the soil, which otherwise would be quickly leached of nutrients in the prevailing climate."

I feel that Dr. Kittleman understates his case. One of the most remarkable ecosystems in the world owes much of its character to the tephra mantle that covers it. This is the Serengeti National Park and its ecological unit in northern Tanzania, which is famous for its landscapes of rolling grassland and acacia parkland, its migrations of more than two million game animals and its abundance of fossil hominids.

The Serengeti ecological unit is bounded on its southeastern side by the Ngorongoro Crater Highlands, which consist of the joined cones of at least six major volcanoes in addition to the Ngorongoro caldera and the still active Oldoinyo Lengai. It has long been realized that the open Serengeti plains to the southeast of the Serengeti ecosystem, and lying at the foot of the Crater Highlands, are covered by a series of layers of volcanic ash, through which the underlying shield protrudes in a series of scenically striking kopjes. The sequence of ash falls has been described in detail by R. L. Hay in relation to Olduvai Gorge, and their influence on the soils of the Serengeti plains has been examined by Hugo de Wit of the University of Wageningen.

It has only recently emerged, however, as a result of the work of Tjapko Jager, also of Wageningen, that the western and northern parts of the Serengeti ecosystem have likewise received

SCIENCE/SCOPE

Novel digital logic circuits employing charge-coupled devices (CCDs) may soon be used in a wide range of military systems, including communications, radar, voice processing, sonar, and guidance. Experimental chips developed by Hughes are five times more compact than similar circuits made with I^2L (integrated injection logic) or CMOS (complementary metal oxide semiconductor) processes. They also can provide up to 10 times the throughput per unit power when structured to perform many different logic operations at the same time.

Indonesia's new Palapa B communications satellites will use flight-proven technology to ensure their eight years of planned service. The spacecraft are modeled after Canada's Anik C and Anik D series, as well as the Satellite Business Systems spacecraft that will serve U.S. businesses. Like the others, the Palapa B craft will have outer cylindrical sleeves of solar cells that will deploy in space. The arrangement nearly doubles the area of cells usable with the same size satellite, thereby increasing total power. Hughes, under contract to Perumtel, Indonesia's government-owned telecommunications company, is building two spacecraft and associated equipment to augment certain ground stations.

Once a laboratory curiosity, the laser is now a workaday tool of immense value. Perhaps the most common use of the laser is in alignment. Construction crews use lasers for surveying and grading land, laying sewer pipe, installing acoustical ceilings, and squaring off buildings. Lasers also help align sawmill blades, guide the cutting of marble slabs, and position patients for medical procedures. Hughes, which achieved the world's first laser action in 1960, is one of the world's largest suppliers of helium-neon lasers for these uses.

Hughes Missile Systems Group, located in Canoga Park, California, is seeking engineers and scientists to work on a growing list of development and production programs. The list includes AMRAAM, Wasp, multimode guidance, TOW, Phoenix, Maverick, and U.S. Roland. Typical openings are in areas of LSI, radars, IR systems, signal processing, pattern recognition, computer software, electronic components, guidance and controls, gyro-stabilized platforms, and digital systems. Please send resume to Hughes Engineering Employment, Dept. SE, Fallbrook at Roscoe, Canoga Park, CA 91304. Equal opportunity M/F/HC.

A new computer system is smoothing out snags in the apparel industry and improving efficiency by monitoring the progress of every production step. The Hughes PC-1 system schedules and tracks production after being programmed with information regarding styles, sizes, quantities to be cut from a bolt of fabric, and the time to do certain operations. It ensures, for example, that one group completes the correct number of sleeves in time for another group to sew them onto jackets. By knowing exact production status and by keeping an inventory of finished and unfinished garments, the system helps management balance production with the work force, as well as predict realistic delivery dates.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY
CULVER CITY, CALIFORNIA 90230

tephra falls from the Crater Highlands. The distances involved (up to 200 kilometers) are well within the ranges mentioned by Kittleman.

The ecological influence of these tephra falls has been profound. It can be seen by comparing the plant and animal communities of the Serengeti ecosystem with those of the Tanzanian shield where it emerges from the tephra mantle to the southwest. The Serengeti ecosystem consists of short, productive grassland, with open, fine-leaved woodland and high densities of medium to small herbivores. The Tanzanian shield is covered by the *Brachystegia* woodland and dambo landscape characteristic of the ancient land surface of the Central African Plateau. Herbivore densities are low, and most of the animal biomass is in large animals.

The mechanism of the difference brought about by tephra falls probably depends first on the fact that the ash has a high content of biologically active substances (calcium, sodium, magnesium, carbonates and so on), which give rise to the production of high-quality vegetation, and second on the effect of the tephra mantle on soil-water dynamics; that is, as has been documented by de Wit and Jager, the tephra tends to give rise to soil horizons that impede water infiltration. This effect favors grassland as opposed to woodland and short grass as opposed to tall grass; in fact, it favors the production of protein as opposed to the production of structural carbohydrate. This combination of factors goes a long way toward explaining the short-grass parklands and the high densities of smaller herbivores characteristic of the Serengeti.

The influence of climate, particularly rainfall, on African wildlife communities has been well documented. For example, the dependence of the Serengeti migrations on the local rainfall pattern is well established [see "A Grazing Ecosystem in the Serengeti," by Richard H. V. Bell; SCIENTIFIC AMERICAN, July, 1971], and several other workers have pointed out relations between wildlife biomasses and rainfall (for example R. M. Watson, and M. Coe, D. Cumming and J. Phillipson). My point here, however, is to emphasize that the geomorphological situation is at least as important an influence on African wildlife community structures as climate, and that in the Serengeti ecosystem the key factor is the mantle of tephra.

RICHARD H. V. BELL

Department of National Parks and Wildlife
Kasungu, Malawi

Sirs:

I am writing you with regard to the article "Chemical Warfare and Chemical Disarmament," by Matthew Meselson and Julian Perry Robinson [SCIENTIFIC AMERICAN, April].

The photograph of the chemical storage area at Tooele Army Depot that appears on pages 38 and 39 misrepresents the facts. The caption under the photograph implies that nerve agents are stored in the open at Tooele Army Depot. This is not so. All the ton containers of agent GB and the spray tanks containing VX were removed from open storage prior to May 10, 1978. The ton containers are now stored in earth-covered bunkers and the spray tanks in concrete sheds. These actions were taken as part of the ongoing security upgrading at all U.S. Army DARCOM chemical storage sites, including Tooele Army Depot. In addition a command policy in effect since February, 1978, prohibits the taking of photographs in chemical storage areas. In effect this means that the photograph you used is at a minimum two years old.

It is regrettable that an article of such importance and as well written could not have been accompanied by photographs with the same credibility.

CHARLES A. HAMMAKER, JR.

Colonel, GS
Chief of Public Affairs
Headquarters U.S. Army Materiel Development and Readiness Command
Alexandria, Va.

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

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Offprint correspondence and orders should be addressed to W. H. Freeman and Company, 660 Market Street, San Francisco, Calif. 94104. For each offprint ordered please enclose 40 cents.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. For change of address, notify us at least four weeks in advance. Send both old and new addresses and enclose an address imprint from a recent issue. (Date of last issue on your subscription is shown at upper right-hand corner of each month's mailing label.)

Name ____________________

New Address ____________________

____

____

Old Address ____________________

____

____

Before there was any other 35, there was Leica.

In every important way, there is still only Leica.

The great Leicas of the past are still eagerly sought after. Not only for their unique appreciative value, but for their ability to create great photographs. This quality in craftsmanship bridges the generations even today. At the same time, Leica lenses have formed an equal and complimentary legend of unsurpassed quality. The Leica® R3 MOT, the Leica of today, with its automation, its whisper-quiet steplessly variable shutter and dual-metering system and its fast near-silent full-motor drive, will see proud service into the next century.

Leica, you see, did more than invent the 35mm still camera. It perfected it. Check the adjacent column for dealer nearest you. Or write for more information.

Leitz means precision. Worldwide.

E. Leitz, Inc., Rockleigh, N.J. 07647
Toll-free (800) 631-2504

1930A

PRADOLUX RT 300.
The Leica of slide projectors.

Now, cover a wall with bright, sharp Leica optical quality. Rugged, high performance construction is compatible with round trays.

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

JULY, 1930: "The new Lawrence Lowell telescope at the Lowell Observatory in Flagstaff, Ariz., has within a year of its installation brought to light the long-sought Planet X. Photographs of moderate exposure covering the region of the heavens around the predicted position were obtained, and on one of them, taken on January 21, Mr. Clyde W. Tombaugh of the Observatory staff found 'a very promising object.' It was identified from its motion past the numerous fixed stars on plates of the same star field being viewed in the blink comparator. This showed that one faint object among many thousands had shifted its place by a certain expected order of distance in the interval between the taking of the two plates. Since that date the object has been carefully followed both photographically by Dr. Carl O. Lampland with the 40-inch reflector and visually by E. C. Slipher and other members of the staff. Its motion in the heavens has been just what might be expected of a trans-Neptunian planet at about the distance predicted by Percival Lowell, and its longitude agrees closely with his predictions. There is no doubt that it is actually a new planet much farther away than any previously known."

"With a production of more than 5,000,000 motor vehicles and a total registration of 26,400,000 in this country last year, and a steadily increasing production each year, the need for more and better roads is impressively indicated. Of the total 3,016,281 miles of highway in the United States only 660,000 are surfaced—slightly more than 20 per cent. Fortunately our people are highway-conscious, and it is certain therefore that the percentage will rapidly mount. This is borne out by the fact that Congress passed this year, with rare speed, a bill for adding $50,000,000 to the present $75,000,000 that is given the states to build highways in the Federal Aid Program."

"Dr. William C. Geer's 'airplane overshoes' give promise of conquering one of the most serious hazards of flying. The adhesion of ice to metal—polished aluminum, for example—is high. In previous attempts to solve the problem the surface of the airplane wing has been covered with oil. The oils used generally become viscous and sticky at freezing temperatures, and moreover the wind forces scrub them off. Profiting by these early mistakes, Dr. Geer has invented a simple mechanical device. It has the form of a light-weight thin rubber 'overshoe' for the leading edge of the wing, with a fabric backing for strength. At the leading edge of the overshoe is an air tube strengthened with an extensible fabric. This inner tube is connected to a pump, either motor- or hand-driven. If ice forms, the pilot turns the air into the tube, thus slightly expanding it. This breaks the ice and removes it. In a test run in Cleveland in March an overshoe was attached to a radio mast behind the cockpit. During the flight ice formed to a thickness of about half an inch over the leading edge. A hand pump was used to expand the tube. When a pressure of two pounds per square inch was applied, the ice suddenly left the overshoe, flying off in chunks. Experiments on a wing and on a strut were also successful."

Scientific American

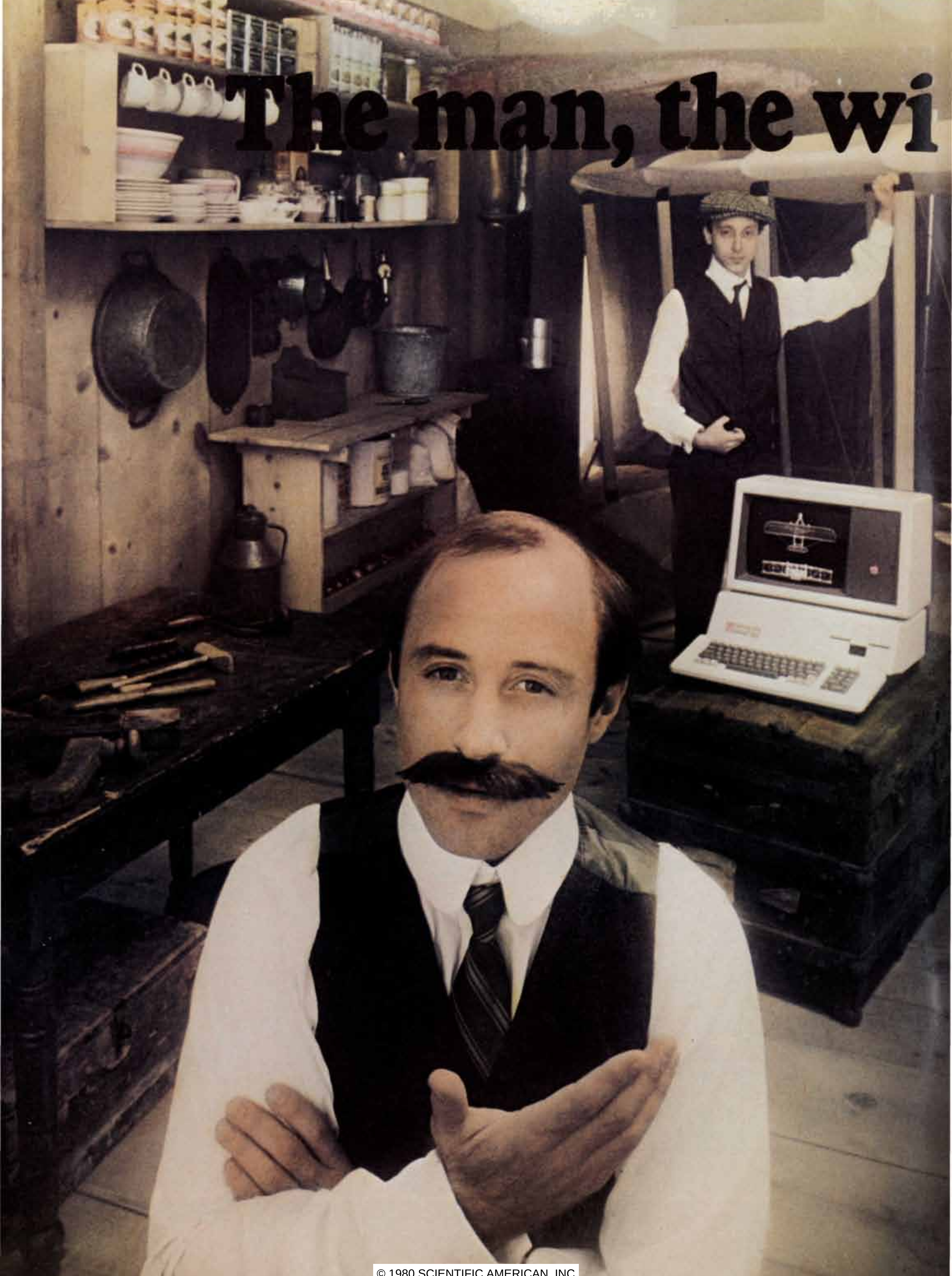
JULY, 1880: "It is a curious circumstance that while from the nonscientific point of view the unpardonable fault of modern science is its 'materialistic' tendency, the actual drift of scientific thought is toward eliminating from the scientific idea of matter everything that answers to the popular notion of it. As our readers are aware, Professor Crookes claims to have demonstrated an ultra-gaseous 'fourth state' of matter, as unlike the other three recognized states of matter as they are unlike one another. The fourth state, according to Professor Crookes, obtains when the gas has been so rarefied that the collisions of the molecules are few compared with the misses. If we follow in our imagination the free molecule in its flight, it loses all known properties of matter and becomes as if it did not exist. For what is a single free molecule in space? Is it a solid, a liquid or a gas? Professor Crookes answers: 'Solid it cannot be, because the idea of solidity involves certain properties that are absent in the isolated molecule. And if the individual molecule is not solid, *a fortiori* it cannot be a liquid or a gas, for these states are even more due to intermolecular collisions than is the solid state.' What we call matter is nothing more than the effect upon our senses of the movements of molecules. From this point of view, then, matter is but a mode of motion."

"The extract from the report of the judges in horology at the Sydney International Exhibition cannot fail to interest our readers. There were 10 exhibitors, and the inherent and comparative merits of the various watches exhibited were rated under 10 heads on the basis of 100 points 'for the highest degree of excellence.' There were British, German, French, Swiss and American competitors, and while the scores of the nine European exhibitors footed up totals ranging from 76 to 686, the total of the Waltham Watch Company was 981. It is the bringing of machinery to every branch of watch-making that is enabling Americans to beat the world in this as well as in many other things. The Americans are showing that they can make better watches than the Swiss or the British, but what is of equal importance, they are showing that they can make them for less money."

"The piercing of the Saint Gothard tunnel in the Alps, the longest tunnel in the world, has been accomplished in seven years five months, the rapidity of its execution being unprecedented in works of this kind. Relative to its length the Saint Gothard has been bored in one fourth the time that was occupied in the boring of the Hanenstein tunnel, and in less than half the time taken to pierce Mont Cenis. The great advance in the art of tunnel driving is due to the improvements in the boring machinery, and particularly to the efficiency of the air compressors invented by Professor Daniel Colladon."

"The extent of the manufacture of coal-tar colors is shown by the following statistics of labor and production at one of the principal coal-tar color works in Germany. There are employed more than 1,000 workmen, in addition to 40 overlookers and branch managers, 25 chemists, one engineer and 30 clerks and accountants. The yearly consumption of coal amounts to 17,000,000 kilograms; anthracene, 825,000; naphtha and benzol, 950,000; chromate of potash, 280,000; caustic soda, 1,245,000; sulphuric acid, 2,250,000; muriatic acid, 4,050,000; nitric acid, 825,000; alcohol, 91,500, and sundry other chemicals, 3,560,000."

"The terrible visitation of yellow fever on the cities along the lower Mississippi has indicated clearly to the United States Medical Department the great need of a better system of quarantine regulation, inspection and disinfection and the want of swift, properly appointed craft to relieve passing vessels of sick persons and to convey them to the quarantine stations along the river. To meet the peculiar requirements of the case a fleet of four steamers was designed by Dr. J. F. Turner, Secretary of the National Board of Health, and these have just been completed at Pittsburg, leaving that city June 23 for duty on the Mississippi between Cairo and New Orleans. The crews will be selected from men who have run the dread gauntlet of 'yellow jack.' The quarantine stations already provided for are located at Cairo, Vicksburg, Memphis and New Orleans and also at the mouth of the Red River."



The man, the wi

ng and the Apple.

If you could talk to Orville Wright, he'd tell you the problems he faced as a turn-of-the-century engineer. You could tell him all about the technological solutions available to today's engineer and scientist... particularly a 20th century phenomenon that tests assumptions and defines models before a project gets off the ground. The Apple personal computer.

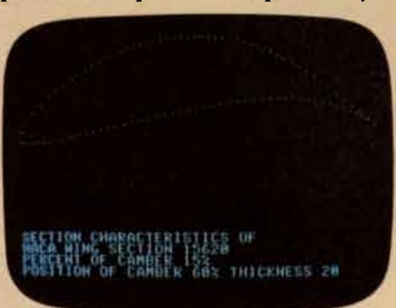
Computation, calculation, analysis...the power to pilot your projects.

With a highly-integrated system from the extensive Apple personal computer family, Orville and brother Wilbur would have increased their productivity. Perhaps even launched the Kitty Hawk Flyer well before 1903.

An Apple in their hangar would have freed them from the time and tedium of crunching numbers by hand.

An Apple in your lab or office will give you the problem-solving capabilities you demand from a big computer... without the time-consuming problems typical of remote processing.

But the Apple system solution doesn't stop there. It keeps on soaring with proven performance, power and expandability

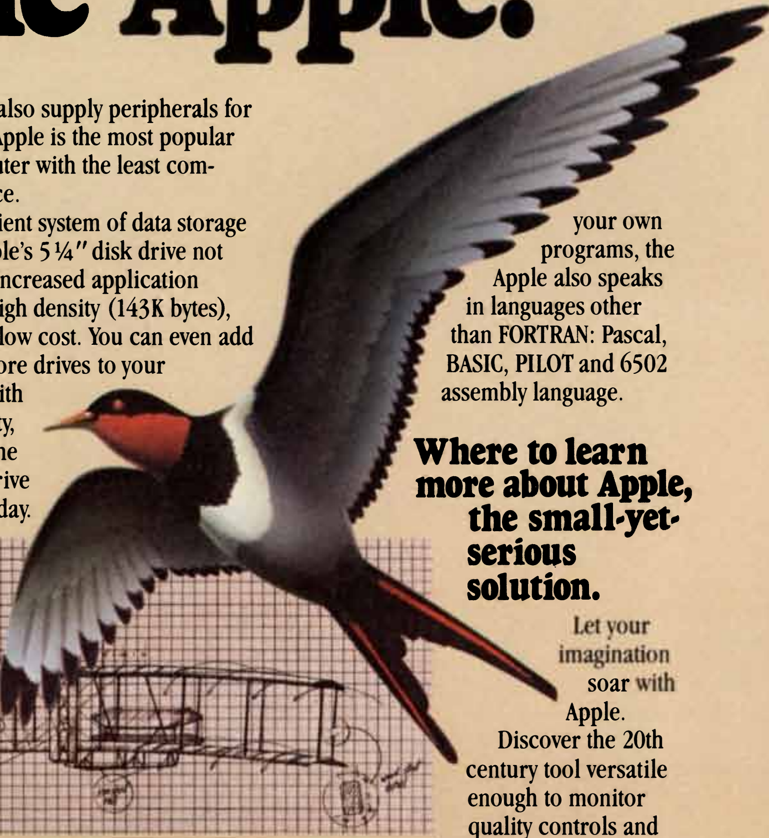


Apple's existing software library includes a program that plots the shape of an airfoil, given its parameters.

that's unparalleled for analyzing alternative paths of design and modeling a wide variety of physical processes.

Want more memory? Depending on your choice of system, Apple has memory expandable to 64K bytes or 128K bytes. Prefer wide displays? Choose 40 or 80 characters. Need to control instruments in the lab? Get on the IEEE 488 bus. Over 100 companies also supply peripherals for Apple because Apple is the most popular personal computer with the least complicated interface.

Want an efficient system of data storage and access? Apple's 5¼" disk drive not only offers you increased application versatility, but high density (143K bytes), high speed and low cost. You can even add up to four or more drives to your Apple system. With proven reliability, no wonder it's the most popular drive on the market today.



Wilbur determined that birds didn't have to constantly flap their wings to fly. With an Apple, he could've determined the fixed-wing design of the Kitty Hawk Flyer much faster.

FORTRAN that helped to design a 20th century flying machine.

Fluent in the same language that helped to design the 747, Apple FORTRAN lets you tackle differential equations at the touch of a key. And since more the 170 companies also offer software for the Apple family, you can have one of the most impressive program libraries ever... including vast subroutine libraries for math, science, engineering and statistics. When you write your own programs, the Apple also speaks in languages other than FORTRAN: Pascal, BASIC, PILOT and 6502 assembly language.

Where to learn more about Apple, the small-yet-serious solution.

Let your imagination soar with Apple. Discover the 20th century tool versatile enough to monitor quality controls and manufacturing schedules, orchestrate tolerance tests and determine alternative parts selection. Learn why Apple emerges as the technological leader of reliable personal computer products that increase your productivity.

Let the Apple dealer show you how, by putting the system of your choice through its paces. He'll tell you about our extended warranty, support and service. And he'll prove that a personal computer is not just a flight of fancy but a serious solution. Don't let history pass you by. Visit your nearest Apple dealer, or call 800-538-9696. In California, 800-662-9238.

apple computer

QUARTZ

Today begins a revolution.

The Contax 139 Quartz: At last, 100% precision.

World's first Quartz controlled camera. The long awaited compact Contax 139 Quartz is here. A unique 35mm SLR, employing a consistently accurate quartz oscillator for unprecedented precision. A quartz accuracy working within the camera extends to all of its electronic accessories. Never before has a camera enjoyed this extreme level of exactness.

Two independent SPD metering systems in one. The metering systems use an ultra-sensitive prism sensor for normal exposures and an instantaneous response film plane sensor for flash photography.

Exposure memory for critical metering. Conveniently located Auto Exposure (AE) control locks exposure settings to get the best possible exposure in unusual lighting situations.

Speed-of-light "pulse" viewfinder display. Quartz accuracy readout uses 16 bright LED's for full-information on all modes of exposure including flash, the AE control and battery status. And the exposure preview automatically turns off after 10 seconds.

Exclusive Auto Winder for easy control. The functional design of this 2 fps winder allows for stable and comfortable handling. It also features a convenient vertical shutter release.

Perfect flash exposures with the TLA-20 Auto Flash. A through-the-lens Direct Flash Control system provides exact exposures at any aperture and total accuracy with any lens accessory, including bellows.

A dramatically innovative Carl Zeiss lens. A 7-element Planar T* 50mm f/1.7 lens is standard. And Zeiss, renowned for optical excellence, offers a full line of lenses including: the Vario-Sonnar T* f/3.5 40 to 80mm zoom; the S-Planar T* f/2.8 60mm macro; the "Olympia" Sonnar T* f/2.8 180mm telephoto, and more.

And the technological revolution continues. An LED flashing Quartz self-timer with reset capability; feather-touch electro/magnetic shutter release; infrared and radio remote control units as well as 200 other highly sophisticated accessories.

See this exciting revolution at your local Yashica/Contax dealer.

Yashica, Inc., 411 Sette Drive, Paramus, New Jersey 07652/In Canada: Yashica Canada, Inc., Ontario.

CONTAX 139 QUARTZ

THE AUTHORS

WILLIAM EPSTEIN ("A Ban on the Production of Fissionable Material for Weapons") is a Special Fellow of the United Nations Institute for Training and Research (UNITAR). He also serves as a consultant on arms control and disarmament to both the UN Secretary General and the Canadian government. For many years Epstein was the senior Canadian national on the permanent staff of the UN Secretariat in New York, working until his retirement in 1973 as director of the Disarmament Division. A graduate of the University of Alberta, where he received his B.A. and LL.B. degrees, he did postgraduate work in international affairs at the London School of Economics. After serving in the Canadian army in World War II he joined the staff of the UN Preparatory Commission Secretariat in London in 1945. He has been associated with the UN ever since. Recently he served as a member of the Group of Consultant Experts appointed by the Secretary General that prepared a report, requested by the General Assembly, on a comprehensive nuclear-test ban. In addition to his work as an international civil servant Epstein has been a visiting professor at several universities.

DAVID KOFFLER ("Systemic Lupus Erythematosus") is professor of pathology and director of laboratory medicine at the Hahnemann Medical College and Hospital of Philadelphia. He is also adjunct professor for immunology at Rockefeller University. As an undergraduate at New York University, Koffler was a prelaw student, but he later switched to medical studies, obtaining his M.D. at the State University of New York College of Medicine in 1958. The profoundest influence on his research career, he says, has been his association with the immunologist Henry G. Kunkel, in whose laboratory at Rockefeller University he got his research training and with whom he has continued to collaborate over the past 10 years. Before taking up his present posts Koffler was professor of pathology at the Mount Sinai School of Medicine and director of pathology at Mount Sinai Hospital.

MARVIN LEVENTHAL and CRAWFORD J. MACCALLUM ("Gamma-Ray-Line Astronomy") are physicists with a common interest in the spectroscopic study of celestial processes that emit gamma radiation at discrete wavelengths. Leventhal is a member of the technical staff at Bell Laboratories. He studied physics at the City College of New York and Brown University, obtaining his Ph.D. from the latter in 1964. He was on the faculty of Yale University until 1968, when he joined Bell Laboratories. Leventhal's main professional fields of interest in addition to gamma-ray astrophysics include nuclear-magnetic-resonance investigations of amorphous solids, spectroscopic measurements of hydrogen ions and the physics of positronium. MacCallum is at Sandia Laboratories. His bachelor's degree is from Princeton University and his doctorate is from the University of New Mexico.

GLYN DANIEL ("Megalithic Monuments") is Disney Professor of Archaeology and head of the department of archaeology at the University of Cambridge. A native of South Wales, he went to Cambridge as an undergraduate in 1932 and has been there ever since except for five years in the Royal Air Force photointelligence service during World War II. His main interests are megalithic monuments and the history of archaeology, and he has written a number of books on these subjects. With his wife Ruth, "a geographer turned archaeologist," he also edits the journal *Antiquity*. In addition to his scholarly work Daniel has written a travel book called *The Hungry Archaeologist in France* and two detective stories, *The Cambridge Murders* and *Welcome Death*.

MANUEL G. VELARDE and CHRISTIANE NORMAND ("Convection") have collaborated at long range on the subject of their article since 1974, when both were working in the theoretical-physics division of the Saclay Nuclear Research Center in France. Velarde is currently professor of statistical mechanics and head of the department of fluid physics at the Autonomous University of Madrid; he is also on the faculty of the Open University in Madrid. The holder of two doctoral degrees, one from the Complutensian University of Madrid (1968) and the other from the University of Brussels (1970), he has worked for extended periods at research laboratories in the U.S., France, Norway, Belgium and Britain. Normand continues her work at the French Atomic Energy Commission's facility at Saclay, where she first went on a fellowship in 1973, shortly after being graduated with a degree in plasma physics from the University of Paris-South. She became interested in the analysis of convective motion soon after her arrival at Saclay, where, she writes, "I met Pierre Bergé and Monique Dubois and saw for the first time their beautiful experiments on Rayleigh-Bénard convection." Since 1976 Normand has been a research associate of the French National Center for Scientific Research (C.N.R.S.).

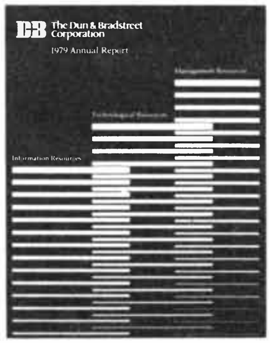
ROBERT DEGABRIELE ("The Physiology of the Koala") teaches biology at Riverina College of Advanced Education, which is at the inland city of Wagga Wagga in New South Wales in Australia. He was educated at the University of New South Wales, receiving his M.Sc. in biology in 1977. His interest in marsupials and their evolution, he writes, "originated as a result of my being a student of T. J. Dawson, the author of the article titled 'Kangaroos' in the August 1977 issue of *Scientific American*."

WILLIAM R. BAUER, F. H. C. CRICK and JAMES H. WHITE ("Supercoiled DNA") approach the supercoiling problem from three quite different backgrounds. Bauer is professor of microbiology at the State University of New York at Stony Brook. He has two bachelor's degrees: a B.S. in chemistry from the California Institute of Technology and a B.A. in animal physiology and biochemistry from the University of Oxford. His Ph.D., in chemistry, is from Cal Tech. Before he moved to Stony Brook in 1973 Bauer taught chemistry for several years at the University of Colorado at Boulder. Crick, the co-winner (with James D. Watson and M. H. F. Wilkins) of the 1962 Nobel Prize in physiology or medicine for his part in the discovery of the double-helix structure of the DNA molecule, is currently Kieckhefer Distinguished Research Professor at the Salk Institute for Biological Studies. He did his undergraduate work in physics at University College London and his graduate work in biology at the University of Cambridge, obtaining his Ph.D. from the latter in 1954. Crick has written a number of articles for SCIENTIFIC AMERICAN; the most recent, "Thinking about the Brain," appeared in last September's single-topic issue on the brain. White is professor of mathematics at the University of California at Los Angeles. A graduate of the University of California at Berkeley, he went on to get his Ph.D. in mathematics from the University of Minnesota in 1968. His specialties include the topology of curves and surfaces in three-dimensional space.

E. EUGENE LARRABEE ("The Screw Propeller") is associate professor of aeronautics and astronautics at the Massachusetts Institute of Technology. He joined the faculty of M.I.T. as a teaching assistant in 1946, he reports, "after wartime work as an aerodynamicist with the Curtiss-Wright Corporation. I have always been interested in airplanes, having been imprinted as a child with the example of Charles Lindbergh. My professional interests are centered on the automobile as well as the airplane, and I have worked on topics that range from driver-steering strategy to the lateral control of sailplanes."

13 Great Books

Here's your opportunity to receive a choice collection of the latest annual reports from leading corporations. Use the handy coupon. There's no obligation.

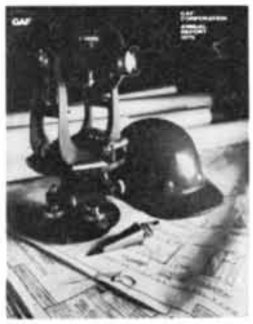


THE DUN & BRADSTREET CORPORATION

The Dun & Bradstreet Corporation provides products and services in four groups: business information, publishing, marketing services and broadcasting. The company is the largest producer and disseminator of business information in the private sector. Earnings per share in 1979 were a record $3.19, up 26.6%, and more than double those of four years ago. Dividends were increased for the 28th consecutive year.

ÆTNA LIFE & CASUALTY

Ætna Life & Casualty—largest investor-owned insurance organization in U.S., also in business financing, real estate development, technology enterprises. 1979 earnings reached new high of $560 million or $6.93 per common share, 21% return on shareholders' equity. Revenues rose 21% to $11.4 billion. Assets, shareholders' equity grew to $30.2 billion and $35.56 per common share. Annual dividend per common share increased 18% to $2.12 with May 15 payment. Annual dividend 194% greater than five years ago.

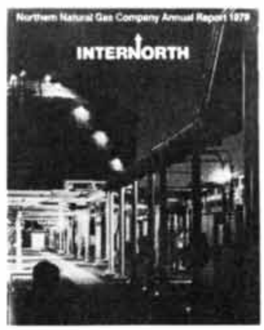


GAF CORPORATION

GAF is expanding advanced technology in its chemical, building materials and reprographic businesses. Its specialty chemicals are gaining new applications in products as diverse as shampoos and auto parts, wines and adhesives. Its roofing, siding, insulation and vinyl floorings are in demand for remodeling and repair work as well as in new construction. GAF's diazo reprographic systems are moving into fields long monopolized by silver-based films and papers. 1979 sales were $1.2 billion.

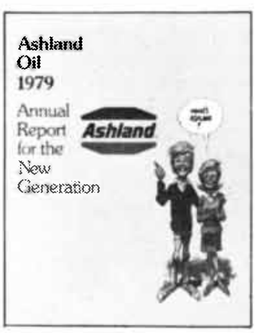
ARMCO INC.

We turned in record sales and earnings in 1979. Our annual report tells how we did it, reports results, both current and historical and describes our strategic plans for the future. We're a diversified five billion dollar corporation, organized with seven lines of business: Oilfield Equipment and Production, Fabricated Metal Products, Industrial Products and Services, Financial Services, Carbon Steel Products, Specialty Steel Products, Mining and Raw Materials.



INTERNORTH (Northern Natural Gas)

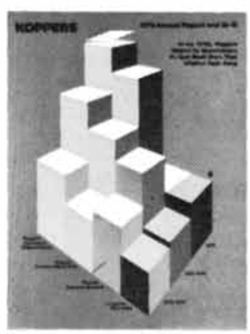
Record earnings and revenues highlighted 1979, Northern's fiftieth year. Net income was a record $185.5 million. Operating revenues totaled a record $2.5 billion. Property, plant and equipment, at cost, now exceeds $3 billion. Earnings per share of $8.17 were our highest ever. Northern is an energy-based company involved in natural gas, liquefied fuels, petrochemicals, oil and gas exploration and coal operations.



ASHLAND OIL

This year Ashland, the nation's largest independent refiner and marketer, published two annual reports. One for today's investors. One for the new generation of investors. This colorful report discusses facts from the 1979 Annual Report—explaining how private investors make our progress possible.

Like these record fiscal 1979 results: Revenues—$6.7 million; net income $526 million. Dividends increased from annual rate of $1.33 in 1978 to a current annual rate of $2.20. Both annual reports will be sent.



KOPPERS COMPANY, INC.

Investment in Koppers over the past five years produced an average total return of 29% per year to shareholders. The highly diversified industrial operations of the company are strategically situated so that 75% of sales are in the construction and capital spending segments of the economy and 25% in intermediate industrial products. This product mix gives Koppers a decided edge in providing future growth and continued high returns to investors.

of the Year. Free.

THE MEAD CORPORATION

1979 another record year: 17% increase in net earnings to $141 million resulted in•record 18.1% return on shareholders' equity. Fully diluted earnings per share reached $5.19 from $2.72 in '74—a compound growth rate of 14%—one of the most impressive in the forest products industry. Dividends have increased for seven consecutive years to annual rate of $1.80 per share.

W.R. GRACE & CO.

Last year our Annual Report was judged best among 2,000 other companies. In 1979 it's even better with net income up 30% to $223 million, record sales of $5.3 billion up 18% and a market price ahead 57%. But we're no one-time champion. Over the decade of the seventies, our earnings have risen at a compound rate of more than 31% per year. Shouldn't you look into GRACE? Our 75,000 shareholders are glad they did.

COMBUSTION ENGINEERING

The Energy Systems Company... Combustion Engineering is a diversified corporation which provides equipment, products and services for oil and gas exploration and production, petroleum and petrochemical processing, environmental control, electric power generation, synthetic fuels, and other energy-related industrial applications. C-E has annual sales approaching $3 billion, and a current annual dividend rate of $2.60, which has been increased eight times in the last five years.

HOW TO READ A FINANCIAL REPORT

Plain talk in this 32 page booklet cuts through the financial mumbo jumbo. Explains what terms like "net fixed assets" and "total shareholders equity" mean. Shows how to use an annual report as a guide for investment decision making.

MERRILL, LYNCH, PIERCE, FENNER & SMITH, INC.

GREAT NORTHERN NEKOOSA CORP.

Great Northern Nekoosa, a major pulp and paper company and the world's largest producer of copier paper, has large, efficient mills which produce low cost business papers, containerboard, newsprint and groundwood printing paper. GNN's assets include nearly 7,000 tons a day of pulp and paper capacity, 2,800,000 acres of timberland, millions of tons of coal reserves, and an extensive hydroelectric system. At just over 70%, GNN's energy self-sufficiency is among the highest in the paper industry.

DOMINION BRIDGE COMPANY, LIMITED

During the Seventies this diversified international company achieved a 28% compounded annual growth rate of net income, while multiplying worldwide sales fivefold—to nearly $1 billion in 1979. During the decade its net income rose 970% and its growth rate exceeded 490 of the *Fortune* 500. Engaged principally in manufacturing steel-based products, engineering and construction activities, Dominion Bridge is traded on the Toronto and Montreal Stock Exchanges.

Annual Reports, P.O. Box 2503, Clinton, Iowa 52735
Or CALL FREE 800-431-1866 Ext. 30
(In N.Y. State 800-942-1917)

I've checked below which of the following companies I want to send me their most recent annual reports. I understand there's no cost or obligation.

- ☐ 1 Aetna Life
- ☐ 2 Armco Inc.
- ☐ 3 Ashland
- ☐ 4 The Dun & Bradstreet Corporation
- ☐ 5 GAF Corporation
- ☐ 6 Internorth
- ☐ 7 Koppers Company, Inc.
- ☐ 8 Mead
- ☐ 9 Combustion Engineering
- ☐ 10 Great Northern Nekoosa
- ☐ 11 Dominion Bridge Company
- ☐ 12 W.R. Grace & Co.
- ☐ 13 Merrill Lynch (How to Read a Financial Report)

Name____________________________

Address____________________________

City______________State________Zip________

This offer expires November 1, 1980.

SA780

MATHEMATICAL GAMES

The pleasures of doing science and technology in the planiverse

by Martin Gardner

"Planiversal scientists are not a very common breed."

—Alexander Keewatin Dewdney

As far as anyone knows the only existing universe is the one we live in, with its three dimensions of space and one of time. It is not hard to imagine, as many science-fiction writers have, that intelligent organisms could live in a four-dimensional space, but two dimensions offer such limited degrees of freedom that it has long been assumed intelligent two-space life forms could not exist. Two notable attempts have nonetheless been made to describe such organisms.

In 1884 Edwin Abbott Abbott, a London clergyman, published his satirical novel *Flatland.* Unfortunately the book leaves the reader almost entirely in the dark about Flatland's physical laws and the technology developed by its inhabitants, but the situation was greatly improved in 1907 when Charles Howard Hinton published *An Episode of Flatland.* Although the style and the characters were flat even for a three-dimensional world, Hinton's story provided the first glimpses of the possible science and technology of the two-dimensional one. His eccentric book is, alas, long out of print, but you can read about it in the chapter "Flatlands" in my book *The Unexpected Hanging and Other Mathematical Diversions* (Simon & Schuster, 1969).

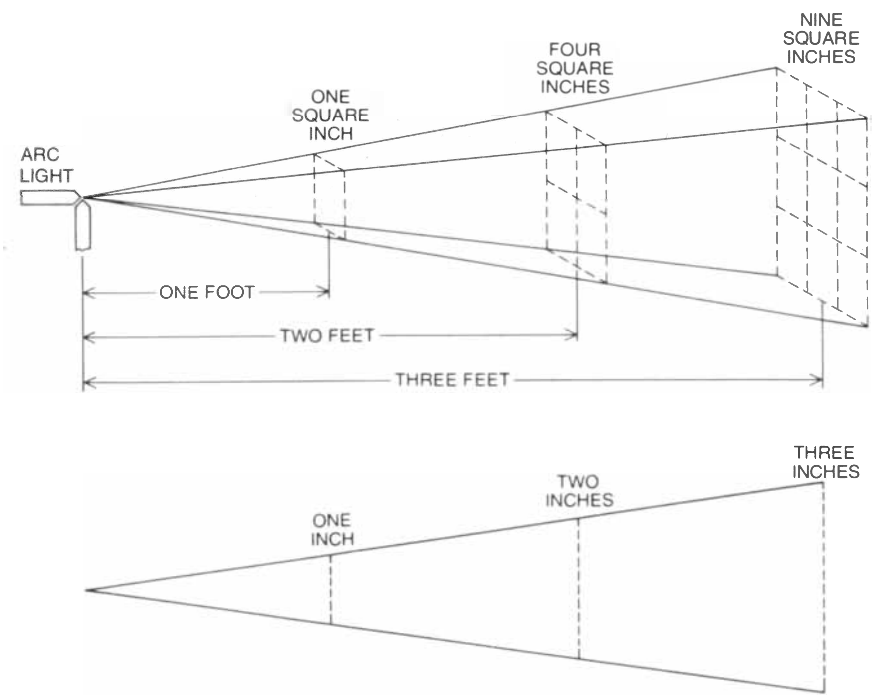
In "Flatlands" I wrote: "It is amusing to speculate on two-dimensional physics and the kinds of simple mechanical devices that would be feasible in a flat world." This remark caught the attention of Alexander Keewatin Dewdney, a computer scientist at the University of Western Ontario. Some of his early speculations on the subject were set down in 1978 in a university report and in 1979 in "Exploring the Planiverse," an article in *Journal of Recreational Mathematics* (Vol. 12, No. 1, pages 16–20; September). Later in 1979 Dewdney also privately published "Two-dimensional Science and Technology," a 97-page tour de force. It is hard to believe, but Dewdney actually lays the groundwork for what he calls a planiverse: a possible two-dimensional world. Complete with its own laws of chemistry, physics, astronomy and biology, the planiverse is closely analogous to our own universe (which he calls the steriverse) and is apparently free of contradictions. I should add that this remarkable achievement is an amusing hobby for a mathematician whose serious contributions have appeared in some 30 papers in technical journals.

Dewdney's planiverse resembles Hinton's in having an earth that he calls (as Hinton did) Astria. Astria is a disklike planet that rotates in planar space. The Astrians, walking upright on the rim of the planet, can distinguish east and west and up and down. Naturally there is no north or south. The "axis" of Astria is a point at the center of the circular planet. You can think of such a flat planet as being truly two-dimensional or you can give it a very slight thickness and imagine it as moving between two frictionless planes.

As in our world, gravity in a planiverse is a force between objects that varies directly with the product of their masses, but it varies inversely with the linear distance between them, not with the square of that distance. On the assumption that forces such as light and gravity in a planiverse move in straight lines, it is easy to see that the intensity of such forces must vary inversely with linear distance. The familiar textbook figure demonstrating that in our world the intensity of light varies inversely with the square of distance is shown at the top of the illustration at the left. The obvious planar analogue is shown at the bottom of the illustration.

To keep his whimsical project from "degenerating into idle speculation" Dewdney adopts two basic principles. The "principle of similarity" states that the planiverse must be as much like the steriverse as possible: a motion not influenced by outside forces follows a straight line, the flat analogue of a sphere is a circle and so on. The "principle of modification" states that in those cases where one is forced to choose between conflicting hypotheses, each one equally similar to a steriversal theory, the more fundamental one must be chosen and the other modified. To determine which hypothesis is more fundamental Dewdney relies on the hierarchy in which physics is more fundamental than chemistry, chemistry more fundamental than biology and so on.

To illustrate the interplay between levels of theory Dewdney considers the evolution of the planiversal hoist in the top illustration on page 21. The engi-



An inverse-square law for our world **(top)** *and its linear analogue for the planar world* **(bottom)**

WHAT'S NEWS IN PATENTS?

A continuing series on progress by GE inventors.

Last year, General Electric inventors were awarded over 600 patents, adding to GE's unsurpassed total of over fifty thousand U.S. patents, past and present. Here, GE reports on some recent patents—and on the inventors who won them.

Faust and Wagenknecht: High and fast vs. low and slow.

There's no such thing as an all-purpose aircraft engine.

Take the turbojet. It performs magnificently at high speeds and altitudes. But for flying lower and slower, the turbofan is a better choice.

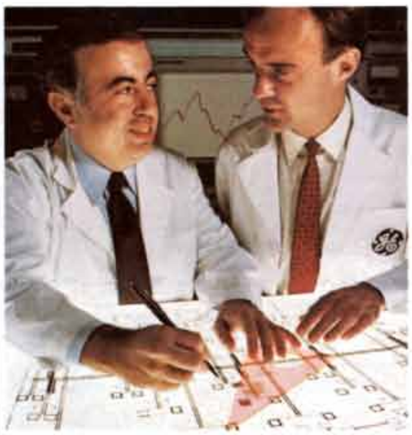
Could an engine be designed to operate in both modes? Many have worked on answering that question.

One recent answer comes from GE engineers Kemp Faust and Conrad Wagenknecht.

Their brainchild is U.S. Patent Number 4,175,384—"Individual Bypass Injector Valves for a Double Bypass Variable Cycle Engine."

Their system splits an engine's internal airflow into three streams that exit from a single, variable exhaust nozzle. The airflow is divided and selectively modulated by a variable area bypass injector. The system helps provide excellent efficiency at both cruising and high Mach speeds; low and high altitudes.

In the search for more versatile engines, Faust and Wagenknecht are part of a GE tradition of progress.



Cline and Anthony: Migratory metals.

"Impurities" in silicon wafers are what give the ubiquitous semiconductor its unique qualities.

It was two GE physicists, Harvey Cline and Tom Anthony, who devised one of the fastest and most precise techniques of introducing such impurities by thermomigration.

In thermomigration, one side of a silicon wafer is cooled, the other is heated. Metal-bearing liquid droplets migrate through the silicon from cold side to hot, leaving the metal dispersed throughout the silicon.

Cline and Anthony's patent, Number 4,184,897—"Droplet Migration Doping Using Carrier Droplets," covers a process for depositing metals like antimony, arsenic and phosphorus. As a result, thermomigration can now be used to fabricate the N-type (electron-rich) portions of semiconductors.

Cline and Anthony are no strangers to progress. Between them, they hold many patents—more than fifty on thermomigration alone.

Germer and Palmer: Leveling the peaks.

Many electric meters can't tell time. They register nothing more than total power consumption.

Too bad. Time-telling meters make it possible for utilities to charge one rate for peak use; a lower one for off-peak use. That tends to flatten the peaks of daily demand.

Enter two General Electric engineers, Warren Germer and Ansell Palmer. They've devised a programmable meter that can tell both time of day and day of the week. (Weekend peaks differ from weekday ones.) Their patent, Number 4,199,717—"Time-of-Day Demand Metering System," is quite a breakthrough.

Heart of Germer and Palmer's meter is a unique large-scale integrated circuit designed for this job alone. Incorporating a digital clock, the circuit can discriminate between peak and off-peak consumption and record both. It even stays on schedule during a power interruption.

Thousands of the new meters are now part of power company test programs, giving many customers a real incentive to level out power use and reduce their electric bills.

To learn more about GE research, write General Electric, Section 11, Fairfield, CT 06431.

We've got a patent on progress.

GENERAL ELECTRIC

AFTER YOU COMPARE IT FOR MILES PER GALLON, COMPARE IT FOR COMFORT PER MILE.

Load up Renault Le Car, and take advantage of more luggage space than you'll find in either a Ford Fiesta or Honda Civic.†

Settle into the anatomically-contoured seats, and luxuriate in more passenger room than either a Civic or a Datsun 210.†

Pick out a really rough road, and watch Le Car's four-wheel independent suspension and standard steel-belted Michelin radials soak up jolts and bumps so well, they give you a smooth, level ride *Motor Trend* says "would do credit to far larger, more expensive cars."

Just visit one of the hundreds of Renault dealers from coast to coast, and take Le Car for a test-drive.

When you read the mileage estimates, you'll know how much a car costs to drive. But only after you test it on the road will you know whether or not it's worth driving.

*Remember: Compare these 1980 EPA estimates to estimated mpg for other cars. Your mileage may vary due to speed, trip length or weather. Your highway mileage will probably be lower. California excluded. †Based on 1980 EPA data.

RENAULT LE CAR

WE BUILD MORE INTO ECONOMY CARS THAN JUST ECONOMY.

Imported by American Motors

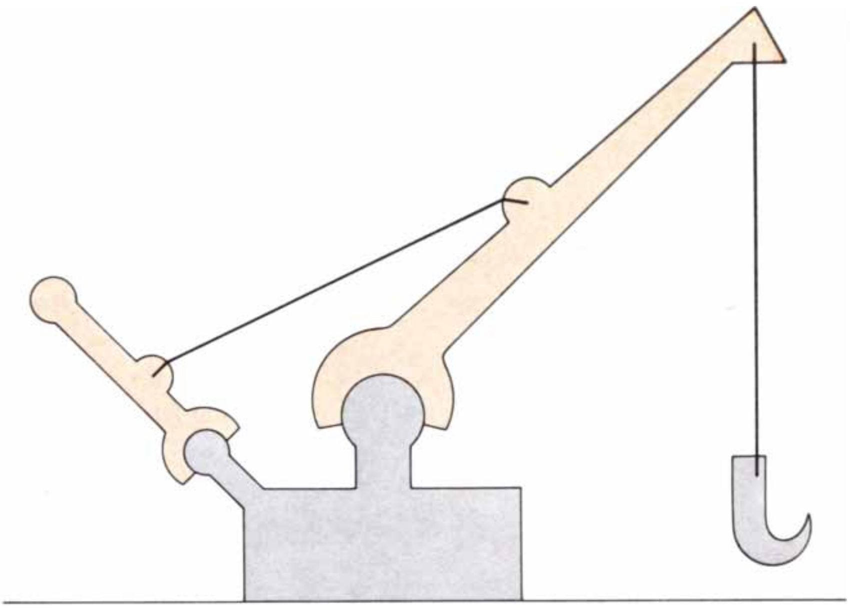
neer who designed it first gave it arms thinner than those in the illustration, but when a metallurgist pointed out that planar materials fracture more easily than their three-space counterparts, the engineer made the arms thicker. Later a theoretical chemist, invoking the principles of similarity and modification at a deeper level, calculated that the planiversal molecular forces are much stronger than had been suspected, and so the engineer went back to thinner arms.

The principle of similarity leads Dewdney to posit that the planiverse is a three-dimensional continuum of space-time containing matter composed of molecules, atoms and fundamental particles. Energy is propagated by waves, and it is quantized. Light exists in all its wavelengths and is refracted by planar lenses, making possible planiversal eyes, planiversal telescopes and planiversal microscopes. The planiverse shares with the steriverse such basic precepts as causality, the first and second laws of thermodynamics and laws concerning inertia, work, friction, magnetism and elasticity.

Dewdney assumes that his planiverse began with a big bang and is currently expanding. An elementary calculation based on the inverse-linear gravity law shows that regardless of the amount of mass in the planiverse the expansion must eventually halt, so that a contracting phase will begin. The Astrian night sky will of course be a semicircle along which are scattered twinkling points of light. If the stars have proper motions, they will continually be occulting one another. If Astria has a sister planet, it will over a period of time occult every star in the sky.

We can assume that Astria revolves around a sun and rotates to create day and night. In a planiverse, Dewdney discovered, the only stable orbit that continually retraces the same path is a perfect circle. Other stable orbits roughly elliptical in shape are possible, but the axis of the ellipse rotates in such a way that the orbit never exactly closes. Whether planiversal gravity would allow a moon to have a stable orbit around Astria remains to be determined. The difficulty is due to the sun's gravity, and resolving the question calls for work on the planar analogue of what our astronomers know as the three-body problem.

Dewdney analyzes in detail the nature of Astrian weather, with its analogues of our seasons, winds, clouds and rain. An Astrian river would be indistinguishable from a lake except that it might have faster currents. One peculiar feature of Astrian geology is that water cannot flow around a rock as it does on the earth. As a result rainwater steadily accumulates behind any rock on a slope, tending to push the rock downhill: the gentler the slope is, the more water accumulates and the stronger the push is. Dewdney concludes that given periodic rainfall the Astrian surface would be unusually flat and uniform. Another consequence of the inability of water to move sideways on Astria is that it would become trapped in pockets within the soil, tending to create large areas of treacherous quicksand in the hollows of the planet. One hopes, Dewdney writes, that rainfall is infrequent on Astria. Wind too would have much severer effects on Astria than on the earth because like rain it cannot "go around" objects.



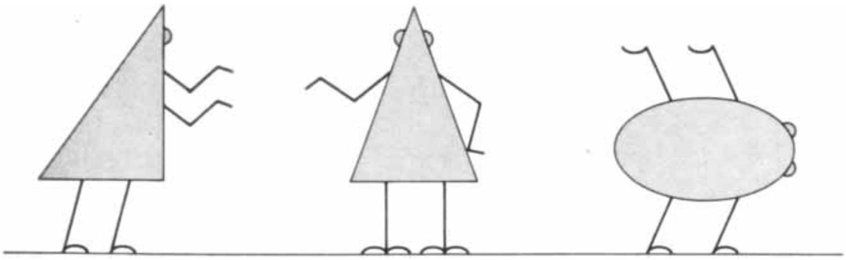
A planiversal hoist

Dewdney devotes many pages to constructing a plausible chemistry for his planiverse, modeling it as much as possible on three-dimensional matter and the laws of quantum mechanics. The illustration below shows Dewdney's periodic table for the first 16 planiversal elements. Because the first two are so much like their counterparts in our world, they are called hydrogen and helium. The next 10 have composite names to suggest the steriversal elements they most resemble; for example, lithrogen combines the properties of lithium and nitrogen. The next four are named after Hinton, Abbott and the young lovers in Hinton's novel, Harold Wall and Laura Cartwright.

In the flat world atoms combine naturally to form molecules, but of course only bonding that can be diagrammed by a planar graph is allowed. (This result follows by analogy from the fact that intersecting bonds do not exist in steriversal chemistry.) As in our world, two asymmetric molecules can be mirror

ATOMIC NUMBER	NAME	SYMBOL	SHELL STRUCTURE 1s	2s	2p	3s	3p	3d	4s	4p . . .	VALENCE
1	HYDROGEN	H	1								1
2	HELIUM	He	2								2
3	LITROGEN	Lt	2	1							1
4	BEROXYGEN	Bx	2	2							2
5	FLUORON	Fl	2	2	1						3
6	NEOCARBON	Nc	2	2	2						4
7	SODALINUM	Sa	2	2	2	1					1
8	MAGNILICON	Mc	2	2	2	2					2
9	ALUPHORUS	Ap	2	2	2	2	1				3
10	SULFICON	Sp	2	2	2	2	2				4
11	CHLOPHORUS	Cp	2	2	2	2	2	1			5
12	ARGOFUR	Af	2	2	2	2	2	2			6
13	HINTONIUM	Hn	2	2	2	2	2	2	1		1
14	ABBOGEN	Ab	2	2	2	2	2	2	2		2
15	HAROLDIUM	Wa	2	2	2	2	2	2	2	1	3
16	LAURANIUM	La	2	2	2	2	2	2	2	2	4

Periodic table for the first 16 planiversal elements



Hinton's Astrian (left), *Dewdney's Astrian* (center) *and an Astrian bug-eyed monster* (right)

images of each other, so that neither one can be "turned over" to become identical with the other. There are striking parallels between planiversal chemistry and the behavior of steriversal monolayers on crystal surfaces [see "Two-dimensional Matter," by J. G. Dash; SCIENTIFIC AMERICAN, May, 1973]. In our world molecules can form 230 distinct crystallographic groups, but in the planiverse they can form only 17. I am obliged to pass over Dewdney's speculations about the diffusion of molecules, electrical and magnetic laws, analogues of Maxwell's equations and other subjects too technical to summarize here.

Dewdney assumes that animals on Astria are composed of cells that cluster to form bones, muscles and connective tissues similar to those found in steriversal biology. He has little difficulty showing how these bones and muscles can be structured to move appendages in such a way that the animals can crawl, walk, fly and swim. Indeed, some of these movements are easier in a planiverse than in our world. For example, a steriversal animal with two legs has considerable difficulty balancing while walking, whereas in the planiverse if an animal has both legs on the ground, there is no way it can fall over. Moreover, a flying planiversal animal cannot have wings and does not need them to fly; if the body of the animal is aerodynamically shaped, it can act as a wing (since air can go around it only in the plane). The flying animal could be propelled by a flapping tail.

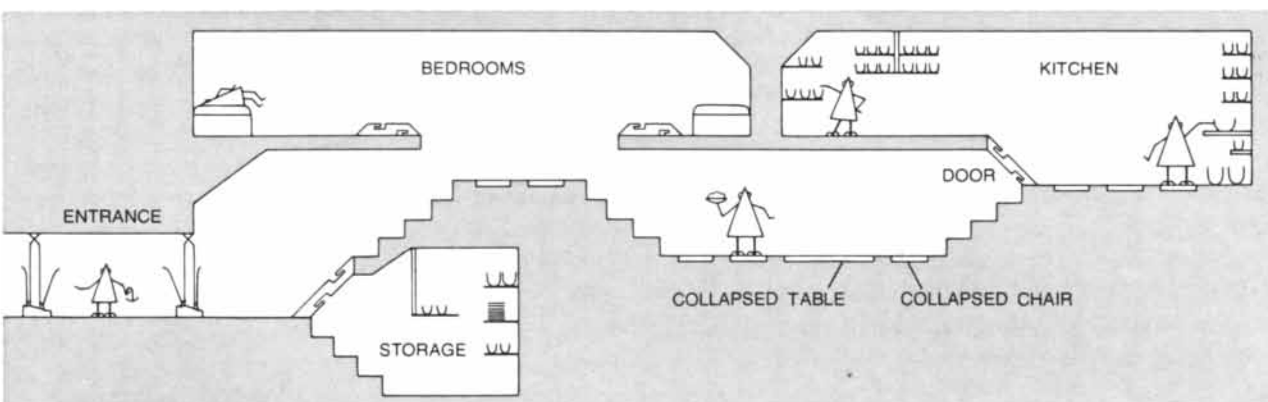
Calculations also show that Astrian animals probably have much lower metabolic rates than terrestrial animals because relatively little heat is lost through the perimeter of their body. Furthermore, animal bones can be thinner on Astria than they are on the earth, because they have less weight to support. Of course, no Astrian animal can have an open tube extending from its mouth to its anus, because if it did, it would be cut in two.

In the appendix to his book *The Structure and Evolution of the Universe* (Harper, 1959) G. J. Whitrow argues that intelligence could not evolve in two-space because of the severe restrictions two dimensions impose on nerve connections. "In three or more dimensions," he writes, "any number of [nerve] cells can be connected with [one another] in pairs without intersection of the joins, but in two dimensions the maximum number of cells for which this is possible is only four." Dewdney easily demolishes this argument, pointing out that if nerve cells are allowed to fire nerve impulses through "crossover points," they can form flat networks as complex as any in the steriverse. Planiversal minds would operate more slowly than steriversal ones, however, because in the two-dimensional networks the pulses would encounter more interruptions. (There are comparable results in the theory of two-dimensional automatons.)

Dewdney sketches in detail the anatomy of an Astrian female fish with a sac of unfertilized eggs between its two tail muscles. The fish has an external skeleton, and nourishment is provided by the internal circulation of food vesicles. If a cell is isolated, food enters it through a membrane that can have only one opening at a time. If the cell is in contact with other cells, as in a tissue, it can have more than one opening at a time because the surrounding cells are able to keep it intact. We can of course see every internal organ of the fish or of any other planiversal life form, just as a four-dimensional animal could see all our internal organs.

Dewdney follows Hinton in depicting his Astrian people schematically, as triangles with two arms and two legs. Hinton's Astrians, however, always face in the same direction: males to the east and females to the west. In both sexes the arms are on the front side, and there is a single eye at the top of the triangle, as is shown at the left in the top illustration on this page. Dewdney's Astrians are bilaterally symmetrical, with an arm, a leg and an eye on each side, as is shown at the center of the illustration. Hence these Astrians, like terrestrial birds or horses, can see in opposite directions. Naturally the only way for one Astrian to pass another is to crawl or leap over him. My conception of an Astrian bug-eyed monster is shown at the right in the illustration. This creature's appendages serve as either arms or legs, depending on which way it is facing, and its two eyes provide binocular vision. With only one eye an Astrian would have a largely one-dimensional visual world, giving him a rather narrow perception of reality. On the other hand, parts of objects in the planiverse might be distinguished by their color, and an illusion of depth might be created by the focusing of the lens of the eye.

On Astria building a house or mowing a lawn requires less work than it does on the earth because the amount of material involved is considerably smaller. As Dewdney points out, however, there are still formidable problems to be dealt with in a two-dimensional world: "Assuming that the surface of the planet is



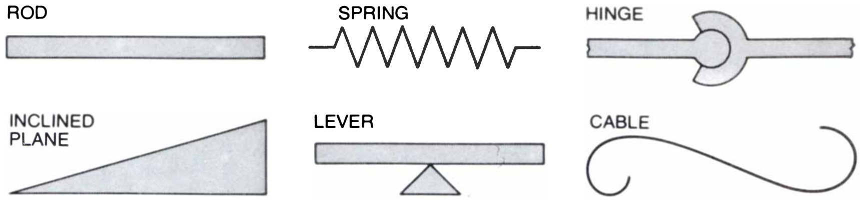
An Astrian underground house

absolutely essential to support life-giving plants and animals, it is clear that very little of the Astrian surface can be disturbed without inviting the biological destruction of the planet. For example, here on earth we may build a modest highway through the middle of several acres of rich farmland and destroy no more than a small percentage of it. A corresponding highway on Astria will destroy *all* the 'acreage' it passes over.... Similarly, extensive cities would quickly use up the Astrian countryside. It would seem that the only alternative for the Astrian technological society is to go underground." A typical subterranean house with a living room, two bedrooms and a storage room is shown in the bottom illustration on the opposite page. Collapsible chairs and tables are stored in recesses in the floors to make the rooms easier to walk through.

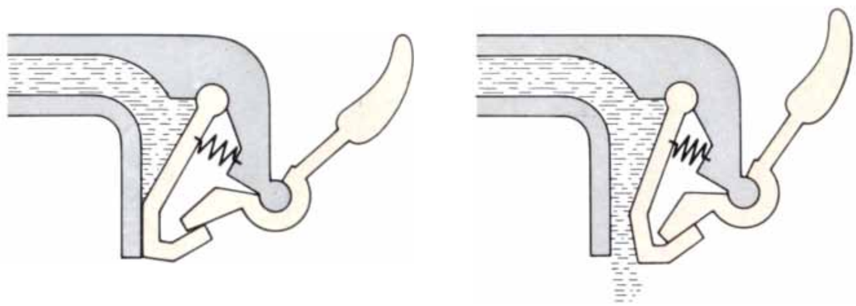
The many simple three-dimensional mechanical elements that have obvious analogues on Astria include rods, levers, inclined planes, springs, hinges, ropes and cables. Wheels can be rolled along the ground, but there is no way to turn them on a fixed axle. Screws are impossible. Ropes cannot be knotted; on the other hand, they never tangle. Tubes and pipes must have partitions, to keep their sides in place, and the partitions have to be opened (but never all of them at once) to allow anything to pass through. It is remarkable that in spite of these severe constraints many flat mechanical devices can be built that will work. A faucet designed by Dewdney is shown in the middle illustration at the right. To operate it the handle is lifted. This action pulls the valve away from the wall of the spout, allowing the water to flow out. When the handle is released, the spring pushes the valve back.

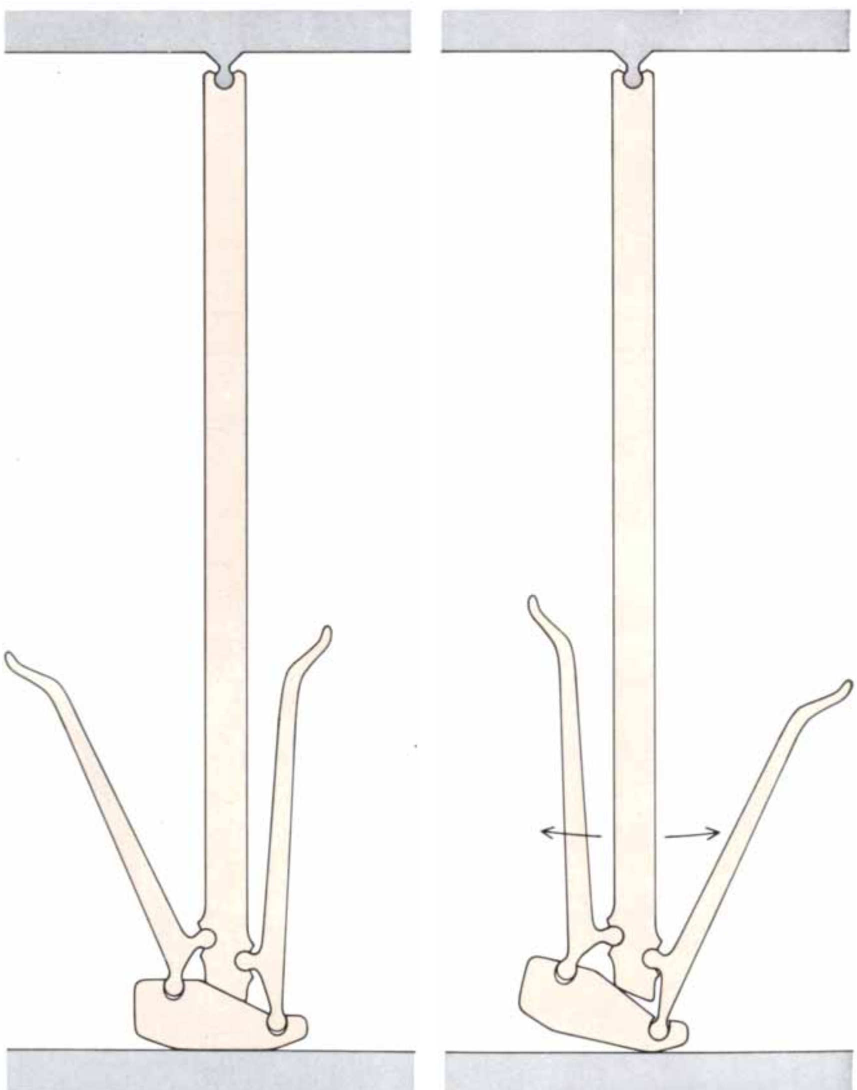
The device shown in the bottom illustration at the right serves to open and close a door (or a wall). Pulling down the lever at the right forces the wedge at the bottom to the left, thereby allowing the door to swing upward (carrying the wedge and the levers with it) on a hinge at the top. The door is opened from the left by pushing up on the other lever. The door can be lowered from either side and the wedge moved back to stabilize the wall by moving a lever in the appropriate direction. This device and the faucet are both mechanisms with permanent planiversal hinges: circular knobs that rotate inside hollows but cannot be removed from them.

The illustration on page 26 depicts a planiversal steam engine whose operation parallels that of a steriversal engine. Steam under pressure is admitted into the cylinder of the engine through a sliding valve that forms one of its walls (*top*). The steam pressure causes a piston to move to the right until steam can escape into a reservoir chamber above it. The subsequent loss of pressure allows the compound leaf spring at the

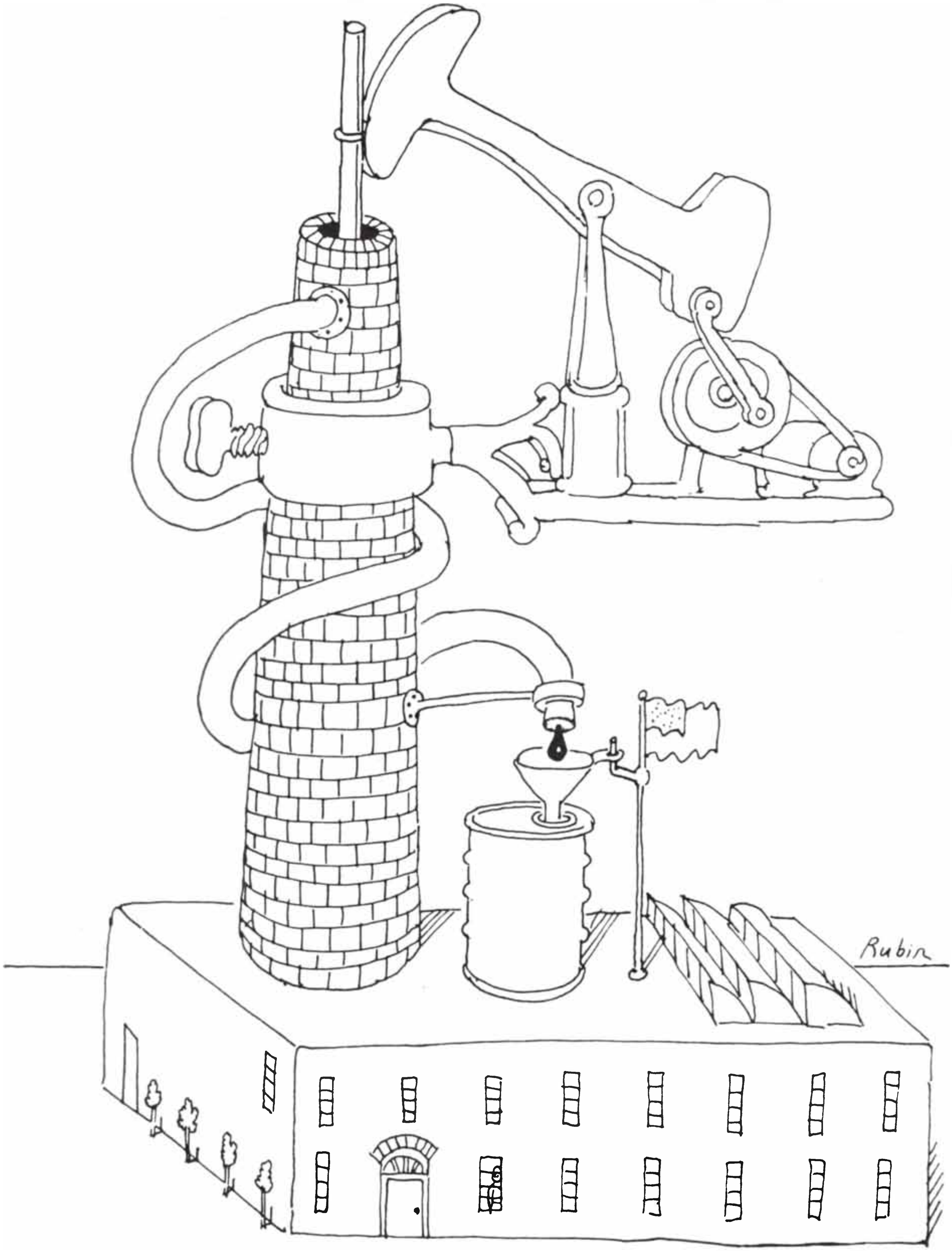


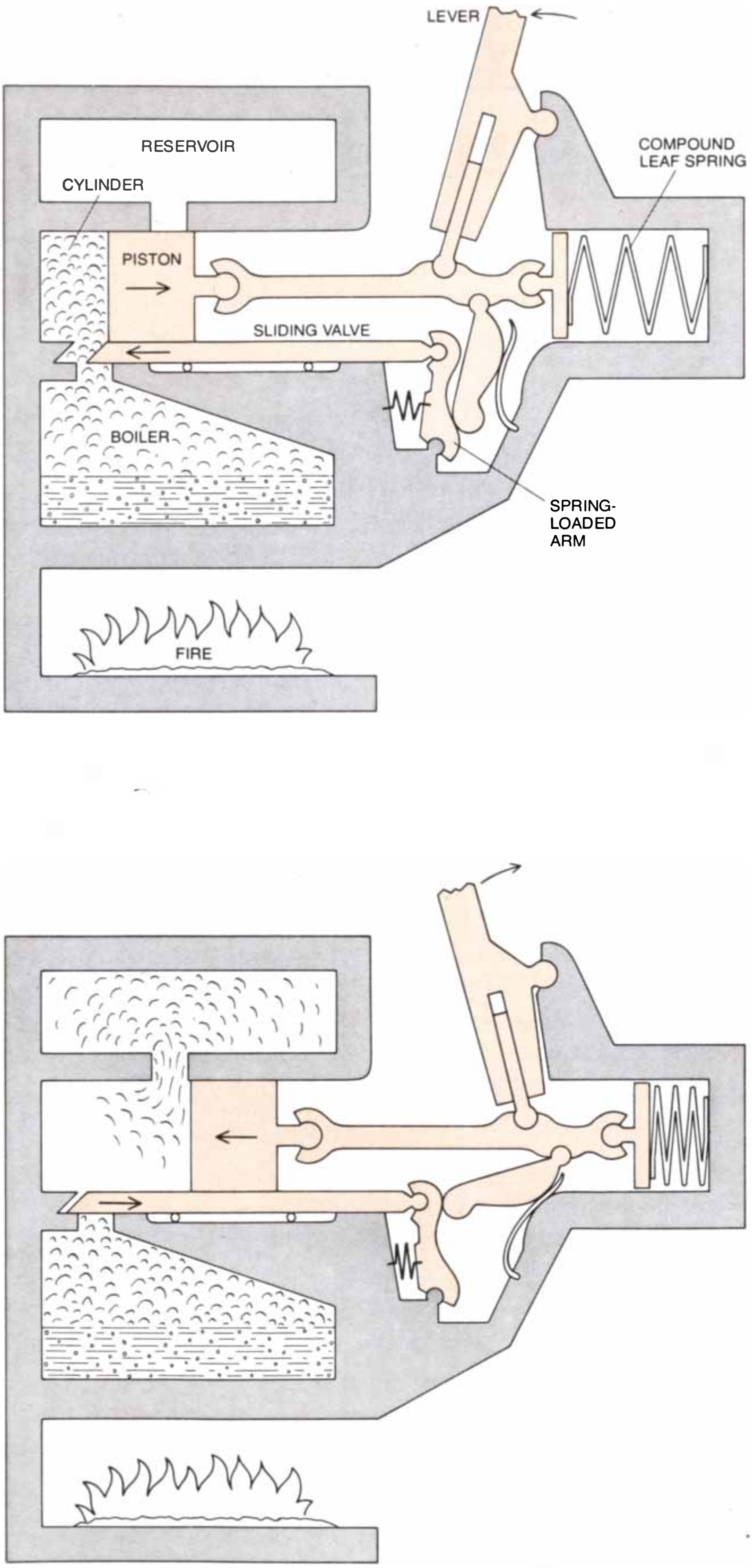
Basic planiversal mechanical elements


An Astrian faucet


An Astrian door-opening mechanism

Imagine a smokestack that turns heat into oil.





A planiversal steam engine

right of the cylinder to drive the piston back to the left (*bottom*). The sliding valve is closed as the steam escapes into the reservoir, but as the piston moves back it reopens, pulled to the right by a spring-loaded arm.

The illustration on the opposite page depicts Dewdney's ingenious mechanism for unlocking a door with a key. This planiversal lock consists of three slotted tumblers (*a*) that line up when a key is inserted (*b*) so that their lower halves move as a unit when the key is pushed (*c*). The pushing of the key is transmitted through a lever arm to the master latch, which pushes down on a slave latch until the door is free to swing to the right (*d*). The bar on the lever arm and the lip on the slave latch make the lock difficult to pick. Simple and compound leaf springs serve to return all the parts of the lock except the lever arm to their original position when the door is opened and the key is removed. When the door closes, it strikes the bar on the lever arm, thereby returning that piece to its original position as well. This flat lock could actually be employed in the steriverse; one simply inserts a key without twisting it. "It is amusing to think," writes Dewdney, "that the rather exotic design pressures created by the planiversal environment could cause us to think about mechanisms in such a different way that entirely novel solutions to old problems arise. The resulting designs, if steriversally practical, are invariably space-saving."

Thousands of challenging planiversal problems remain unsolved. Is there a way, Dewdney wonders, to design a two-dimensional windup motor with flat springs or rubber bands that would store energy? What is the most efficient design for a planiversal clock, telephone, book, typewriter, car, elevator or computer? Will some machines need a substitute for the wheel and axle? Will some need electric power?

There is a curious pleasure in trying to invent machines for what Dewdney calls "a universe both similar to and yet strangely different from ours." As he puts it, "from a small number of assumptions many phenomena seem to unfurl, giving one the sense of a kind of separate existence of this two-dimensional world. One finds oneself speaking, willy-nilly, of *the* planiverse as opposed to *a* planiverse.... [For] those who engage in it positively, there is a kind of strange enjoyment, like [that of] an explorer who enters a land where his own perceptions play a major role in the landscape that greets his eyes."

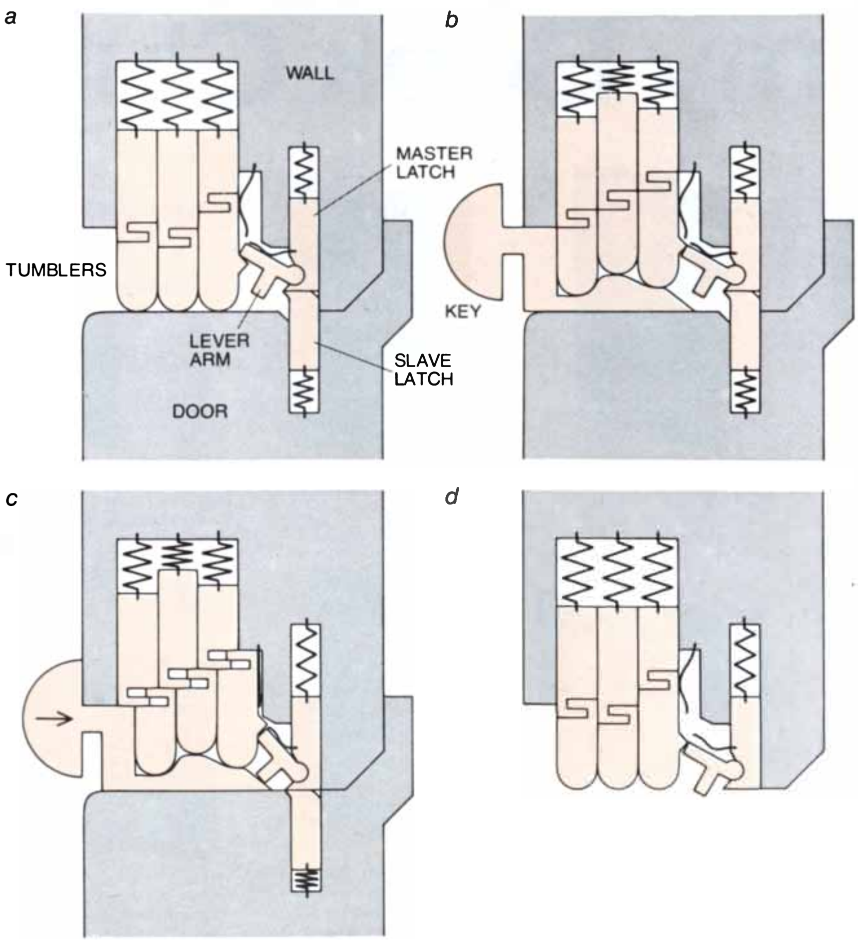
Some philosophical aspects of this exploration are not trivial. In constructing a planiverse one sees immediately that it cannot be built without a host of axioms Leibniz called the "compossible" elements of any possible world, which allow a logically consistent structure. Yet as Dewdney points out, science in our universe is based mainly on observations and experiments, and it is not easy to find any underlying axioms. In constructing a planiverse we have nothing to observe. We can only perform *gedanken* experiments (thought experiments) about what might be observed. "The experimentalist's loss," observes Dewdney, "is the theoretician's gain."

I urge all who want to participate in this delightful enterprise (or to learn more about it) to obtain Dewdney's monograph. He tells me he will send it postpaid for $5 to anyone in the U.S. or Canada, with an additional mailing charge for other countries. Checks or money orders should be made out to Dewdney and sent to him at the Department of Computer Science, University of Western Ontario, London, Ont., Canada N6A 5B9. If enough people become engaged in planning a planiverse, Dewdney intends to publish their contributions on an annual basis.

A marvelous exhibit could be put on of working models of planiversal machines, cut out of cardboard or sheet metal and displayed on a surface that slopes to simulate planiversal gravity. One can also imagine beautiful cardboard exhibits of planiversal landscapes, cities and houses. Dewdney has opened up a new game that demands knowledge of both science and mathematics: the exploration of a vast fantasy world about which at present almost nothing is known.

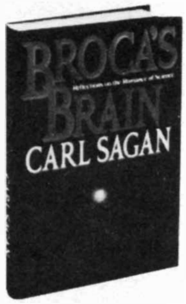
It occurs to me that Astrians would be able to play two-dimensional board games but that such games would be as awkward for them as three-dimensional board games are for us. I imagine them, then, playing a variety of linear games on the analogue of our 8-by-8 chessboard. Several games of this type are shown in the illustration on page 31. Part *a* shows the start of a checkers game. Pieces move forward only, one cell at a time, and jumps are compulsory. The linear game is equivalent to a game of regular checkers with play confined to the main diagonal of a standard board. It is easy to see how the second player wins in rational play and how in misère, or "giveaway," checkers the first player wins just as easily. Linear checkers games become progressively harder to analyze as longer boards are introduced. For example, which player wins standard linear checkers on the 11-cell board when each player starts with checkers on the first four cells at his end of the board?

Part *b* in the illustration shows an amusing Astrian analogue of chess. On a linear board a bishop is meaningless



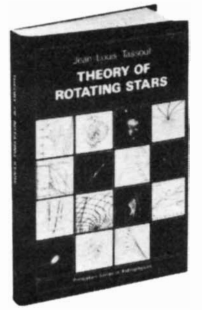
A planiversal lock and key

80200-2. STARS AND CLUSTERS. *Cecilia Payne-Gaposchkin.* Over 120 line drawings and 100 photos follow the life of a star from protostar to black hole. *Counts as 2 of your 3 books.* **$22.50**



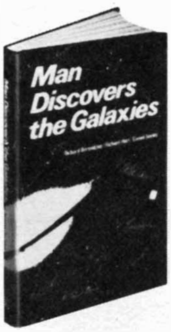
36875. BROCA'S BRAIN: Reflections On The Romance Of Science. *Carl Sagan.* Lucid account of recent astronomical discoveries. Debunks UFO's and other pseudoscientific ideas. **$12.95**

44930-2. THE ENCYCLOPAEDIA OF IGNORANCE. *Edited by Duncan and Weston-Smith.* 58 internationally renowned scientists highlight topics ranging from evolution to black holes. *Counts as 2 of your 3 books.* **$30.00**

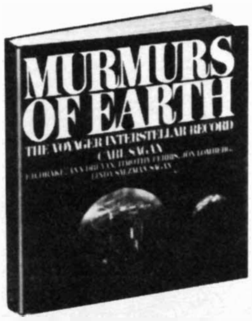


83850-3. THEORY OF ROTATING STARS. *Jean-Louis Tassoul.* The entire field of stellar rotation, from the basic observational data to theories on structure, stability and evolution of rotating stars. *Counts as 3 of your 3 books.* **$40.00**

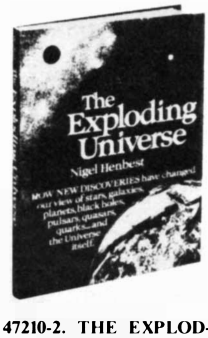
57373. KNOWLEDGE AND WONDER: The Natural World as Man Knows It. *Victor F. Weisskopf.* Revised classic primer of twentieth century science. From astrophysics and cosmology to particle physics, biochemistry and genetics. With 70 illustrations. **$15.00**



59800. MAN DISCOVERS THE GALAXIES. *Richard E. Berendzen et al.* A study of what is known about the galaxies, and how we came to know it. Over 200 illustrations. **$15.95**

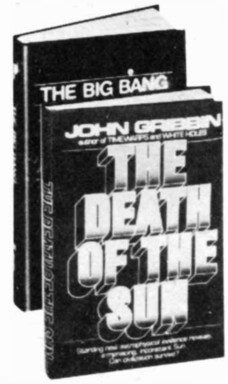


63217. MURMURS OF EARTH: The Voyager Interstellar Record. *Carl Sagan et al.* The story of the recorded messages carried by the *Voyager* space probe. **$15.00**

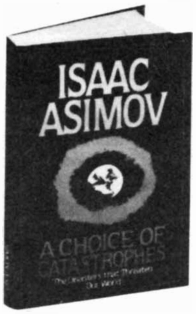


47210-2. THE EXPLODING UNIVERSE. *Nigel Henbest.* How new discoveries have changed our view of stars, galaxies, planets, black holes and more. Spectacularly illustrated. *Counts as 2 of your 3 books.* **$19.95**

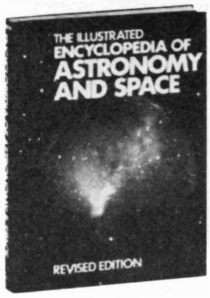
69303. PLANETARY ENCOUNTERS. *Robert M. Powers.* The case for all-out exploration of our solar system in this century. **$13.95**



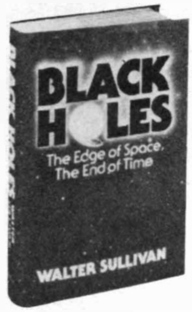
36170-2. THE BIG BANG/THE DEATH OF THE SUN. Two books. *First,* the creation of the galaxies, stars, and planets, and their probable fate. *Then,* the latest ideas on the origin, mechanics, and influence of our nearest star. *Counts as 2 of your 3 books.* **$26.95**



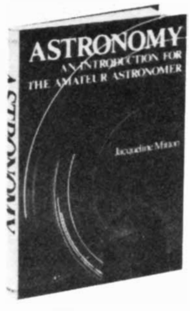
37953. A CHOICE OF CATASTROPHES. *Isaac Asimov.* Supernovae, ice ages, meteors, another Skylab, many more. **$11.95**



55001. THE ILLUSTRATED ENCYCLOPEDIA OF ASTRONOMY AND SPACE. **$17.95**

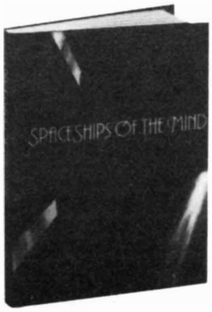


36400. BLACK HOLES: The Edge of Space, The End of Time. *Walter Sullivan.* **$17.95**

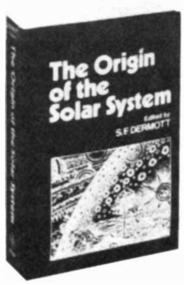


34661. ASTRONOMY: An Introduction for the Amateur Astronomer. *Jacqueline Mitton.* **$7.95**

55116-2. IN SEARCH OF ANCIENT ASTRONOMIES/THE NEW ASTRONOMY. Two books. *Counts as 2 of your 3 books.* **$22.95**

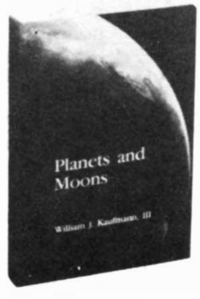


79735. SPACESHIPS OF THE MIND. *Nigel Calder.* Proposals for man's self-directed evolution, his colonization of space, and the restoration of the earth to a more natural state. **$14.95**



66275-3. THE ORIGIN OF THE SOLAR SYSTEM. *S.F. Dermott.* Planet formation, galactic processes, gravitational effects, and much more. *Counts as 3 of your 3 books.* **$39.00**

81403. SUN, MOON, AND STANDING STONES. *John Edwin Wood.* The latest findings from researchers in archaeology, astronomy, climatology, and physics. Illustrated. **$14.95**



69327. PLANETS AND MOONS. *William J. Kaufmann III.* **$13.00**

and . . .

34683-2. ASTRONOMY OF THE ANCIENTS/GALAXIES AND QUASARS. Two books. *Counts as 2 of your 3 books.* **$25.50**

If the reply card has been removed, please write to
Astronomy Book Club
Dept. L-AC6, Riverside, N.J. 08370
to obtain membership information and an application


CHEESE BOUTIQUE
Fromages de France
9 Caledonia
OPEN
grande
grande

and a queen is the same as a rook, and so the pieces are limited to kings, knights and rooks. The only rule modification needed is that a knight moves two cells in either direction and can jump an intervening piece of either color. If the game is played rationally, will either White or Black win or will the game end in a draw? The question is surprisingly tricky to answer.

Linear go, played on the same board, is by no means trivial. The version I shall describe was invented 10 years ago by James Marston Henle, a mathematician who is now at Smith College. Called pinch by Henle, it is published here for the first time.

In the game of pinch players take turns placing black and white stones on the cells of the linear board, and whenever the stones of one player surround the stones of the other, the surrounded stones are removed. For example, both sets of white stones shown in part *c* of the illustration are surrounded. Pinch is played according to the following two rules.

Rule 1: No stone can be placed on a cell where it is surrounded unless the move serves to surround a set of enemy stones. Hence in the situation shown in part *d* of the illustration White cannot play on cells 1, 3 or 8, but he can play on cell 6 because this move serves to surround cell 5.

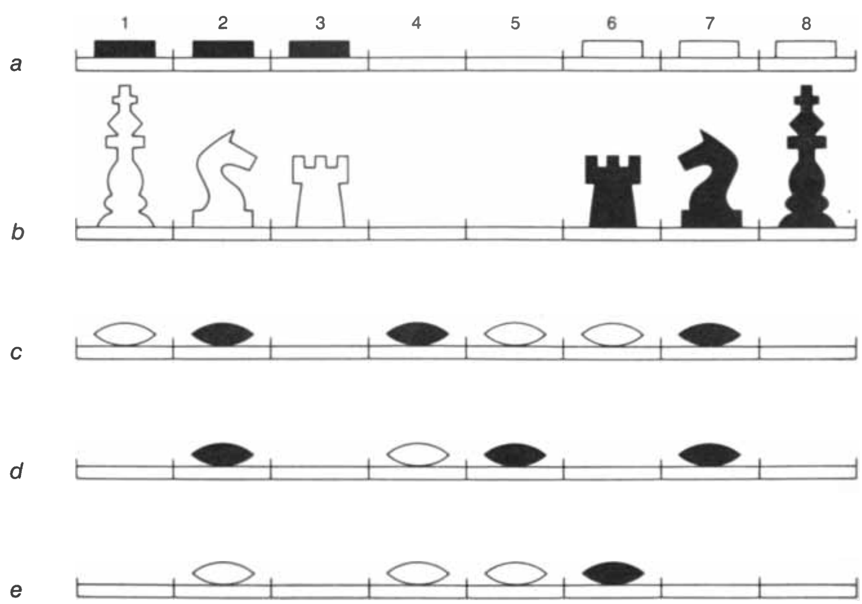
Rule 2: A stone cannot be placed on a cell from which a stone was removed on the last play if the purpose of the move is to surround something. A player must wait at least one turn before making such a move. For example, in part *e* of the illustration assume that Black plays on cell 3 and removes the white stones on cells 4 and 5. White cannot play on cell 4 (to surround cell 3) for his next move, but he may do so for any later move. He can play on cell 5, however, because even though a stone was just removed from that cell, the move does not serve to surround anything. This rule is designed to decrease the number of stalemates, as is the similar rule in go.

Two-cell pinch is a trivial win for the second player. The three- and four-cell games are easy wins for the first player if he takes the center in the three-cell game and one of the two central cells in the four-cell one. The five-cell game is won by the second player and the six- and seven-cell games are won by the first player. The eight-cell game jumps to such a high level of complexity that it becomes very exciting to play. Fortunes often change rapidly, and in most situations the winning player has only one winning move.

Next month I shall disclose which player has the win in eight-cell pinch and answer questions about the other three linear games.

Last month's problem was to determine whether the three models of a 4-group were examples of the cyclic 4-group or the Klein 4-group. All three are Klein 4-groups. An easy test for determining the character of a given 4-group is checking to see whether each operation in the group is its own inverse. If it is, the group is a Klein 4-group.

The two groups of order 4 provide the most basic example of the way simple groups, the building blocks of all finite groups, do not behave like primes. Both 4-groups include the same set of simple subgroups: the identity group and three order-2 groups, each consisting of the identity operation and one of the other three operations. Yet the two 4-groups are not the same. In other words, although the product of a set of primes is always a unique number, the same set of simple groups can combine to create more than one composite group.



Linear checkers (a), *chess* (b) *and go* (c, d, e)

BOOKS

From two million cases of smallpox to none, and reactions to the coming of the white men

by Philip Morrison

THE ERADICATION OF SMALLPOX FROM INDIA, by R. N. Basu, Z. Jezek and N. A. Ward. South-East Asia Regional Office, World Health Organization, WHO Publications Centre U.S.A., 49 Sheridan Avenue, Albany, N.Y. 12210 ($21.60). Lakshmi, beautiful consort to great Vishnu, is the generous source of prosperity and good fortune. Born in the Ocean of Milk, she has placed her sign in milk. No one who seeks her favor would turn aside any offer of a foamy beaker of milk, set at every good meal in all India. There are less benign deities in the pantheon of the subcontinent. Mother Shitala is represented as a woman of commanding presence, large-eyed and grave. She rides the land past rich and poor on her humble vehicle, the donkey. In her right hand she holds a pitcher of water, in her left a dry broom. On her head she balances with womanly dexterity as Indian as the summer monsoon a basket filled with grain. As she rides she may happen to shake her powerful head. Each grain she spills turns into a smallpox pustule; her marked victim will survive if she cleans away the spilt grain with a swish of water, but if she uses only the dry broom, the pox will take a life. Her temples dot the land; she has been propitiated in painted image whenever in India smallpox has come, surely for thousands of years.

The bounty of Lakshmi is one outcome of the ancient symbiosis between cattlekind and humankind, important in India since the Neolithic. The ugly burden imposed by arbitrary Shitala may well, however, be another element of that same tie between our two mammalian species. It is at least an attractive surmise that variola, the smallpox virus, first came to mankind from the cattle we have so long tended, after some particular modifying genetic event. Smallpox is Old World, unknown in the Americas before Columbus, and deadlier there than all the cannon and cavalry of the conquistadors. It is invariably virulent in man, without any carriers or invisible infection; what you get, you see. That bespeaks the novelty of the parasitism; there has been no time for the evolution of an enduring temperate relation between the virus and the host.

Smallpox needs a sizable population to offer new hosts; otherwise it dies out through the long immunity it confers on those who recover. Hunters might have acquired the new virus from other primates, in which similar pox viruses are known. But hunting bands are small, and their contact with the other primates is very old. Rather, the cowpox vaccine that in the experience of Americans has ended smallpox gets its very name from the gentle milch cow, whence Edward Jenner believed he had taken a mild surrogate for the dread disease. Nowadays it is thought the modern strains of the cowpox vaccinia virus are variants of the human variola virus itself, although they are still harvested in quantity from the skin lesions of inoculated calves or sheep, or from growing chick embryos.

The news is victory. Shitala Mata has decided to hold her head uncommonly steady these days. Across the entire country, from alpine Ladakh up against Tibet to the forested Andaman Islands in tropical seas, not one grain is spilt. No case of smallpox is to be found in all India. The last-reported smallpox patient there was discharged from the hospital on July 4, 1975; her pocked face confronts the reader. Saiban Bibi, a migrant beggar woman from Bangladesh, developed a fever and a rash "while living on the platform of the Karimganj railway station." Shivering with ague, subsisting for a few days on the tea she could get from the third-class waiting-room stall 10 meters away, she finally walked to the hospital, to enter history.

Polyglot and wonderfully diverse, some 600 million people live under the rule of Delhi, in about 115 million dwellings of every kind. Many millions wander the roads and the streets, pilgrims, migrants, homeless. Thirty years ago some 250,000 cases of smallpox were reported there every year, the majority of all the world's human hosts to variola. There is excellent reason to estimate that the cases were then massively under-reported; it is likely there were more than two million sufferers each year, a fourth of them fatal. A graph shows the cases reported from 1950 on; there are peaks every four to seven years, time to accumulate a pool of new susceptibles "following the high birth rate" in densely populated areas.

In 1961 war was declared on smallpox by India: the National Smallpox Eradication Programme was set in motion. Hopes were high; the U.S.S.R. gave India half a billion doses of its newly developed freeze-dried vaccine. Smallpox remained vigorously endemic. By 1966 there had been half a billion vaccinations, but the case rate had not materially decreased. To be sure, there were many states, particularly in the south, where smallpox was now rare and almost entirely imported from elsewhere. But mass vaccination could never reach all the teeming inhabitants; the disease prospered among the many who were missed.

The Intensified Campaign began in 1973. WHO had fixed on and demonstrated a new worldwide strategy; India was the most challenging theater of the war, particularly the populous Gangetic valley. The plan was powerful. Its base was the sure intelligence that this disease cannot hide; its spread is solely by close, although not necessarily intimate, human contact. Every infection shows up within a couple of weeks, and people can spot the fever with rash every time. True, it takes more skill to distinguish smallpox from other contagious rashes, but in any case the alarm has been sounded. The logistics were favorable; the old liquid vaccine, temperature-sensitive and often nothing but glycerine and dead virus, was no longer in service. Four Indian centers were in full production by 1973, making enough of the new quick-frozen and vacuum-dried vaccine to serve India and four of its neighbors. A central laboratory in Delhi and a WHO laboratory abroad monitored the output against world standards of potency and bacterial sterility.

Finally came the bifurcated needle, the right weapon for the front-line vaccinator. It is sparing of ammunition; a quarter of a cubic centimeter of vaccine could support a long day's work for an energetic young person, more than 100 vaccinations. The take rate was close to 100 percent among primary vaccinees. A vaccinator moving from house to house needs only a vial of diluted freeze-dried vaccine and a plastic tube full of bifurcated needles. With a new vial of vaccine and the same needles, simply boiled after hours to sterilize them, the soldier is armed for another day. "In 20 minutes an untrained worker could be converted into a skilled vaccinator."

The tactical plan was clear, soundly based on epidemiology. Search out every outbreak and then contain it. When the season of disease opened in the spring of 1973, the forces were ready. Their first task was a periodic national search for cases. In each of the nearly 400 administrative districts of India the primary health-center staff was drafted for search, a week per month in endemic states, less in the nonendemic areas. Out they spread, each with a photograph showing a smallpox-affected child, to ask city people and villagers, teachers, mailmen, doctors of ayurvedic medicine and leaders of every kind if they knew of a case of smallpox. The weekly markets were regularly targeted, together with "bus stands, . . . schools, railway stations, tea and betel shops, mosques, temples and migrant camps and brick kilns."

One hundred and fifty thousand people took part in the searches. They did not neglect travelers on the road, dockside, rail platform, forest camp and river island. A typical district includes some 200 villages, about the equivalent urban population in a dozen municipal wards. Once an outbreak was found (any village or ward with one or more cases in the preceding six weeks was defined as an outbreak) it was contained. Came the vaccinators, one full-time person for 20,000 people on the average, and one noncom, an experienced supervisor, for each squad of four. They followed the chain of which the reported case made one link. The people in the nearest houses were all vaccinated, if possible within 24 hours. All other contacts that could be established were followed up for vaccination; if the chain left the district, urgent notice was sent in person or by wire. As long as outbreaks were many, these measures were necessarily minimized. At first only 50 nearby houses could be visited with needles; then as cases dwindled the number of houses grew to 200. Later the entire village was included in the treatment, and finally a 10-mile radius was.

By the end of 1973 the maps show the endemic regions in a clear light. Four populous states held 80,000 cases among them; the rest of the country had only 5,500. In the little-affected areas the campaign could go on the offensive. There house-to-house searches could be undertaken in every village that reported an outbreak, instead of relying on the reports of leaders. Forces concentrated in the worst areas. The local staff was reinforced, often with international epidemiologists and special mobile teams. There was dramatic progress after the monsoon of 1974, yet in the worst places a bitter fight remained. Visitors came too often to affected houses; even patients were found to move from place to place. Now, however, bigger teams of vaccinators could be sent to any outbreak. Local watchmen were hired and posted in pairs at the door on a 24-hour basis. Food was provided for the family within. By the end of 1974 the map shows outbreak dots in only a few dozen districts.

The year of Operation Smallpox Target Zero was 1975. It is the chief pleasure of this well-illustrated, detailed and candid volume that the authors offer persuasive evidence that no smallpox case was left in India after the summer of that year. The argument springs really from continuity. Down the outbreak numbers steadily fall. The workers, no longer vaccinating so much, can post or stencil signs on every school, indeed on 10 or 15 million houses in all India, offering a cash reward for notice of a smallpox case. The reward grows; the payoffs are progressively fewer. The smallpox workers seek cases of fever with rash, and they see plenty of them. But these patients have chickenpox, not smallpox, and the virologists confirm the judgment.

A million cases of fever and rash, mainly chickenpox and measles, were reported in 1975; hundreds of them resulted in death. But the presence of a smallpox vaccination scar, the determination of contact with someone else suffering from another infection or other evidence made it clear that smallpox was not implicated in a single fatal case. In 1972 one district had deliberately concealed its epidemic of more than 1,000 cases. Of course, the event finally came to light, and the "transmission was arrested." The rewards for case notification were extended to reward both the public informant and the health worker who made the first report. In another district chickenpox was the diagnosis for more than a dozen suspected smallpox outbreaks. The verification of the diagnoses was then set into routine operation.

The great steel complex at Jamshedpur became for the month of May, 1974, "the world's greatest exporter of smallpox." The steel towns are prosperous; they attract beggars, transients, poor relations. A third of the third-class passengers at the station were unauthorized, ticketless travelers whom no one tried to stop. Responsibility was diffuse. Once the problem came to light two months of stringent control—barricaded roads, checkpoints, companies asked to deny travel leave to unvaccinated employees and their families—ended the matter. Finally, long after public notice of the last case in India, the coroner's court in Bombay declared after an autopsy that a four-year-old boy had died from "complications of smallpox." That protocol of autopsy was done in secret by the police. Burial was at noon one day. By five in the afternoon the international commission of assessment had asked for exhumation of the body. Fluid was collected from some 50 vesicles. The virus laboratory reported the samples negative for smallpox. The fatal illness was chickenpox; the Bombay coroner was wrong.

The war is won worldwide, at least for a good while. Smallpox is possibly to be found somewhere in the war-torn Horn of Africa; it is under guard in some laboratories and nowhere else as far as is known. In India even the laboratories made their final gesture of peace; all four research centers known to hold variola virus destroyed their stocks. All other medical schools and virus laboratories in India have been checked, and all published literature issuing from India on smallpox has been examined to locate other possible virus workers. There is no known virus left. A tragedy like that of Birmingham—escaped laboratory virus—cannot happen in India. Maybe there is some wild primate vector, although it is not expected. Smallpox has been routed in its ancient home; Shitala Mata will ride head high, unseen forever.

It is a grand triumph. Science, reason, energy and devotion have won. There were of course those who resisted vaccination. Indeed, there is a real risk of death from vaccination (one in about 10 million in Indian circumstances), although those who resist it do not depend on such arguments. Most of the 3 percent of initial resisters were women past 70. Persuasion, tact, the use of local workers, of senior staff or of female team members, particularly in Muslim areas, generally overcame their refusal. In the end resistance had no substantial effect.

After a stringent series of all-India verification surveys an international commission declared India free of smallpox on April 23, 1977. They had good reason. This volume is the required record of the entire campaign; the three authors are themselves veterans of the war, Delhi-based officers of WHO. (They are surely among the group here photographed in handsome T-shirts inscribed Smallpox Zero.) The cost was $95.4 million, the auditors say, with a saving for India of $91 million per year, not costing out bereavement, disfigurement and pain.

Human beings are able to stand united against a virus. Divided by nation, caste and class, they are less able to ameliorate more human problems. Saiban Bibi is immune to smallpox now; one wonders if she is yet immune to sleeping on railway platforms. Still, there rises from these lively pages, with all their maps, forms and graphs, the distant fragrance of rational hope.

CHRISTO: RUNNING FENCE, SONOMA AND MARIN COUNTIES, CALIFORNIA 1972–76. Narrative text by David Bourdon. Photographs by Gianfranco Gorgoni. Chronicle by Calvin Tomkins. Harry N. Abrams, Inc., Publishers ($200). The inception is brilliantly documented. "RUNNING FENCE project for the West Coast. Woven synthetic fabric forming Fence, attached to an upper steel cable and to 20 feet high poles, spaced 60 to 80 feet, running 18 to 20 miles, from the Pacific Ocean towards Inland, through the Landscape, over the crest of the Hills, into a small City, along and across the Highways, county and dirt roads." The pencil-and-crayon drawing of the concept over the typed

The known universe is expanding at the rate of 40 trillion cubic light-years per second. And, not only is the universe physically expanding, but our understanding of it is undergoing an unprecedented explosion as well: we have gained more new knowledge of the universe in the last two decades than in ALL PRIOR HUMAN HISTORY!

The pace of discovery is now so rapid that most — if not all — of the astronomy books in your local library are badly out-of-date — even if they are only a few years old!

Star & Sky is a magazine that pays particular attention to the fast-breaking news and keeps you on top of it like an astronomy newspaper . . . in it you can find consistently accurate, detailed reports of what's going on today (and *tomorrow*) in space exploration, astronomical research and discovery.

The January 1980 issue of **Star & Sky** broke Richard C. Hoagland's report on the possibilities of life on the Jovian moon, Europa. Details unravelled the facts of how the findings of the recent Voyager spacecraft explorations indicated that Europa is likely to have an ocean of water under its icy surface.

Hoagland traces back to Europa's beginnings, when it may have had ice-free global oceans — just as Earth did billions of years ago. This — combined with evidence from the Voyager findings, and from the bottom of the oceans right here on Earth — makes Europa the solar system's most likely abode for alien life.

This lengthy and abundantly illustrated article created a furor that reverberated well beyond the astronomical community. Worldwide press coverage soon followed, including front-page articles in ten of the twenty largest newspapers in America.

By subscribing now, you can receive the January 1980 issue, with this highly acclaimed report on Europa, absolutely FREE

Here is what two scientists — both involved with the search for life beyond Earth — had to say about the article:

". . . a very fine piece of science writing, but it also presents an idea that is, to my knowledge, completely original and that seems to me to have great merit."

Robert Jastrow
NASA Institute for Space Studies
New York, NY

". . . one of the finest pieces I've ever seen in an astronomy magazine. . . . He (Hoagland) produced an article stimulating to amateur and professional astronomer alike."

Michael H. Hart
Trinity University
San Antonio, Texas

Star & Sky has reporters at mission control for the latest planetary encounters so we can bring you the detail and full color illustrations that the general science magazines don't have room for. We reported

David Healy

on what the Voyager and Pioneer spacecraft actually found at Jupiter and Saturn and in later issues augmented it with late-breaking data emerging from further analyses.

We cover the scientific meetings at the world's great observatories and universities to bring you details of the most recent discoveries about quasars, galaxies, black holes and such mind-stretching concepts as the origin and destiny of the universe. In addition, our monthly report on manned space activities by James Oberg at NASA's Johnson Space Center ensures that you won't miss a thing.

UNLIKE ANY OTHER MAGAZINE IN THE FIELD we often publish Special Issues — entire issues devoted to in-depth treatment of single topics.

These *Special Editions* give you the most thorough, in-depth and up-to-date accounts available anywhere: the facts, the findings, old and new theories, photos, computer-generated art, schematics and diagrams, upcoming projects and the riddles they hope to solve — all presented to you by the greatest astronomy minds on Earth!

Star & Sky is a blend of visuals and precise but easy-to-digest text. Jargon is held to a minimum yet we don't skimp on comprehensive coverage.

And art: What does a black hole look like? How does it work? How do the rings of Jupiter, Saturn and Uranus differ? Where photography is impossible, we turn to art. Intelligent renderings — made by astronomer-artists who keep abreast of the newest theories and discoveries — portray visually the probabilities and possibilities we can't see.

Star & Sky brings you top authors in both professional and amateur astronomy:

Welcome to the expanding universe...

Isaac Asimov, James Oberg, Richard C. Hoagland, William J. Kaufmann, Barry Parker, Wallace Tucker, John Gribbin, Patrick Moore, John Sanford, Leo Henzl, Terence Dickinson, Ray Villard, Joel Davis, Trudy E. Bell, George T. Keene, James Mullaney, and many others.

Star & Sky covers the entire field of astronomy: IT IS THE COMPLETE REFERENCE FOR EVERYONE INTERESTED IN THE UNIVERSE.

FOR THE PRICE OF A TYPICAL HARDCOVER BOOK YOU CAN ENJOY 12 DETAILED AND UP-TO-DATE ISSUES OF STAR & SKY

And as a SPECIAL OFFER available to Scientific American readers, the BONUS JANUARY 1980 Issue, containing Richard C. Hoagland's headline-producing article, is yours — FREE — by subscribing to Star & Sky — if you enclose payment with your order, we will send your BONUS ISSUE out immediately. If you prefer to be billed, we will send out your BONUS ISSUE when payment is received.

Act today to reserve your BONUS ISSUE — supplies are limited.

STAR & SKY Magazine, P.O. Box 5172, Westport, CT 06881

I don't want to miss any reports of Voyager's encounter with Saturn; the first launch of the Space Shuttle; preparations for the return of Halley's Comet; plans for the launch of a 94-inch telescope into Earth orbit; etc., etc. . . . but I especially don't want to miss any future **Special Issues** *on important and timely topics in astronomy!*

☐ Yes, send me my BONUS ISSUE right now. I am enclosing my check or money order.

☐ Please bill me. I understand that when my payment is received, Star & Sky will send me my BONUS ISSUE.

☐ 3 years for $33 (36 issues) — a 54% savings over the newsstand price.
☐ 2 years for $27 (24 issues) — a 44% savings over the newsstand price.
☐ 1 year for $15 (12 issues) — a 38% saving over the newsstand price.

Canada and Mexico add $3 — all other foreign add $6 PER YEAR OF SUBSCRIPTION.
Allow 6-8 weeks for delivery of first issue. (Subscription must begin with current issue.)

1SA780

Name ____________________

Address ____________________ Apt. # ________

City ____________________ State ________ Zip ________ label "Christo, October 1972" is convincing; the long sweep of the curtain over the tawny, rolling lands and then out of sight along the ridgeline is a pretty good representation of the reality, which began to grow in two counties in the summer of 1976. In the end it stretched out of Bodega Bay 24.5 miles across the landscape, from noon of September 10 until its complete removal on October 23, 1976. Christo Javacheff and his wife Jeanne-Claude, the principals of the Running Fence Corporation, had finished what they set out to do. They created a beautiful work of art, both visual and social, another Christo project of his uniquely playful and aesthetic engineering.

The money? Working capital came from the cosponsors, museums and collectors, who were solicited with the proposal in 1973 and were offered a hefty discount off the market price on some drawing or collage by Christo. The bottom line came to about $3 million, most of it for materials and labor, construction vehicles and legal, engineering and insurance fees. The physical fence was some 2,000 big panels of woven nylon fabric (surplus from early projects for air bags in automobiles, and a bargain). Each pure white panel was 18 by 68 feet and weighed less than 60 pounds. A panel was fitted on all four sides with grommets, to hang on hooks from steel cable spanning steel posts each 18 feet high (plus three feet belowground).

Those posts did not stand in concrete footings; each was supported vertically by a pair of "shoe angles," bolted steel *L* beams that kept the post from sinking. Guy cables braced the posts laterally; a single long, tapered panel ran from the beach down into the sea, where it was strongly anchored. There were a few special panels with archways cut to pass a road or free a parking lot here and there. Ninety miles of cable was used. "At the end of the viewing period, the anchors were driven below plow level; all holes were backfilled with sand and reseeded with grass." All other materials were removed and given to the ranchers whose land the fence had crossed; the county experts inspected the entire site with approval, and on December 14 the Sonoma County Landmarks Commission designated Pole No. 7–33 as Historic Landmark No. 24.

This book, made in the luxurious and handsome manner of well-illustrated presentations of the work of an artist or an architect, records the project. It begins with the chronology of events, includes an adroit essay on the entire tale and continues with documents, drawings, letters, invoices, reports, calculations, hearings. There is even a page-size sample of the fabric itself. The last half of the big volume mainly presents color photographs of the fence both in construction and during the time of viewing, seen as a work of art, flame-touched by the sunset, filled delicately by the wind, a mark of hand and mind across the land, evocative of drapery, regattas and the Great Wall of China.

The engineering was modest but real. There were desert tests of poles and fittings, and plenty of worry about the seagoing end. The route was crucial; it was laid out by the artist himself, who walked the fence line for three weeks in the fall of 1974. He carried a bag of wood stakes, and he dropped one every few hundred feet for the crew that followed. They hammered in the stakes at the future post intervals of 62 feet; the stakes and posts were numbered and the terrain was recorded. The route was not traveled sequentially; Christo worked rather with the landholders at their convenience, with concern for their livestock and their vehicles. Sharp curves were structurally difficult (nine feet offset from pole to pole was the maximum without special anchors). Christo made frequent revisions as the surveyors and the contractor came along his original zigzagging trail.

The 50-odd landholders along the line were natural proponents of the work. Indeed, they were won over first by the personal sincerity of the Christos and then by the audacity and beauty of the task. The 60 skilled professional hard hats who worked at driving anchors, laying cable and erecting poles enjoyed the project, and so did the 360 extras, the young men and women who were taken on for fabric-panel installation in late August. For them, many caught up in the counterculture way of life, the hard, sharp work and the beauty it created made a little epiphany.

There were plenty of opponents, including even artists, together with scientists, certain conservationists and resident squares. The corporation employed nine lawyers, and in the end they satisfied 15 governmental agencies, from both the counties and the state, with highway, fish and game, forest, water quality, land and coastal jurisdictions. One transgression of the laws marked this project: the permit to use the coastal section was granted by the regional division, but an appeal against it was taken to the statewide commission. That appeal meant the automatic revocation of the permit, and fatal delay. Christo boldly went ahead; the application for an injunction sought by the commission to uphold its powers against this vulnerable, delightful and entirely transient project was carefully read by a wise judge with his blind eye. He set his hearing weeks later; by then the fence had flowered and gone. "I completely work within American system by being illegal, like everyone else," Christo said. "And no make-believe, remember. We challenge, and we pay the consequences."

This volume is a fine record for anyone who wants to see into how things are done in our society, how the world works, both the newer world of environmental-impact reports and the older world of wire, wild flowers and wind. It lacks only one thing: it has no visual report of the death of the fence, which is surely part of its life. If you would enjoy your share of this contribution to the G.N.W.B. (the gross national well-being), seek out this expensive volume. Let the artist lower the curtain: he was watching a quarter mile of cloth rippling in the moving air. "'Look there, Ted,' he said, his voice rising in excitement. 'See how it describes the wind.'"

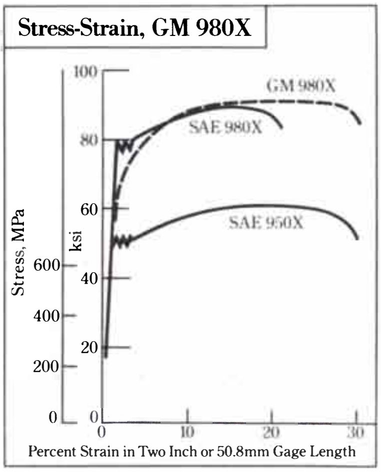
THE WHITE MEN: THE FIRST RESPONSE OF ABORIGINAL PEOPLES TO THE WHITE MAN, by Julia Blackburn. The New York Times Book Co. ($14.95). "When I was a boy, the Sioux owned the world; the sun rose and set on their land; they sent ten thousand men to battle. Where are the warriors today? Who slew them? Where are our lands? Who owns them?" From the days of Captain Cook through the 19th century of European colonial expansion and the irresistible westering of the U.S. up to the post–World War II opening of New Guinea and Melanesia hundreds of smaller societies have known apocalyptic times. Troubled minds sought meaning. The undreamed-of white men suddenly arrive; their rule is based on strange secrets of artifact and action, on writings of power, on new weapons and new beasts. The strangers must be fitted somehow into a cosmology that had held no such beings; their coming "brought . . . countless and hitherto unknown ways of dying," from new disease to hopeless battle. Finally, the changed life can be managed to allow spiritual survival notwithstanding the mysterious presence.

This book is a poignant and beautiful anthology, remarkably fresh in its materials, collecting around each of five meaningful themes a chapter of some 10 texts as told by those to whom intruders once came. The cultures cited do not include those great states of the Andes or the Indus or the Yangtze where long ago men from Europe confronted societies of commensurate weight, even though they lacked the cannon or the rapacity of the venturers. That is a longer and richer tale. Here we read texts by people within the small polities, none larger than some kingdoms of South Africa, some as small and isolated as the foraging hunters of the Australian desert. The texts are all in translation, of course. They are somehow accidental; they are on record here only because someone who had the power of writing and detached observation could yet engage prophets and seers and chroniclers on other than hostile terms. In our times, when apocalypse does not seem remote from the most favored cities, these accounts are dazzling mirrors of our com-

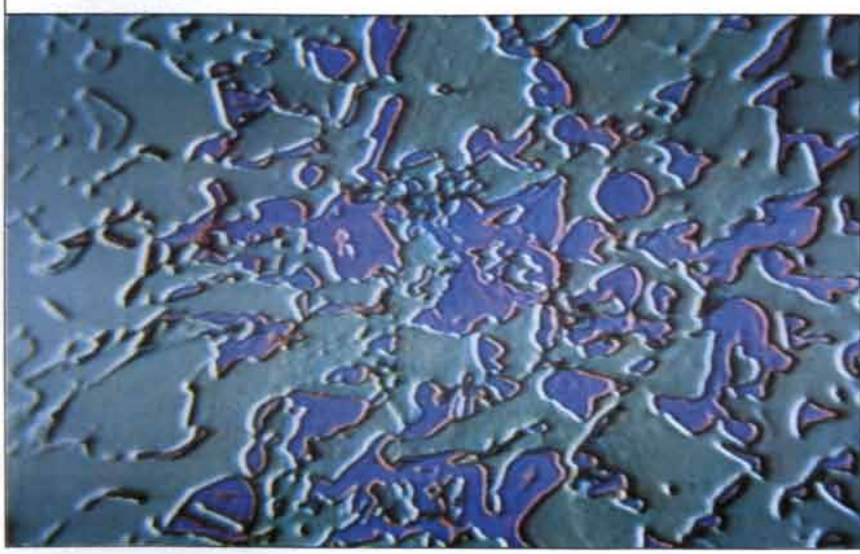

The Ductility Factor

The Ductility Factor

The use of high strength, low alloy steel has been severely limited, due to its low ductility. Now, a simple heat treating and controlled cooling process, developed at the General Motors Research Laboratories, has successfully enhanced formability properties without sacrificing strength.



A comparison of the stress-strain behavior of GM 980X, SAE 980X, and SAE 950X steels. GM 980X offers greater ductility at the same strength as SAE 980X, and greater strength at the same ductility as SAE 950X.



Scanning electron microscope micrograph of dual phase steel at a magnification of 2,000. The matrix (background) is ferrite; the second phase is martensite.

For some time, automotive engineers and designers have been faced with the challenge of building cars light enough to get good gas mileage, but still roomy enough to comfortably transport four or five passengers. One technique which has proved fruitful is materials substitution.

Lighter materials, such as aluminum alloys and plastics and high strength, low alloy steels (HSLA), are being phased into new vehicle designs to replace certain plain carbon steel components. Each, though, has displayed inherent problems which limit its utilization.

Unlike plastics and aluminum, however, HSLA steels have the same density as plain carbon steel. Weight reduction is achieved because thinner sections (less volume) can be used to carry the same load. Since the formability (ductility) of most high strength steels is poor, though, it has only been possible to form simple shapes from it. This has severely limited the widespread use of HSLA steels (such as SAE 980X) for auto components. New hope for the increased utilization of HSLA steel has arisen, however, with the development of a new dual-phase steel, GM 980X, at the General Motors Research Laboratories.

General Motors is not in the steel business, and GM 980X is not a brand of steel. GM 980X is the designation for a type of steel displaying mechanical properties similar to those of the samples first formulated at the General Motors Research Laboratories. "GM" in the designation indicates that the steel is a variation of the conventional SAE 980X grade. In the standard SAE system for material identification, "9" designates that the steel is HSLA. "80" is the nominal yield strength of the metal in thousands of pounds per square inch. The "X" denotes a micro-alloyed steel—one containing on the order of 0.1% of other metals such as vanadium, columbium, titanium, or zirconium as a strengthening agent.

GM 980X displays the same strength, after strain hardening, as SAE 980X steel, but has far more ductility. This characteristic allows it to be formed into various complex shapes which were previously thought to be impossible with HSLA steels. The superior formability of GM 980X has substantially increased the utilization of HSLA steel in the manufacturing of automotive components such as wheel discs and rims, bumper face bars and reinforcements, control arms, and steering coupling reinforcements.

Dr. M.S. Rashid, discoverer of

the technique to make GM 980X steel, comments, "I was working on another project using HSLA steel, when I noticed that if SAE 980X steel is heated above its eutectoid temperature (the temperature at which the crystalline structure of metal is transformed) for a few minutes, and cooled under controlled conditions, the steel developed significantly higher ductility and strain-hardening characteristics, with no reduction in tensile strength."

FURTHER experiments proved that the key variables to make GM 980X are steel chemistry, heating time and temperature, and the rate at which the steel is cooled. Specimens of SAE 980X were heated in a neutral salt bath, then cooled to room temperature with cooling rates ranging from 5° to 14°C/sec. (9° to 26°F/sec.). Dr. Rashid notes, "We found that the maximum total elongation resulted when the cooling rate was 9°C/sec. (16°F), and the lowest total elongation resulted from the highest cooling rate (14°C or 26°F/sec.)."

GM 980X steel has a high strain-hardening coefficient or n value, accompanied by a large total elongation. The n value gives a measure of the ability of the metal to distribute strain. The higher the n value, the more uniform the strain distribution and the greater the resistance of the metal to necking (localized hour-glass-shaped thinning that stretched metals display just prior to breaking). Tests have proved that GM 980X distributes strain more uniformly than SAE 980X, has a greater resistance to necking, and thus has far superior formability.

"The superior formability of GM 980X compared to SAE 980X steel appears to depend on the nature of two microstructural constituents, a ferrite matrix (the principal microstructural component) with a very high strain-hardening coefficient, and a deformable martensite (the other crystalline structure) phase. In the SAE 980X, failure occurs after the ferrite becomes highly strained, but when the GM 980X ferrite is highly strained, strain is apparently transferred to the martensite phase, and it also deforms.

"Therefore, voids leading to failure do not form until after more extensive deformation has occurred and the martensite phase is also highly strained. Obviously, the exact nature of these constituents must be important, and any variations in the nature of these constituents could influence formability. This is the subject of ongoing research."

Dr. Rashid's discovery represents a significant breakthrough in the area of steel development. His findings have opened the door to a new class of materials and have completely disproved the commonly held belief that high strength steel is not a practical material for extensive automotive application. "At GM, we've done what was previously thought to be impossible," says Dr. Rashid, "and now we're hard at work to find an even stronger and more ductile steel to meet the needs of the future."

THE MAN BEHIND THE WORK

M.S. Rashid is a Senior Research Engineer in the Metallurgy Department at the General Motors Research Laboratories. He was born in the city of Vellore in Tamil Nadu (Madras), India, and attended the College of Engineering at the University of Madras–Guindy. He came to the United States in 1963 and was awarded a Ph.D. in Metallurgical Engineering from the University of Illinois at Urbana-Champaign in 1969. After a three year Post-Doctoral Fellowship at Iowa State University, he joined the staff of the General Motors Research Laboratories.

Dr. Rashid is continuing his investigations into the development of even more ductile high strength, low alloy steels. When not in the lab, he enjoys relaxing by playing tennis and racquetball with his wife, Kulsum.


"Last year I switched to rum.
This year I graduated to Myers's Rum."
White rums may be what you learn on. But Myers's dark rum will advance your education. It will teach you just how good tasting rum can be. Because with Myers's Rum you get a smoother, softer taste that comes from master-blending.
What makes Myers's precious imported rum cost more, makes Myers's taste better. In cola, soda, fruit juice or any of your favorite mixers.
MYERS'S
WORLD FAMOUS
• IMPORTED •
MYERS'S
RUM
Original Dark
THE PLANTERS' PUNCH BRAND
100% FINE
JAMAICA RUM
DISTILLED IN JAMAICA
UNDER GOVERNMENT SUPERVISION
FOR MYERS RUM COMPANY LTD.
NASSAU, BAHAMAS, SUCCESSORS TO
FRED L. MYERS & SON
FOUNDED 1879
KINGSTON, JAMAICA
WEST INDIES
750 ML (25.4 FL.OZ.) · 80 PROOF
IMPORTED AND BOTTLED BY FRED L. MYERS & SON CO. BALTIMORE, MD

mon humanity in the glare of inevitable change.

In 1786 the ships of La Pérouse came to the narrow Alaskan inlet of Lituya Bay; they stayed for four weeks to observe and to trade. In 1949 a Tlingit from Yakutat recounted the traditional story of that alien arrival, earlier recorded in 1886. "One spring a large party of men from the big village of Kaxnuwu went to get copper from the people at Yakutat. Four canoes were lost... and the first chief of the party was drowned. While the survivors were still mourning, two ships rounded the bay. The Indians thought they were two great birds with white wings, perhaps Raven himself. They fled to the woods. After a time they came back to the shore and looked through tubes of rolled-up skunk cabbage leaves, like telescopes, for if they looked directly at Raven they might turn to stone.... They were so frightened when thunder and smoke came from the ship that their canoe overturned....

"Then a nearly blind old man said his life was behind him, and he would see if Raven really turned men to stone. He... induced two of his slaves to paddle him to the ship. When he got on board, his eyesight was so poor that he mistook the sailors for crows.... He traded his fur coat for a tin pan.... The people were surprised to see the old man alive. They smelled him to make sure that he had not been turned into a land-otter man, and they refused to eat the food he had brought. The old man finally decided that it must be ships and people he had seen, so the Indians visited the ships and traded their furs." Some of these details are confirmed in the French records; oral tradition can hold high events for centuries.

The most tragic account is that of the powerful Xhosa. A group of 10 strange young men appeared to a young girl, Nongquase (the text is her court testimony), at the river's edge where she was fetching water. She was fearful, but her uncle Mhala could see that they were the dead ancestors of the Xhosa. They promised victory over the English, if the corn was burned and the cattle were killed to feed the dead. The ancestors will soon return, and "they will rise at different kraals with cattle, corn, guns and assegais, and they will drive the English out of the country and make them run into the sea."

Both grain and cattle were indeed sacrificed over the whole of the Xhosa region. Three-fourths of the people died of this self-inflicted starvation between the spring of 1856 and the spring of 1857. "They waited for two red suns to rise over the hills in the east. Then the heavens would fall on the English and crush them. But the sun rose and set as usual and the ancestors did not come. My uncle... went down towards the sea where he could try to live on roots and shellfish. And when we were starving, I often heard him say that he regretted killing the cattle and destroying his corn."

The volume includes a striking collection of photographs, some of them portraits of the very prophets whose texts are quoted, such as Short Bull, a leader of the Ghost Dance. He looks steadily across 90 years at the reader in grave and charismatic nobility. Most of the photographs are less closely related to the texts; they present fascinating sculpture and graphics from many peoples showing the white man and his works as they appeared to the artist. The layout and presentation are particularly attractive, and the author, an English student of oral tradition, has sought her material in a wide variety of unusual sources, carefully cited. The famous report on the Ghost Dance by James Mooney and the material on the Tlingit are well known, but not many general readers will have encountered the Xhosa story, related in a book published in Kingwilliamstown, not far from the port of East London in the Republic of South Africa. An air of timelessness surrounds this work; in these words ring echoes of the tragedies of Hellas.

THE ASTRONOMICAL COMPANION, by Guy Ottewell. Care of the Department of Physics, Furman University, Greenville, S.C. 29613 ($12). The top of the big page is a glossy black square, nearly a foot on edge, above a few inches of text on white paper. On the black ground, in an artist's lively freehand white line, is drawn a big sphere, represented by only a few sweeping circles. The reader is looking down on a transparent equatorial plane, sketched in with an open grid. "The pictures are white-on-black because, in the nature of astronomical illustration, we mostly have to show not matter but light." In careful yet informal projection we view the place in which we live on the grand scale; the sphere has a radius of one megaparsec, and we regard it from a few megaparsecs out. What we see is the Local Group of galaxies, chiefly our centered home, the Milky Way galaxy, and its counterpart in Andromeda, around each of them a little cluster of satellite galaxies. Maffei I and IC 6822 are outliers that stand well above and below the plane on lightly drawn stalks. The three dimensions are vivid; it is not a page we are inspecting but a spatial volume.

This floppy, large-page paperbound book offers more than 30 such spheres, an atlas of the glowing furniture of space. The largest sphere (shown with a little distortion to confess to paradox) depicts the whole of visible space, out to a radius of 15 billion light-years; the smallest, with a radius of a mere two astronomical units, is centered on the sun and shows the orbits of the inner planets, with a couple of small friends, the asteroid 433 Eros and a well-known periodic comet whose name is a "volley of syllables." A few sphere pages are spent on more abstract purposes, to define coordinate systems, to map the constellations and so on. Throughout the book there are about as many pages of text as there are of maps, and auxiliary sketches and diagrams abound.

The book is the work of an artist, an almanac maker and no mean poet. Ottewell yearly publishes a guide to the shifting celestial scene; this companion holds permanent material that goes beyond (outside?) the season's repertory of the unaided-eye heavens. It is an original and perceptive introduction to astronomy, by a man whose delight and talent is the imaginative but faithful representation of the visual world. The text is excellent, full and clear, with almost no formal mathematics beyond a sharp eye for geometry. Atoms and spectra and radio hardly enter; here is no synoptic textbook but a strongly visually centered place "to look in for explanations, for reference, or for mere enjoyment... for those who already enjoy either astronomy itself or the idea of it."

The precession of the earth's axis and its effect on the look of the sky are well discussed; the usual pole-circle map is augmented by a sphere drawing and by a very helpful graphical account of how much the stars appear to shift in each part of the sky. The moon's orbit, moonlight and earthshine, the geometry of eclipses and considerably more moon lore are painstakingly described. Knowing sketches and paragraphs aid each other; the tough geometry is here and there allayed by a poet's image. ("In those years when Indonesian volcanoes girdle the earth with stratospheric dust ... the eclipsed moon is a drop of almost black blood, or an Arkansas apple.") Comets and meteors are also fully and freshly treated. A detailed plot of star temperature with respect to luminosity (the Hertzsprung-Russell diagram) is included, with a couple of hundred particular stars marked on it. (The diagram is the work of Douglas Roosa.) There are lists of constellations, of planets and satellites, of star names and their meaning, of important lengths, bright stars, comets and meteor showers. Hardly an equal sign is to be found; there is no algebra, but there is geometry aplenty.

The entire work is a tour de force, the product of understanding and taste, not just another able reworking of a long pedagogical tradition. The data are quite reliable, although one could here and there take some exception to a number. (Supernovas are brighter than they are plotted.) The heavy covers are in meaningful and attractive color: the front cover shows our stellar neighborhood out to some 10 light-years, to Procyon; the back one, the stars in the wedge of sky called Taurus out to the clustered Pleiades.

Lose Citicorp Travelers Checks in Maputo and you're not up the Zambezi without a paddle.

Just contact our Maputo refund center and we'll have you back on safari before you know it.

If refunds are this easy here, imagine how easy they are in places you *have* heard of.

In all, we have more than 40,000 locations here and abroad where you can get fast refunds (with emergency funds in the U.S. through 8,100 Western Union offices, many of them open 24 hours a day, 365 days a year). We even have our own Citibanks in 64 countries.

Where can you spend Citicorp Travelers Checks? Where *can't* you? They're as welcome as money in literally millions of hotels, restaurants and stores around the world.

Citicorp Travelers Checks—backed by America's leading banking institution worldwide.

Citicorp. Travel the world with us.

CITICORP®
TRAVELERS CHECKS

NOTE: Travelers checks under our old name, First National City, will continue to be accepted indefinitely.

Established 1845 SCIENTIFIC AMERICAN July 1980 Volume 243 Number 1

A Ban on the Production of Fissionable Material for Weapons

Such a "cutoff" could serve two purposes: to stall the further buildup of nuclear arms by the present nuclear-weapons states and to prevent their spread to the non-nuclear-weapons states

by William Epstein

The suspension of the SALT II Treaty ratification proceedings in the U.S. Senate, following the military intervention of the U.S.S.R. in Afghanistan, does not spell the end of international efforts to restrain the nuclear-arms race between the two superpowers. Outside the bilateral SALT process various multilateral arms-control negotiations are either in progress or planned, and the Carter Administration has pledged the continued participation of the U.S. in these endeavors. For example, talks are currently under way among representatives of the U.S., the U.S.S.R. and the U.K. on a comprehensive treaty to ban all nuclear test explosions. Final preparations are also being made for the second five-year conference to review the operation of the Nonproliferation Treaty, scheduled to convene next month in Geneva.

Both the comprehensive test ban and the Nonproliferation Treaty were put at the top of the list of other important arms-limitation issues in the joint communiqué issued by the U.S. and the U.S.S.R. at the time the SALT II Treaty was signed last year. That document stated that "definite progress" had been made in the negotiations for a comprehensive test ban and confirmed the intentions of the two superpowers to work together with the U.K. "to complete preparation of this treaty as soon as possible." With respect to the Nonproliferation Treaty, President Carter and President Brezhnev merely noted "the profound threat posed to world security by the proliferation of nuclear weapons" and agreed that the nuclear-weapons states "bear a special responsibility to demonstrate restraint" in this regard; they also "affirmed their joint conviction that further efforts are needed" to strengthen the nonproliferation regime.

The two chiefs of state were clearly right in concluding that further efforts are needed, not least to ensure the success of the upcoming Nonproliferation Treaty Review Conference. The non-nuclear-weapons states are far from impressed by the results of SALT II, which even if it were to be ratified would not halt the production of a single new strategic weapon now under development and would allow each of the superpowers to add thousands of new strategic nuclear warheads to their already excessive arsenals. Although most of the non-nuclear-weapons states seem to regard the SALT II agreement as a step forward, they certainly do not consider it adequate compliance with Article VI of the Nonproliferation Treaty, which commits the three signatory nuclear-weapons states (the U.S., the U.S.S.R. and the U.K.) to a "cessation of the nuclear arms race at an early date."

In the circumstances what can be done to curb both "vertical" proliferation (the increase in the numbers and kinds of nuclear weapons in the hands of the nuclear-weapons states) and "horizontal" proliferation (the further spread of nuclear weapons to nations that do not already have them)? In my view, based on more than three decades of work on international arms-control matters, there are at present two measures that could have a salutary effect not only on the strategic-arms competition between the two superpowers but also on the attitude of the non-nuclear-weapons states toward the entire nonproliferation regime. One would be the early achievement of a Comprehensive Test Ban Treaty. The other would be a commitment on the part of all concerned parties—nuclear-weapons states and non-nuclear-weapons states alike—to start negotiations soon on a treaty to ban the production of fissionable material for nuclear weapons.

It is hardly necessary to discuss here the details of the pending comprehensive test ban. Since 1963, when the Partial Test Ban Treaty (prohibiting nuclear explosions in the atmosphere, under water and in outer space) was concluded, a comprehensive test ban has meant in effect an underground test ban. The comprehensive test ban has been the subject of more discussion and negotiation than any other topic in the history of arms control and disarmament. In the text of the Partial Test Ban Treaty and again in the Nonproliferation Treaty the three signatory nuclear-weapons states promised to seek "the discontinuance of all test explosions of nuclear weapons for all time." The non-nuclear-weapons states regard the achievement of a comprehensive test ban as a crucial test of the seriousness of the intentions of the nuclear-weapons states to halt the nuclear-arms race; it was their chief demand at the first Nonproliferation Treaty Review Conference, held in Geneva in 1975. Indeed, if the nuclear-weapons states had succeeded in achieving a comprehensive test ban in 1963, they might well have avoided the proliferation predicament in which they now find themselves.

The damage done then might still be reparable if a comprehensive test ban could be achieved soon. The three negotiating parties are said to be very close to agreement and had expected to bring their negotiations to a successful conclusion soon after the ratification of the SALT II Treaty. Failure to ratify the treaty could only have an adverse effect on the prospects for a comprehensive test ban and for nonproliferation.

On the face of it the achievement of a comprehensive test ban would appear to be a comparatively easy step, since most of the military benefits likely to be gained from a large number of nuclear-weapons tests have already been incorporated into the nuclear arsenals of the two superpowers. To be sure, some military officers and nuclear-weapons designers on both sides might not want to see the door closed on all further nuclear-weapons testing, since they presumably would like to continue refining and improving the large variety of sophisticated nuclear weapons now in existence and developing new ones for future generations of delivery vehicles. Any conceivable benefits from continued testing would be insignificant, however, compared with the serious proliferation risks that would be presented by such a policy.

An equally important step in the direction of curbing the arms race, it seems to me, would be to ban the production of fissionable material for nuclear weapons. (For most practical purposes this means natural uranium enriched to about 90 percent in the readily fissionable isotope uranium 235 and reactor-generated plutonium containing a fairly high concentration of the readily fissionable isotope plutonium 239.) This proposal has also been discussed for many years, although no formal negotiations on it have ever taken place. Originally the concept was put forward as a way to "cut off," or halt, the manufacture of weapons-grade fissionable material in special military-production facilities within the existing nuclear-weapons states. In recent years, however, as a number of other countries have acquired the capability of producing plutonium and enriched uranium in the course of their civil nuclear-power programs, it has become more accurate to speak of a broader prohibition on the production of weapons-grade material by both nuclear-weapons states and non-nuclear-weapons states. The term "cutoff" has nonetheless continued to be used in arms-control circles to cover both contingencies.

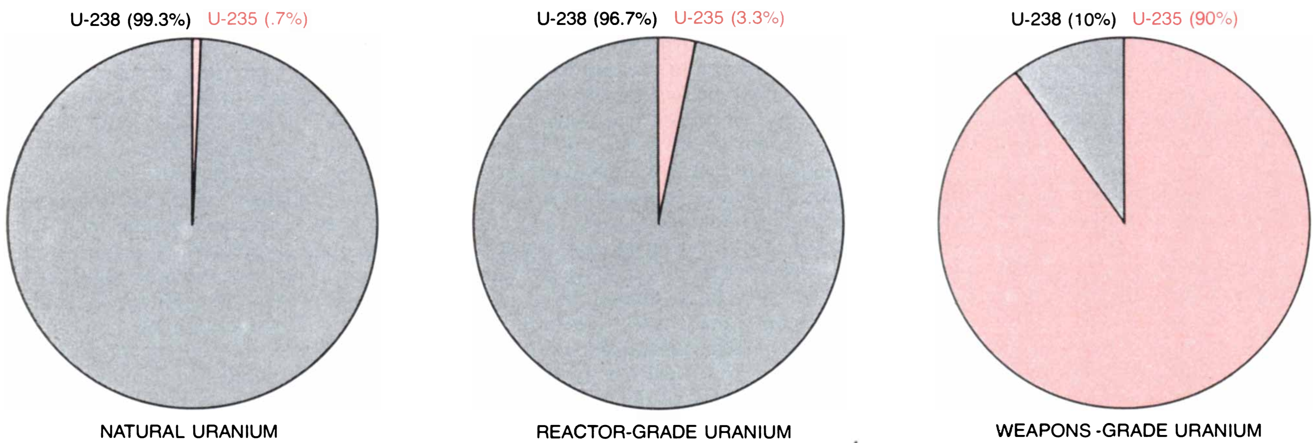
The proposed production ban would be a logical and feasible next step in the nuclear-arms-limitation process. Indeed, the problems involved in its negotiation would be fairly simple, particularly when they are compared with the special difficulties and complexities facing any future SALT negotiations. More important, a treaty incorporating such a ban would help to restrain both the vertical and the horizontal proliferation of nuclear weapons. The confluence of a number of circumstances has made the time ripe for this step. Accordingly it should be put at the top of the international arms-control agenda, on a par with the comprehensive test ban.

The idea of a cutoff in the production of fissionable material for weapons purposes was first put forward in President Eisenhower's "Atoms for Peace" address to the General Assembly of the United Nations in 1953. The President's proposal, which arose out of the failure of the postwar Baruch plan for the international control of atomic energy, incorporated not only the idea of a production cutoff but also a call for "contributions from the stockpiles of normal uranium and fissionable material" to a new international agency empowered to promote and regulate the use of atomic energy for peaceful purposes. (This part of the proposal led to the establishment of the International Atomic Energy Agency, or IAEA, in 1957.)

The officials responsible for the American "Atoms for Peace" proposal were attempting a new approach to nuclear-arms control and disarmament. They advanced the idea of the cutoff as a way of placing a ceiling on the amount of fissionable material available for nuclear weapons and hence on the total number of nuclear weapons that could be made, and they viewed the transfer of such material from military stockpiles to the new international agency as a way of beginning to reduce the total number of nuclear weapons already in existence. They regarded the two parts of the proposal as interrelated measures that could proceed in parallel with other efforts to prevent the further proliferation of nuclear weapons both vertically and horizontally.

In 1956 Secretary of State Dulles, acting on behalf of the U.S., the U.K., Canada and France, presented a working paper titled "Proposals for Partial Measures of Disarmament" to the five-member subcommittee of the UN Disarmament Commission, which was then meeting in London. Among the provisions included in the working paper was the proposal that "all future production of fissionable material will be used under international supervision, exclusively for non-weapons purposes," and that "the parties undertake to provide, under international supervision, for equitable transfers, in successive increments, of fissionable material from previous production to non-weapons purposes." The U.S.S.R. rejected the working paper in its entirety on the grounds that such a ban on the production of fissionable material for nuclear weapons without the outlawing of nuclear weapons themselves would be impractical.

The following year the UN General Assembly adopted, over the opposition of the U.S.S.R., a resolution based on a U.S.-sponsored draft that urged the

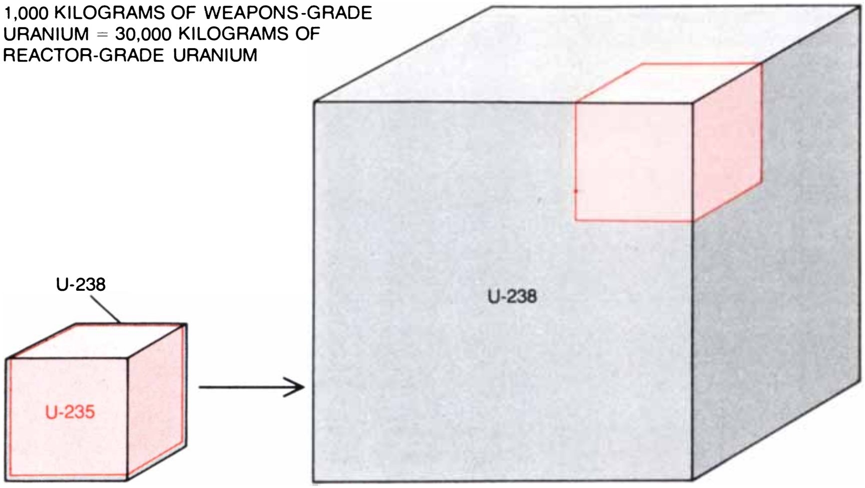


URANIUM in its natural state (*left*) consists almost entirely of the isotope uranium 238, with a small admixture (.7 percent) of the readily fissionable isotope uranium 235. Low-enriched uranium suitable for use as a fuel in a typical nuclear-power reactor (*center*) contains approximately 3.3 percent uranium 235. High-enriched weapons-grade uranium (*right*) contains more than 90 percent uranium 235.

member states to give priority to a disarmament agreement providing for a number of partial disarmament measures, including "the cessation of the production of fissionable material for weapons purposes." That was the first time the General Assembly had adopted any resolution dealing specifically with the cutoff.

The issue was raised again in 1964, when President Johnson, in a message to the UN Committee on Disarmament, proposed a "freeze" on the nuclear-arms race, accompanied by a separate agreement for a cutoff that could begin with verified plant-by-plant shutdowns. Later that year the U.S., the U.S.S.R. and the U.K. separately announced their unilateral decision for an immediate reduction (or "cutback") in the production of fissionable material for weapons purposes. The U.S. explained that its unilateral curtailment of the production of fissionable material, to be carried out over a period of four years, would (when added to previous reductions) represent overall decreases of 20 percent in the production of plutonium and 40 percent in the output of enriched uranium. The U.S.S.R. in turn announced its decision to stop forthwith the construction of two large reactors for the production of plutonium, to reduce substantially over the next few years the production of uranium 235 for nuclear weapons and to allocate more fissionable material for peaceful uses. The U.K. responded by pointing out that it had pursued a policy along the same lines; it had earlier announced its production of uranium 235 had ceased and its output of plutonium was ending gradually. It is noteworthy that none of the three parties ever complained that any of the others had failed to abide by its unilateral promises.

The U.S. subsequently submitted a detailed working paper in support of a complete cutoff, outlining nonintrusive verification procedures and proposing that it convert 60,000 kilograms of fissionable material (and the U.S.S.R. 40,000 kilograms) to peaceful purposes. The U.S.S.R. again expressed doubt about whether a separate agreement on a cutoff was possible outside of a general disarmament agreement.



CONVERSION of 1,000 kilograms of stockpiled weapons-grade uranium to peaceful purposes would yield some 30,000 kilograms of reactor-grade uranium, enough to fuel a 1,000-megawatt power reactor of the light-water type for approximately a year. The exact amount of weapons-grade fissionable material currently set aside in the military stockpiles of the nuclear-weapons states is of course not a matter of public knowledge. Some idea of the potential supply of nuclear fuel from this source, however, can be gathered from the fact that in 1964 the U.S. offered to transfer 60,000 kilograms of fissionable material from its military stockpiles to peaceful uses, if the U.S.S.R. would similarly convert 40,000 kilograms. The U.S.S.R., which at that time had far less weapons-grade material on hand than the U.S., turned down the offer.

It is not surprising that the U.S.S.R. opposed the proposal for a cutoff in the 1950's and 1960's. All nuclear powers restrict information about the size of their stockpiles of fissionable plutonium and enriched uranium, and they continue to guard this information as a military secret. The possession of such information would give an adversary an accurate basis for estimating the size of a nation's nuclear arsenal. It was generally recognized at the time, however, that the U.S., with its long head start and greater technical competence, had stockpiled far more fissionable material than the U.S.S.R. Hence any cutoff in the 1950's or 1960's, even if it had been accompanied by larger transfers of U.S. material from military stockpiles to peaceful purposes, would surely have worked to the disadvantage of the U.S.S.R.

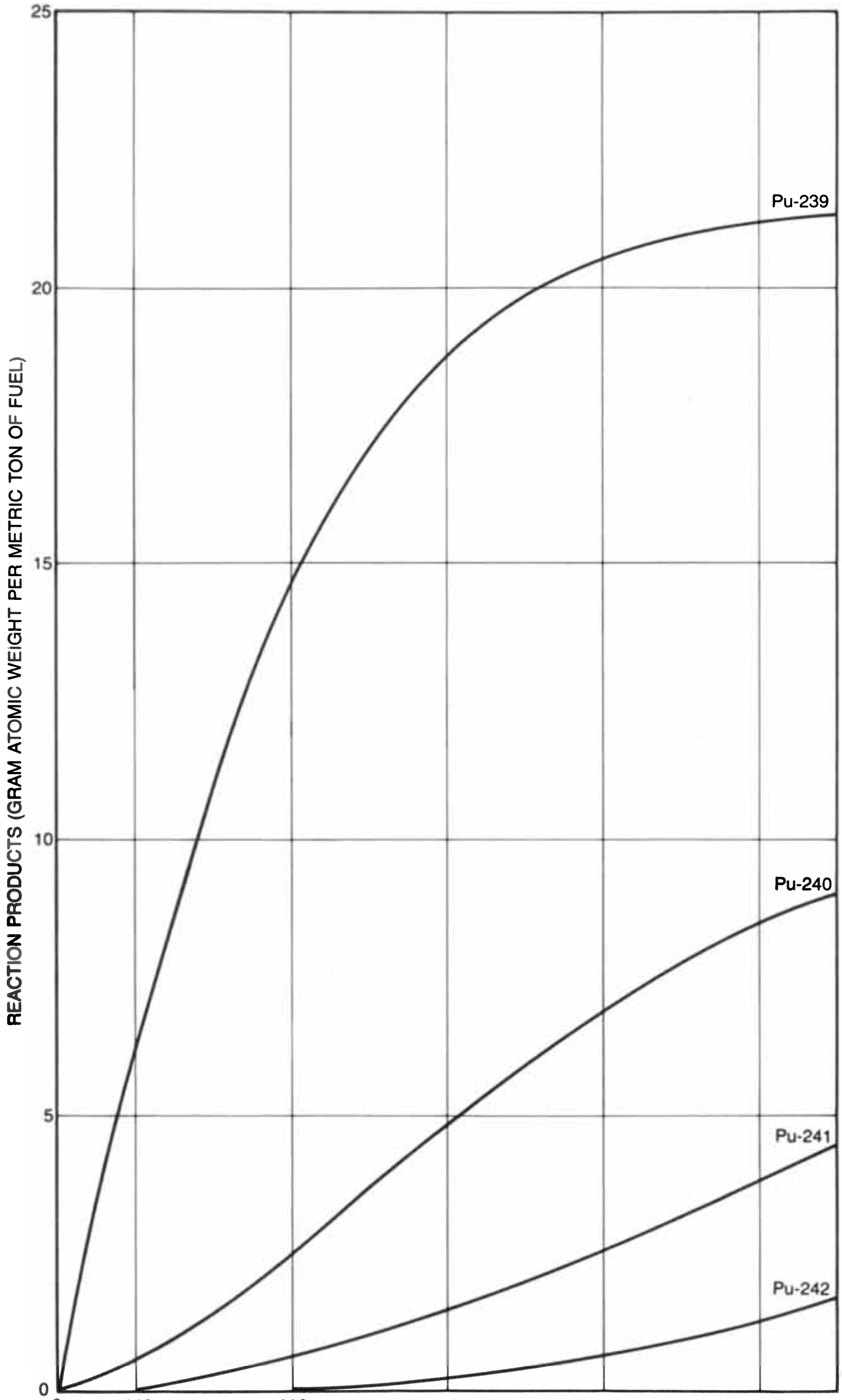
The U.S. persisted, however, and in 1966 presented three working papers outlining in further detail proposals for the transfer of fissionable material obtained from the destruction of nuclear weapons and for the provision of inspection systems to monitor shut-down plutonium-production facilities and shut-down nuclear reactors. The matter was not pursued for the next two years, however, because both the U.S. and the U.S.S.R. had in the meantime become preoccupied with negotiating an agreement on the Nonproliferation Treaty.

In 1969 the U.S. again stressed the importance of the cutoff. President Nixon listed it, along with a Comprehensive Test Ban Treaty, as one of the items the U.S. would press for in the Geneva meetings of the Committee on Disarmament. In order to meet the objections of the U.S.S.R. to the verification provisions proposed earlier, the U.S. announced that it would no longer ask for "adversary inspection" under a cutoff agreement but would rely on the IAEA safeguards system. The Nixon Administration had evidently reached the conclusion that the IAEA safeguards system, even if it was applied only to "declared" facilities or to shut-down ones, would provide adequate verification of Russian compliance. The U.S. also emphasized that in the context of a cutoff agreement the nuclear-weapons states could accept the same IAEA safeguards as were required for the non-nuclear-weapons states under the Nonproliferation Treaty. Finally, the U.S. maintained, the Russian argument that a cutoff would not be worthwhile because it did not deal with existing nuclear weapons was invalid; it was tantamount to saying no steps toward halting the arms race were worthwhile if they did not completely eliminate existing nuclear arsenals, a thesis that was unacceptable. The U.S.S.R. remained adamant in spite of the fact the new American approach received wide support among the non-aligned countries.

After the beginning of the SALT negotiations in 1969 the issue of multilateral negotiations for a cutoff (and indeed for most other nuclear-arms-limitation measures) was treated as if it had lost its urgency, and it lapsed into abeyance. It was not until the UN Special Session on Disarmament almost a decade later that the idea of a cutoff was revived.

The UN Special Session on Disarmament, held in New York in May and June of 1978, was the largest conference of its kind ever convened. It was attended by 20 heads of government, four vice-presidents or deputy prime ministers and 49 foreign ministers. After six weeks of debate and intensive negotiations the session reached a consensus on a lengthy final document containing a declaration of principles, a program of action for disarmament and an improved procedural mechanism for negotiations.

Several participating countries supported the cutoff as a promising mea-



PLUTONIUM 239, another readily fissionable isotope widely used in the fabrication of nuclear weapons, is created in a typical uranium-fueled reactor as the end product of a particular sequence of nuclear reactions in which a nucleus of uranium 238 in the initial fuel captures a neutron, becoming uranium 239, which then decays radioactively in two steps to yield plutonium 239. A nucleus of plutonium 239 may in turn fission when it is struck by a neutron, or it may absorb the neutron to become plutonium 240. The leveling off of the production curve for plutonium 239 means that at the end of the effective life of the fuel load in a typical pressurized-water reactor this isotope is being consumed almost as fast as it is being created. Plutonium 241, the product of neutron capture by plutonium 240, may also fission, or it may capture another neutron to become plutonium 242. Weapons-grade plutonium, loosely defined as plutonium containing a high concentration of plutonium 239, is obtained from special military-production reactors simply by removing the fuel elements from the reactor core after a comparatively short "burnup" time. Under certain circumstances, however, reactor-grade plutonium (obtained after a burnup time of three years, say) could also be used to make nuclear explosives; special provisions would therefore have to be incorporated in any international agreement prohibiting the production of this material for the purpose of making nuclear weapons.

sure of nuclear-arms limitation. One of the strongest proponents was Prime Minister Trudeau of Canada, who outlined a four-point strategy to "suffocate" the nuclear-arms race, a strategy consisting of a comprehensive nuclear-test ban, a ban on the flight testing of all new strategic delivery systems, a ban on the production of fissionable material for weapons purposes and a reduction in military expenditures on new strategic-weapons systems. Concerning the cutoff he said: "The effect of this would be to set a finite limit on the availability of nuclear-weapons material.... It would have the great advantage of placing nuclear-weapons states on a much more comparable basis with non-nuclear-weapons states than they have been thus far under the dispensation of the Non-Proliferation Treaty." As eventually agreed to by the special session, the final document called for a "cessation of the production of all types of nuclear weapons and their means of delivery, and of the production of fissionable material for weapons purposes."

At the regular session of the General Assembly in the fall of 1978 Canada followed up the initiative of its prime minister by proposing a resolution on the cutoff as an important measure both for limiting the nuclear-arms race and for preventing the further proliferation of nuclear weapons. The resolution asked that the Committee on Disarmament, at an appropriate stage of its work, consider urgently "an adequately verified cessation and prohibition of the production of fissionable material for weapons purposes and other nuclear explosive devices." The language of the proposed cutoff thus explicitly covered both the cessation of current production and the prohibition of new production; in addition it was deliberately extended to ban all "peaceful" nuclear explosives. The U.S.S.R. and the other members of the Warsaw Pact bloc (except Romania) opposed the resolution as not going far enough, since it did not also call for a halt in the production of nuclear weapons. The U.S. and the U.K., although they were no longer such strong advocates of the cutoff as they had once been, voted for the resolution, and it was adopted by an overwhelming majority of the nations present and voting. At the 1979 session of the General Assembly, Canada again proposed a resolution calling on the Committee on Disarmament to pursue its consideration of the cutoff, and the resolution was adopted by a vote even larger than the one in 1978. These large affirmative votes have encouraged Canada to continue to press for the cutoff, which is once again clearly an outstanding item on the international agenda of disarmament issues.

It is not clear why the U.S. should now demonstrate less enthusiasm for an arms-limitation measure it had initiated and ardently supported for more than two decades. It is of course possible that for the thousands of new warheads it will be allowed to manufacture under the SALT II Treaty even the massive stockpile of weapons-grade fissionable material the U.S. has built up over the years may be thought by some to be inadequate. When one remembers, however, that the U.S.'s unilateral cutback in the production of fissionable material in 1964 amounted to a decrease in production of 20 percent for plutonium and 40 percent for highly enriched uranium, and that in addition the U.S. offered to convert 60,000 kilograms of such material (enough to make thousands of nuclear warheads) from military uses to peaceful ones, it is difficult to believe the U.S. might now fear a shortage of nuclear explosive material. It seems reasonable to look elsewhere for the real motivation.

Of course, any proposal as far-reaching as a ban on the production of fissionable material for nuclear weapons can be expected to stir up opposition based on perceived or imagined strategic inequities. The old specter of the U.S.S.R.'s gaining some advantage would no doubt be resurrected. Actually it is the Russians who might still be concerned about the possibility that the U.S. could gain some advantage as the result of a cutoff. At the time the SALT II Treaty was signed the U.S. had close to 10,000 strategic warheads and bombs, whereas the U.S.S.R. was reported to have about 5,000. Under the terms of the treaty each of the two powers would be legally entitled to build up to a maximum of some 17,000 strategic warheads and bombs by 1985, although neither apparently intends to build that many. Current plans are for the American arsenal to increase to about 13,000 warheads and for the Russian arsenal to increase to about 10,000 warheads. (In addition the U.S. is reported to have some 22,000 tactical nuclear weapons and the U.S.S.R. a somewhat smaller number.)

In neither case, however, would there be any real cause for concern that one side would gain any advantage over the other as a result of a cutoff. The SALT II Treaty has fixed equal overall ceilings for strategic launchers and warheads. Any cutoff agreement, if it is to be acceptable, would also have to provide equal ceilings for both powers. If there is some disparity between the present stockpiles of weapons-grade fissionable material, it could be adjusted by asymmetrical conversion and transfers from military purposes to peaceful ones. Therefore, whereas a cutoff coupled with a transfer of military material to peaceful purposes in the period of the 1950's and 1960's might not have helped to stabilize the nuclear balance between the two superpowers, in today's circumstances it would help. In addition if the initial production ban and transfers proceeded according to agreement, future transfers could be progressively increased and could provide another way of achieving actual reductions in strategic nuclear weapons. Thus the cutoff could become a significant disarmament measure in itself.

Questions would no doubt be raised about the adequacy of verification of such a production ban. Because of the remarkable progress made in the past two decades in satellite surveillance and other "national technical means," verification no longer presents the same problems it once did, in terms of either effectiveness or intrusiveness. That was undoubtedly a factor in the decision by President Nixon to abandon adversary inspection in 1969 and to rely on the less onerous international safeguards system of the IAEA for declared plants and facilities that were to be shut down. Modern means of verification would ensure that the large plants and facilities needed to produce significant quantities of highly enriched uranium and plutonium could not escape detection. Even the secret diversion of spent fuel from power reactors to the production of plutonium, or the development of new processes for enriching uranium, which might allow the secret production of weapons-grade fissionable material in small clandestine installations, would not create any serious instability in the nuclear balance between the superpowers. The amounts that could be produced secretly in this way would be so small as to make no real difference, considering the magnitude of the existing American and Russian stockpiles.

Special attention would have to be given to the problems of verification that would arise in connection with the production of plutonium. Although pure plutonium 239 is the most efficient weapons-grade fissionable material, it can be exploded even when it is in a mixture with smaller amounts of other fissionable and nonfissionable plutonium isotopes. Furthermore, it is extremely difficult to deliberately "denature" plutonium 239. Even when it is associated with oxides and isotopes of other elements, the plutonium can be extracted from spent reactor fuel by various chemical-separation techniques. The verification problem would be complicated further, of course, if the world were to move in the direction of a "plutonium economy" through the widespread adoption of plutonium recycling and breeder reactors.

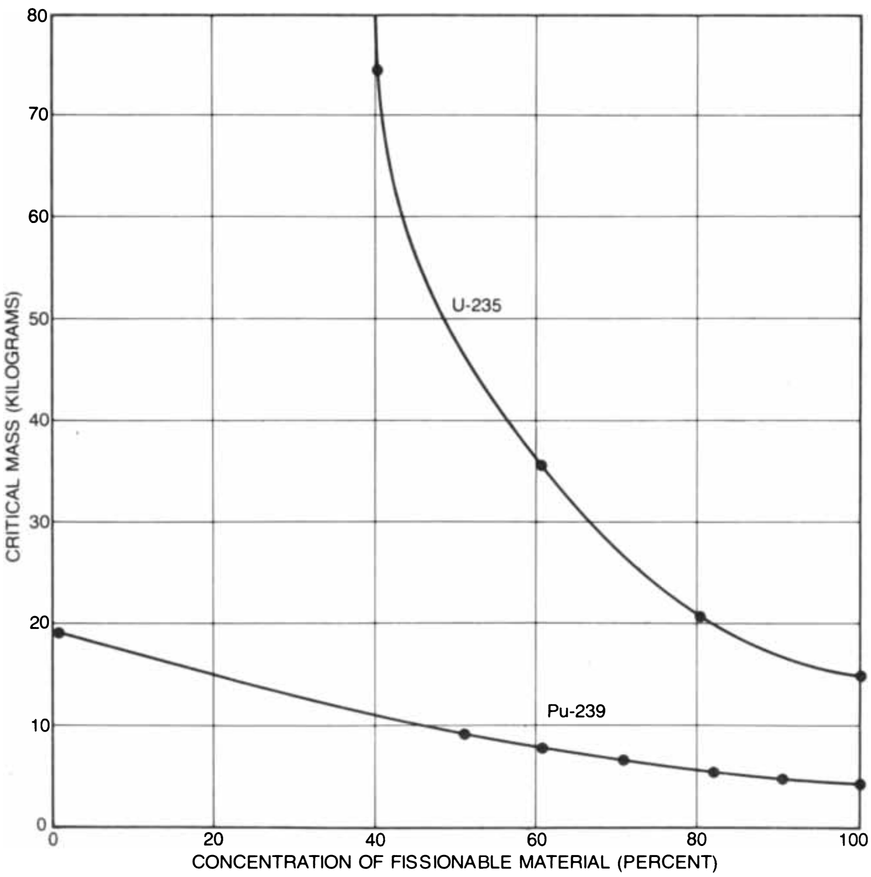
Two different approaches to dealing with this problem have been put forward. A number of observers have suggested that all reprocessing of plutonium be stopped, a solution that in itself creates many problems. The other approach that has been proposed is to place all plutonium under IAEA safeguards and all plutonium stockpiles in IAEA custody. This would require that the IAEA safeguards, which are now based on accounting, containment and surveillance systems and periodic on-site inspections, be strengthened and that special stockpiling facilities be established. Whichever approach is adopted, the details of any system that called for the verification of a ban on the production of weapons-grade plutonium, or for its control in special cases (such as its use as a fuel in certain nuclear reactors or in nuclear-powered submarines and other ships), would have to be carefully negotiated, and special rules and precautions would be required. The same would be true of highly enriched uranium produced for special nonexplosive purposes. In any case full use could also be made of national technical means of verification and of consultative commissions to deal with ambiguous events, following the precedent set in the various SALT agreements.

Another problem would arise if, as seems likely, China and France would not agree to a cutoff. Both countries have repeatedly insisted they would not agree to a comprehensive test ban until they closed or at least reduced the nuclear-weapons gap between them and the superpowers. They would no doubt take the same position with respect to a cutoff. China and France are so far behind in nuclear weaponry, however, that for a number of years it would make little difference from the point of view of the superpowers whether these two secondary nuclear powers agreed or not. The U.S., the U.S.S.R. and the U.K. were willing to conclude the 1963 Partial Test Ban Treaty and the Nonproliferation Treaty (and to negotiate a comprehensive test ban) without the participation of China and France. They should be equally willing and able, without jeopardizing their security or undermining nuclear deterrence, to agree to a cutoff without China and France, in the expectation that eventually these two holdouts would also agree. Indeed, any cutoff agreement would almost certainly include what has now become a customary provision of arms-control treaties, namely calling for a review of the operation of the treaty after a stated period of years to evaluate how expectations have been fulfilled.

It might also be argued that a cutoff and the beginning of nuclear disarmament would result in a massive buildup of conventional arms, which could lead to a large increase in military expenditures. In the first place, it must be reiterated that, dangerous and destructive though conventional weapons are, they do not present anything like the threat to civilization and human survival that nuclear weapons do; it would take years for any conventional war to cause the destruction that could be inflicted in hours or even minutes by a nuclear war. Moreover, this argument tends to ignore the vastly different world climate that would be produced by any sign of real restraint in the nuclear-arms race. The resulting reduction of fears and tensions, the revival of détente between the U.S. and the U.S.S.R. and increased international confidence could lead to greater international cooperation and might create the conditions where far-reaching measures of conventional disarmament as well as of nuclear disarmament would become possible. There is of course no guarantee this would be the result; history shows, however, that the continuing development of nuclear weapons does not lead to conventional disarmament but rather feeds the arms race in both directions.

Finally, it could be argued that a cutoff would not end the nuclear-arms race because the U.S. and the U.S.S.R. could still conduct a technological race to increase the accuracy and effectiveness of their delivery systems, presenting an even greater threat to deterrence and international security than the mere production and accumulation of more weapons. The cutoff by itself would not halt research and development, or the modernization of nuclear systems or the transfer of fissionable material from older to newer types of nuclear warheads. Therefore unless it were followed by more direct measures for nuclear disarmament and by curbs on new technological developments (such as a ban on underground nuclear tests) the impact of a production ban would in time begin to lose its effect.

These arguments have some merit. They could of course be equally applied to a ban on the production of additional nuclear weapons. Nevertheless, there are a variety of indirect benefits to be had from limiting the amount of nuclear-explosive material and reversing the open-ended race to increase the number of nuclear weapons. As the U.S. argued in 1969, it is an unacceptable thesis to maintain that intermediate steps toward halting the arms race are not useful if they do not completely eliminate all nuclear weapons. In order to reverse any race it is first necessary to bring it to a halt. Furthermore, the very act of ending the production of explosive material necessary to produce additional weapons could create a more favorable climate for ending the production of additional nuclear weapons.



CRITICAL MASS of fissionable material needed to produce a nuclear explosive is the smallest amount in which a chain reaction can be sustained by fast neutrons alone. The critical mass depends on a number of factors, including the physical purity of the material in question. For a given material the critical mass is smallest when the fissionable material is in its pure isotopic form, and it is surrounded by a suitable neutron-reflecting material, such as natural uranium. (A critical mass of 4.4 kilograms of 100 percent plutonium 239, plotted here under those conditions, would be about the size of an orange.) The variation of critical mass with concentration of fissionable material in it is obviously less pronounced for plutonium than it is for uranium.

The successful verification of a cutoff would have great psychological and political importance. It might profoundly affect conventional beliefs that nothing can be done to halt the arms race. It would have some effect in de-emphasizing the importance of nuclear weapons in international relations and could lead to a significant reduction in tensions between the U.S. and the U.S.S.R. It could also provide a demonstration in international verification that would be useful for other measures of nuclear-arms control and disarmament.

At present 114 countries are parties to the Nonproliferation Treaty, leaving more than 40 that are not. Most of the latter have no capability or desire to "go nuclear," but a number of them already have or will soon have such capability; they include not only India, which has already exploded a nuclear device, but also Pakistan, Israel, Egypt, South Africa, Spain, Argentina and Brazil. Among the countries that are parties to the Nonproliferation Treaty some three dozen are potential nuclear-weapons powers. Under the terms of the treaty they can withdraw on three months' notice if they feel their vital interests are in jeopardy. Whether any of them decide to withdraw and go nuclear will depend in part on whether any other non-nuclear-weapons states (and which ones) decide to acquire nuclear weapons. It is therefore of crucial importance for all states interested in strengthening the nonproliferation regime to do everything possible to reduce the incentives for additional states to go nuclear.

There are a number of things the nuclear-weapons states can do in this regard, one of which is the cessation of their vertical proliferation of nuclear weapons, which will take some time to achieve. A more immediate step would be the elimination, or reduction to the extent possible, of the discriminatory elements of the Nonproliferation Treaty. When this agreement was being negotiated in the 1960's, many of the non-nuclear-weapons states were uneasy about the inherent imbalance of the treaty, which decreed a permanent division of the states of the world into two categories: those with nuclear weapons and those without them. Most countries were willing to accept the situation as a temporary fact of international life, provided that the nuclear-weapons states committed themselves to halting and reversing the nuclear-arms race (according to Article VI of the treaty) and assisting the non-nuclear-weapons states in the development of the peaceful uses of nuclear energy (according to Article IV). The problem of the lack of equal rights and obligations between nuclear-weapons states and non-nuclear-weapons states was made more difficult by the provisions of the treaty that required only non-nuclear-weapons states to subject their entire nuclear programs to the safeguards system of the IAEA. So great was the resentment of the non-nuclear-weapons states over what many of them regarded as an inequitable and discriminatory situation that the U.S. and the U.K. (but not the U.S.S.R.) agreed voluntarily to accept IAEA safeguards over their nonmilitary nuclear programs and facilities. Although this concession went some way toward achieving greater balance between the two categories of states, it did not go far enough to satisfy such countries as India, Brazil and Argentina, which continue to regard the treaty as discriminatory.

Even the non-nuclear-weapons states that are parties to the Nonproliferation Treaty are dissatisfied with what they consider the failure of the nuclear-weapons states to carry out their commitments under the terms of the treaty. At the first Nonproliferation Treaty Review Conference in 1975 the nonaligned non-nuclear-weapons states (known as the Group of 77) were insistent in their demands that the nuclear-weapons states live up to their commitments and that in addition they provide security assurances to the non-nuclear-weapons states by pledging not to use or threaten to use nuclear weapons against them. The compromise declaration that was finally adopted by consensus left most of the non-nuclear-weapons states dissatisfied.

In the five years that have passed since the first Nonproliferation Treaty Review Conference little progress has been made in satisfying the demands of the non-nuclear-weapons states. Although the member states of the Nuclear Suppliers Group, which met in London after 1975, did succeed, on their own initiative and in their own interests, in agreeing to tighten the restrictions and safeguards on the export of nuclear material and equipment, the non-nuclear-weapons states tended to regard this as a one-sided effort that would intensify the discriminatory aspects of the nonproliferation regime. The International Nuclear Fuel Cycle Evaluation (INFCE) has just completed a detailed two-year inquiry, with the participation of both the importing countries and the suppliers, into the many problems of nuclear energy and their relation to nuclear-weapons proliferation. The INFCE report, which was released in February, contains a great deal of valuable information and analysis. It was strictly a technical study, however, not a political one. Accordingly it has not yielded any solutions to the main problems of nuclear energy and nuclear-weapons proliferation, which are largely political in nature. Only a combination of both technical and political measures holds out any promise of success in halting the further proliferation of nuclear weapons.

Many of the political measures required are related to halting and reversing the vertical proliferation of nuclear weapons as a way of facilitating efforts to prevent their horizontal proliferation. In this context the most important immediate steps that could be taken would be the comprehensive test ban and the cutoff. The impact of a comprehensive test ban is obvious: it would place the nuclear-weapons states and the non-nuclear-weapons states on a more equal footing if both were barred from conducting any nuclear explosions of either a military or a peaceful nature; it would also constitute an effective curb on the technological improvement of nuclear warheads. Finally, it would have an immediate symbolic and psychological effect by demonstrating the willingness of the nuclear-weapons states to accept some limitation on the qualitative nuclear-arms race.

The impact of the cutoff, which would be similar to that of the comprehensive test ban, requires more explanation. It would, of course, still be possible for the nuclear-weapons states to produce additional nuclear warheads from their stockpiles of fissionable material or to transfer explosive material from larger to smaller warheads or from older to newer ones. The cutoff would still, however, constitute an important step toward curbing the vertical proliferation of nuclear weapons; it would slow the production of nuclear weapons and could be a very useful step toward actually ending the production of all nuclear weapons. It would therefore help to reduce the discriminatory features of the Nonproliferation Treaty.

Moreover, once the production of fissionable material for weapons purposes was halted, future production of fissionable material would have to be for peaceful civilian purposes. As the U.S. stressed in 1969, this would make it possible for the nuclear-weapons states to accept the same IAEA safeguards that were required of the non-nuclear-weapons states under the Nonproliferation Treaty, since there would be few military secrets to protect. Thus at a stroke one of the most irritating elements of discrimination between the nuclear-weapons states and the non-nuclear-weapons states could be removed. Special precautions would have to be taken so that certain secrets of uranium enrichment would not be disclosed, but these need not be much different from the precautions now taken under IAEA safeguards to prevent the disclosure of technological and commercial secrets in the case of peaceful nuclear reactors and facilities. If additional verification procedures were required to ensure that uranium-enrichment and plutonium-reprocessing plants in the nuclear-weapons states were not being used to produce fissionable material for nuclear weapons, these procedures would be a necessary concomitant of the greater nuclear capabilities of the nuclear-weapons states and would not detract from the essential equity of the treaty, as long as the safeguards on the nuclear-weapons states were no less thorough or effective than those on the non-nuclear-weapons states.

By taking a concrete step toward nuclear-arms limitation and reduction, by holding out hope for further measures of nuclear disarmament and by removing the unequal treatment regarding international safeguards, the proposed production ban could help to strengthen the Nonproliferation Treaty and remove or at least weaken some of the arguments used against it. If a cutoff agreement were negotiated as a multilateral treaty, it would constitute a direct legal barrier to the spread of nuclear weapons to additional countries that accepted the treaty. In order for such a treaty to apply to non-nuclear-weapons states that were producing fissionable material, it would have to provide for the prohibition as well as the cessation of the production of fissionable material for weapons purposes and for other nuclear explosive devices. A prohibition of this kind would go considerably beyond the provisions of the Nonproliferation Treaty and the pending comprehensive test ban in imposing a prior legal barrier to the spread of nuclear weapons and explosive devices to countries that do not yet have them. If the treaty were to provide for safeguards to be applied equitably to both nuclear-weapons states and non-nuclear-weapons states, and also for the conversion and transfer of existing stocks of weapons-grade fissionable material to peaceful civilian uses, in particular for the benefit of the non-nuclear-weapons states, then these provisions would serve as incentives for non-nuclear-weapons states to become parties to it.

There are several different options that could be pursued in negotiating a ban on the production of fissionable material for nuclear weapons. Ideally it would be better if it could be negotiated from the beginning as a multilateral treaty covering both nuclear-weapons states and non-nuclear-weapons states. From the practical point of view, however, it might be necessary to begin on a bilateral basis between the two superpowers or on a trilateral basis, including the U.K. Any preliminary agreement reached in this way could be submitted to the Geneva Committee on Disarmament so that the non-nuclear-weapons states would have an opportunity for a full discussion and review of the treaty; the agreement could then be submitted to the UN General Assembly. Hence all states would be able to participate in the elaboration of the treaty, a procedure that would facilitate the widest possible adherence to it.

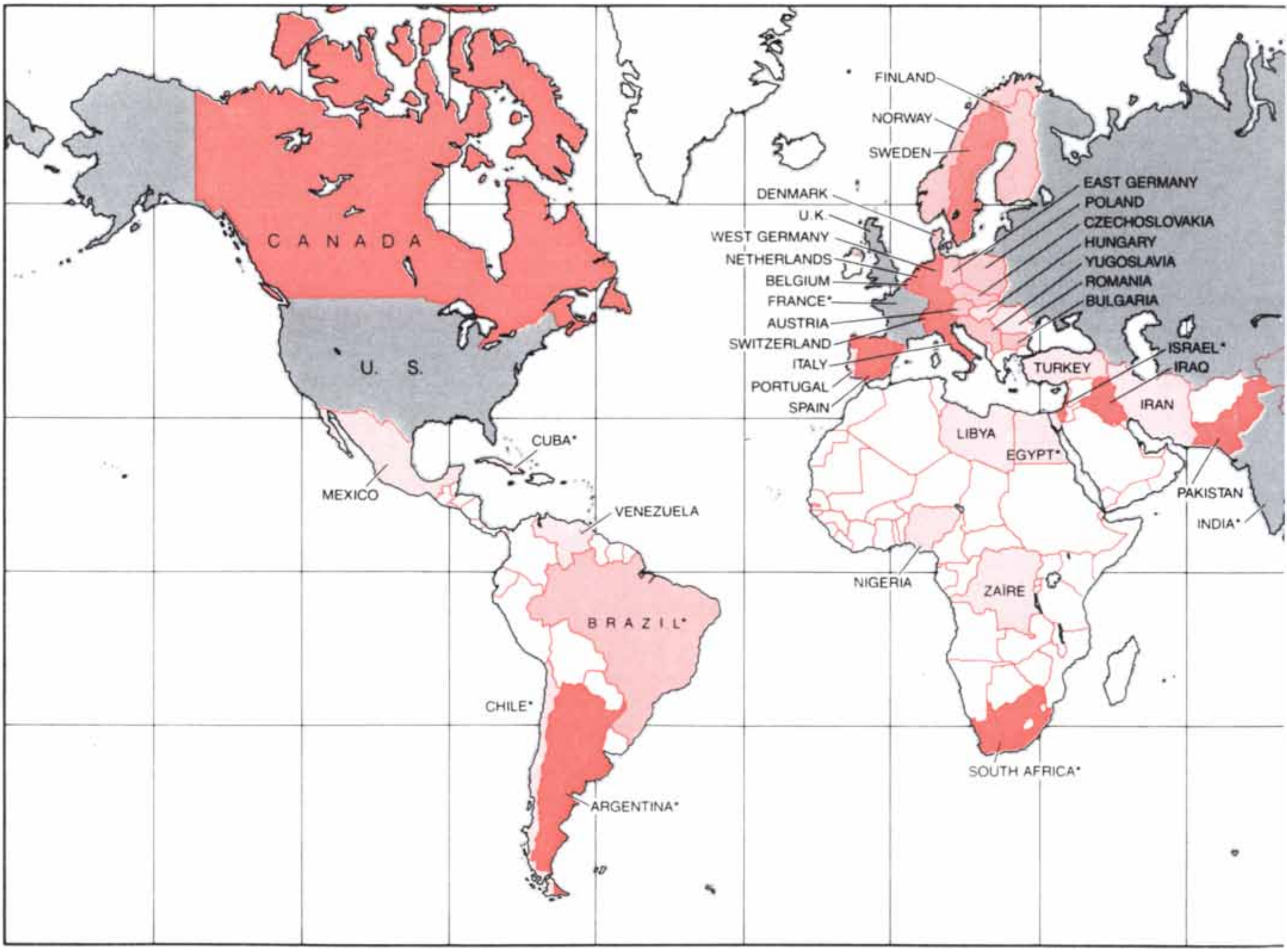
The verification procedures applied

to the nuclear-weapons states could provide in the first instance for complete plant-by-plant shutdowns on a reciprocal or matching basis, as the U.S. proposed in the 1960's, with available verification techniques, including the various national technical means of verification. The more complex procedures required for ensuring the conversion of weapons-grade fissionable material to peaceful uses could be verified by the IAEA safeguards system, which now monitors all the nonmilitary nuclear facilities of the non-nuclear-weapons states under the terms of the Nonproliferation Treaty. Special, more stringent procedures would have to be negotiated for the nuclear-weapons states in cooperation with the IAEA with respect to the various problem areas discussed above. Procedures would also have to be worked out for the conversion of military stocks to peaceful purposes.

Another option open to the U.S. would be to take the initiative on a cutoff unilaterally. Because of the painfully slow pace of negotiations for any international agreement and the complications of obtaining Senate approval, there is much to be said for a unilateral approach. There is also ample precedent for such national action intended to facilitate international reciprocal action or negotiations. The U.S. might well consider repeating its initiative of 1964, when it unilaterally announced a cutback of its production of fissionable material for weapons and invited the U.S.S.R. to reciprocate. This U.S. action led to the mutual cutbacks announced by the U.S.S.R. and the U.K., which came to be known as "reciprocated unilateral acts" or, as the Russians called the process, action by "mutual example."

In a similar action President Nixon announced in 1969 the unilateral decision of the U.S. to renounce all use of biological agents and weapons, to stop the production of all biological agents for weapons purposes, to dispose of all existing stocks of biological weapons and to confine all future biological research to defensive purposes. The U.S. hoped that other states too would renounce biological weapons, and it urged that such decisions be converted into international legal obligations through a convention of some kind. This unilateral action by the U.S. led to the negotiation and successful conclusion in 1972 of the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction.

A unilateral cutoff by the U.S. of the production of fissionable material for weapons purposes, at least in the form of a suspension, would not appear to



POTENTIAL NUCLEAR-WEAPONS STATES, shown in different shades of color on this world map, could each have the technical capability of exploding a nuclear device by 1990, if they should decide to do so. The darkest-colored countries could acquire this capability within a year or two, the medium-colored ones within two to six years and the lightest-colored ones within six to 10 years. Nations that have already tested nuclear explosives are shown in gray. Asterisks are used to designate nations that are not parties to the Nonproliferation Treaty. Although neither Israel nor South Africa has yet exploded a nuclear device, it is reported that they both now have the ca-

place it in any disadvantageous position compared with the U.S.S.R. if it were done on a temporary trial basis and the Russians were called on to take similar unilateral action and to begin negotiations for a cutoff treaty. The U.S. might therefore consider whether it could advance the cause of curbing both the vertical and the horizontal proliferation of nuclear weapons by initiating such unilateral action. If the initiative were carefully prepared in private through diplomatic channels, it might lead to a result as successful as the unilateral cutback in the production of fissionable material for weapons in 1964 and the unilateral cutoff of the production of biological agents for weapons in 1969. Such unilateral national action has the great merit of speed and timeliness in its initiation and implementation.

The two superpowers agreed last year in their Joint Statement of Principles and Basic Guidelines for Subsequent Negotiations on the Limitation of Strategic Arms that they would pursue "significant and substantial reductions in the numbers of strategic offensive arms [and] qualitative limitations on strategic offensive arms, including restrictions on the development, testing and deployment of new types of strategic offensive arms and on the modernization of existing strategic offensive arms." It is obviously in their mutual interest to do so, not only as a means of containing the nuclear-arms race and reducing the threat of nuclear war but also as a possible avenue for restraining the spread of nuclear weapons to other countries.

When the cutoff was first proposed by President Eisenhower in 1953, the situation in the world was vastly different from what it is now. The U.S.S.R. was far behind the U.S. in all areas of nuclear weaponry, and the danger of the proliferation of nuclear weapons was still rather remote. Today there is a rough parity between the two superpowers, and the threat of the further spread of nuclear weapons to additional countries, with all the incalculable problems that it would present to the superpowers and to the relative stability of the international situation, is a very real danger.

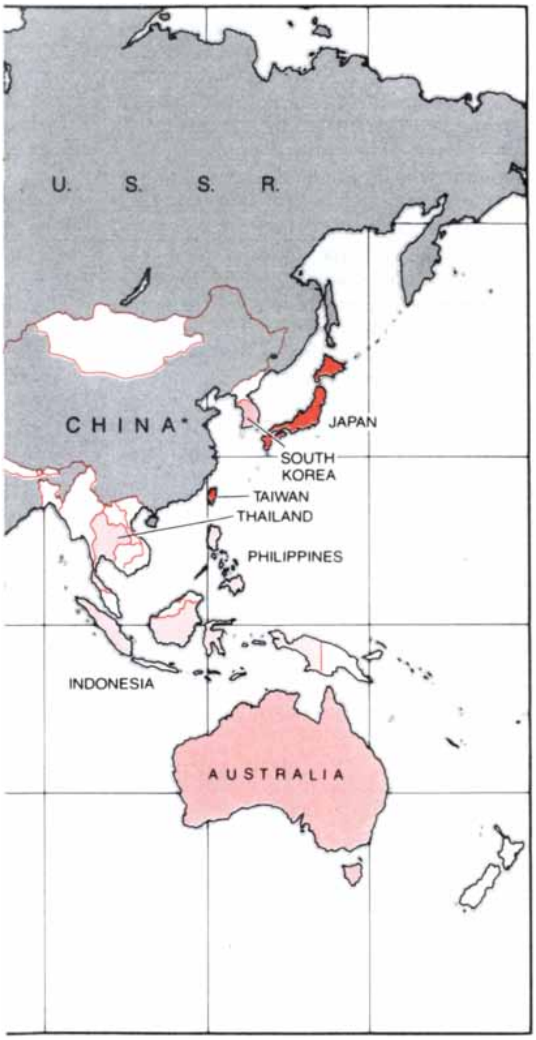
There is no assurance that if the superpowers were able to agree on a ban on the production of fissionable material for weapons or other explosive nuclear devices, that act alone would prevent the proliferation of nuclear weapons to additional countries. Nevertheless, a cutoff treaty could, particularly if it closely followed a comprehensive test ban, provide concrete evidence that the superpowers were determined to halt and reverse the nuclear-arms race. It might also create sufficient positive feedback to facilitate other measures that would halt both vertical and horizontal proliferation.

Apart from the major considerations outlined here there is an additional reason for proceeding with a cutoff. In spite of the opposition to nuclear power plants that has developed in recent years it seems fairly clear that the world will have to rely on nuclear power as an important and perhaps growing source of energy in the future. The fears of a few years ago of a shortage of uranium for nuclear reactors, which seem to have subsided for the moment, may very well revive. The high-enriched uranium stockpiled for weapons could provide a source of low-enriched uranium for power reactors. Every 1,000 kilograms of weapons-grade uranium could be converted into some 30,000 kilograms of nuclear fuel for reactors. This energy source would by itself help to postpone the advent of commercial breeder reactors and the plutonium economy. Furthermore, if all fissionable material produced in the future were to be used solely for peaceful purposes, there would be less fear that if the world does in fact move toward breeder reactors and plutonium recycling, it would lead to nuclear-weapons proliferation.

There is of course the ever present matter of the politics of the cutoff. The difficulties of achieving agreement on the SALT II Treaty and the looming threat of the Senate's failure to give its consent to ratification are so great that they create doubts for even the most optimistic proponents of arms control. The current mood of the American public and Congress verges on hostility toward arms-control measures. This antipathy, however, makes it all the more necessary to seek new ideas and openings that could keep the process going. Under the most pessimistic circumstances short of war it is almost inconceivable that the two superpowers would abandon their efforts to halt the nuclear-arms race and restrain the proliferation of nuclear weapons. Even during the "cold war" the U.S. took imaginative initiatives such as the Baruch plan in 1946 and the "Atoms for Peace" proposal in 1953. Any initiative in favor of the cutoff would be much simpler and far less radical than either of those initiatives.

The fact that the U.S.S.R. has in the past opposed a halt in the production of fissionable material for weapons purposes without a corresponding halt in the production of nuclear weapons should not be regarded as an insuperable obstacle. In the case of almost every multilateral arms-limitation agreement the U.S.S.R. at first took a radical position and called for total prohibitions. When it became apparent that the overwhelming majority of states were opposed to an all-or-nothing approach and considered half a loaf better than none, the U.S.S.R. has in almost all cases accepted a partial solution, often along the lines urged by the U.S. This same pattern has been repeated in many aspects of the SALT negotiations. Hence on the grounds of historical experience as well as of logic there is reason to hope that the U.S.S.R. would agree also to the cutoff as a first step toward further measures of nuclear-arms limitation.

The recent initiative taken by Canada has already received the support of nearly all the Western and nonaligned countries. If the leaders of the U.S. would once again apply their minds to the cutoff, the problem could be on the way to solution. Apart from the early achievement of a Comprehensive Test Ban Treaty, it is difficult to imagine a more promising new step than a treaty banning the production of fissionable material for nuclear weapons. Such an effort could also do much to restore the eroding credibility of the U.S. commitment to nuclear-arms control and nonproliferation.



pability of doing so. Libya is included as a potential nuclear-weapons state by virtue of its special ties to Pakistan. The map is an updated version of one published by the author in the April 1975 issue of Scientific American.

Systemic Lupus Erythematosus

A rheumatic disease with varied symptoms, SLE is a disorder of the immune system. A patient makes autoantibodies that bind to such antigens as DNA, forming immune complexes that can injure tissue

by David Koffler

One of the early manifestations of systemic lupus erythematosus (SLE) is the erosive inflammation of the skin, typically in the form of a "butterfly rash" over the nose and cheeks, that gave the disease its original name: lupus erythematosus, or "red wolf." That name does not begin, however, to reflect the true scope and the multisystem nature of this puzzling rheumatic disease that afflicts perhaps one U.S. woman in 500 (and about a tenth as many men). It is indeed a systemic disease, affecting not only the skin but also, in different patients and at different times, the joints, the blood vessels, the heart, the lungs and the brain—and most significantly the kidneys.

Whatever the symptoms may be in an individual case, the disease is essentially a disorder of the immune system. For some time SLE was considered an "autoimmune" disease, one in which the patient forms antibodies that interact with antigens in her own tissues. That may be the case for some manifestations of the disease. It now appears, however, that SLE is primarily an "immune complex" disease: the patient's antibodies combine with antigens (whether foreign antigens or the patient's own is still not known) to form complexes that actually mediate the tissue damage.

A primary function of the immune system is to distinguish self from nonself and thereby protect the body's internal environment against infection and other challenges from the external environment. In man the system consists of perhaps a trillion (10^{12}) of the cells called lymphocytes and some 10^{20} immunoglobulin molecules, or antibodies. There are two kinds of lymphocytes: *B* cells and *T* cells. Each clone, or line, of *B* cells is genetically programmed to make a particular antibody with combining sites that recognize a particular epitope: a region of an antigen molecule with which the antibody interacts.

In its resting state a *B* lymphocyte displays some of its antibody as receptors on its surface. Contact of a receptor with its specific antigen stimulates the *B* cell to proliferate and differentiate, forming a clone of antibody-producing plasma cells that manufacture millions of identical antibody molecules. There are five major classes of antibodies, two of which are of primary importance in SLE and other rheumatic diseases: the immunoglobulins (or gamma globulins) IgG and IgM. Antibody combines, either in the blood or in the tissues, with antigen, thereby forming immune complexes and setting in motion various processes that can destroy cells carrying the antigen or that can damage innocent bystanders: the organism's own tissues.

The important immune complexes appear to be those that fix complement: a nine-component group of serum proteins. A portion of the first component designated C1q interacts with the antibody part of a complex, initiating a series of reactions that bring on an inflammatory response. Factors formed by complement can bind to neutrophils (the commonest white blood cells), facilitating phagocytosis (ingestion of particles) and the release of enzymes that break down protein. Small molecules derived from complement attract other phagocytic cells. Other factors increase the permeability of blood vessels by releasing such substances as histamine from certain cells, thus increasing the influx of inflammatory cells and also enhancing the deposition of immune complexes.

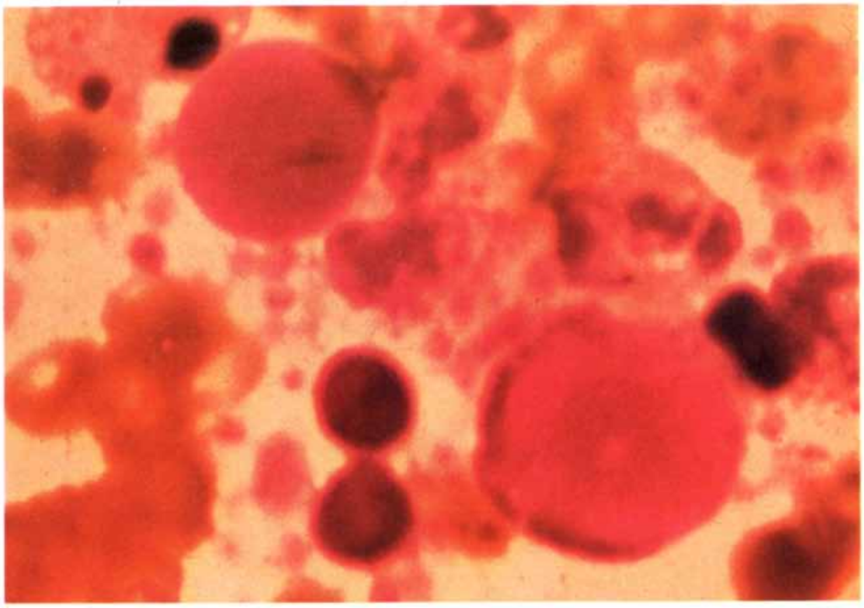
Whereas the *B* lymphocytes mediate humoral immune mechanisms involving soluble antibodies, the *T* lymphocytes mediate cellular immunity. They too have specific receptors on their surface that recognize epitopes on antigen molecules. The recognition of an antigen prompts the release from the *T* cells of lymphokines, chemical substances that have multiple effects: they injure cells, inhibit the replication of viruses and attract inflammatory cells. Other *T* lymphocytes influence the production of antibody by *B* cells. One type, the helper cells, facilitates the proliferation of *B* cells and another type, the suppressor cells, retards it. In SLE *T* cells seem to be important because their regulatory function is disturbed rather than because they damage tissue directly.

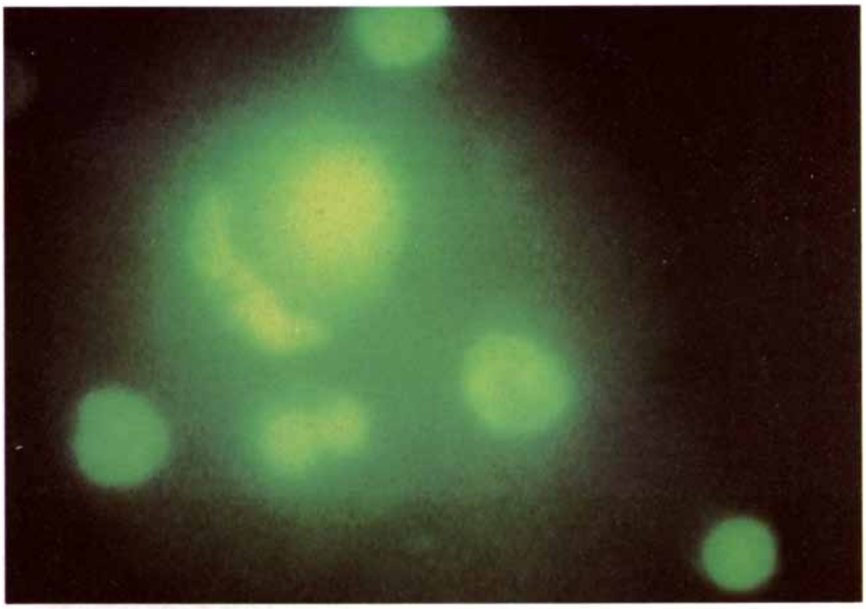
Although the immune system generally distinguishes self from nonself, the distinction is by no means absolute. As far back as 1902 Paul Ehrlich called attention to the potential for self-injury through the immune system. Ehrlich assumed that a *horror autotoxicus* must protect against this possibility: that "the organism has contrivances" preventing the immune response "from reacting against the organism's own elements." That is not the case, however; there are many situations in which just such a response takes place. Perhaps the earliest autoimmune disease caused by physicians was an encephalomyelitis (inflammation of the brain and spinal cord) that was inadvertently induced in people who had been bitten by a dog and who received multiple injections of the Pasteur vaccine to prevent rabies. The vaccine, prepared by growing rabies virus in the rabbit spinal cord, contained antigens that "cross-reacted" with antigens in the human central nervous system. That is, the rabbit and human antigens had at least some epitopes in common, and the antibodies whose production the vaccine stimulated recognized and interacted with the human epitopes.

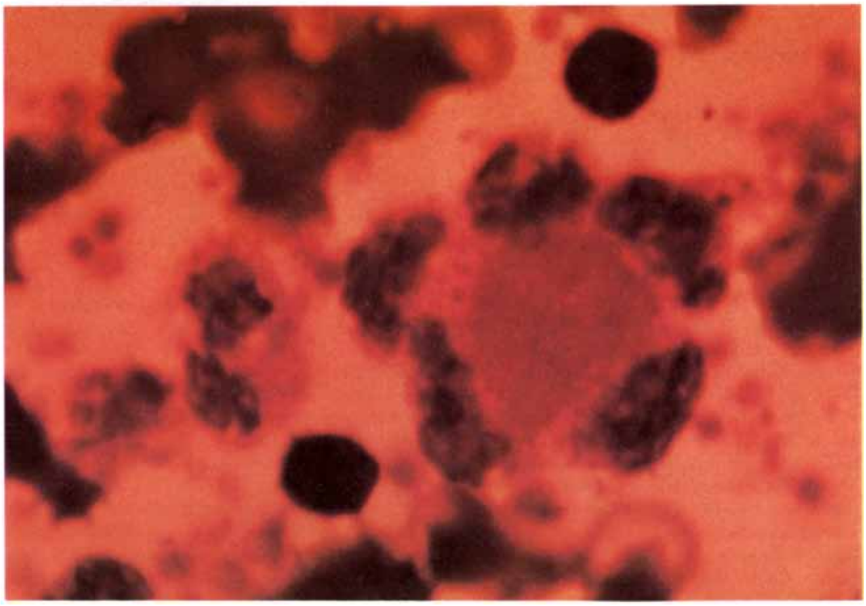
In 1933 Thomas M. Rivers of the Rockefeller Institute for Medical Research demonstrated that monkeys given a series of injections of homogenized nervous tissue developed a similar encephalomyelitis. It was not until 1956, however, that Ernest Witebsky and Noel R. Rose, working at the University of Buffalo School of Medicine, clearly defined autoimmune disease and reported a reproducible method whereby the natural tolerance of the immune system toward its own antigens could be abrogated. They immunized rabbits with extracts of the rabbit thyroid gland mixed with an adjuvant (a substance that enhances the immune response); the rabbits developed thyroiditis, or inflammation of their own thyroid gland. Witebsky and Rose defined criteria for an autoimmune disease: immunization with an antigen such as a particular thyroid protein initiates an immune response to the antigen followed by pathological changes that are limited to the organ containing the antigen. Later William O. Weigle of the Scripps Clinic and Research Foundation in La Jolla, Calif., found that a similar thyroiditis could be produced by immunizing rabbits with cross-reacting thyroid proteins from different species; the proteins were recognized as being foreign but had epitopes identical with epitopes on the rabbits' own thyroid gland. The mechanism of autoimmune disease, then, is a direct attack by the body's immune apparatus, most probably involving both antibodies and sensitized *T* cells, on the body's own tissues.

There is another mechanism for self-injury by the immune system. It was recognized in the early 1900's when Clemens von Pirquet of the Johns Hopkins University School of Medicine speculated about the cause of the "serum sickness" seen in people who had been treated (to counteract the effects of tetanus toxin, for example) with an antiserum prepared in horses. Von Pirquet suggested that serum sickness resulted from the interaction of the patient's antibodies with foreign antigens—proteins in the horse serum—to form a toxic substance. He was right. The substance is an immune complex, and it is responsible for the rash, joint pains and other symptoms of serum sickness. The mechanism of tissue injury in an immune-complex disease involves the formation of toxic antigen-antibody complexes, and such complexes can be made in the course of an immune response to an antigen.

Both autoantibodies and immune complexes are ubiquitous in mammalian serum. Their presence is almost surely not just an accident; they appear to be integral components of the immune system that have a vital role in preserving homeostasis, or the stability of the body's internal environment. According to the network theory proposed by Niels Kaj Jerne of the Basel Institute for Immunology, for example, a





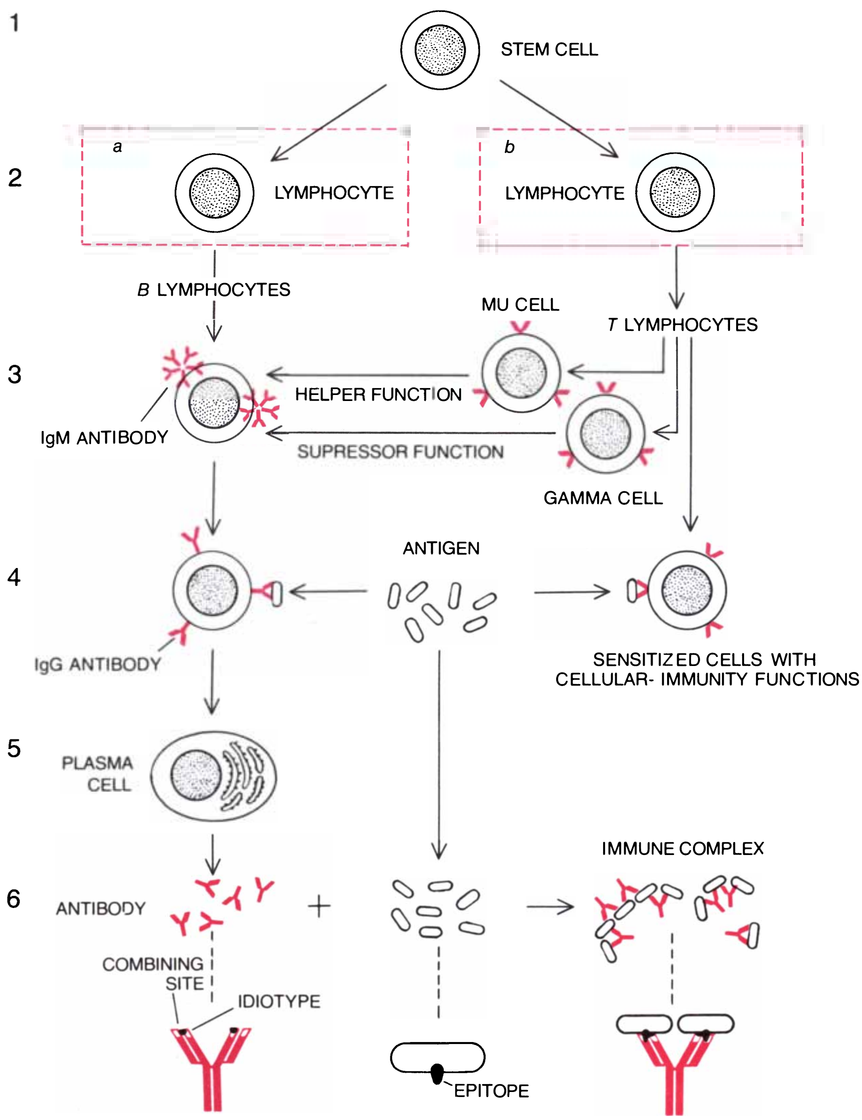


AUTOANTIBODIES that induce the formation of a lupus erythematosus (LE) cell provided the first diagnostic test for SLE. White blood cells are shaken in a test tube to damage them and are then incubated with a patient's serum. If the serum contains antinuclear antibodies, characteristic of SLE, that react with the DNA-protein complex of the nucleus, the antibodies penetrate damaged cells and cause their nuclei to swell, degenerate and be extruded as free nuclear bodies (*upper left in top micrograph*). A neutrophil may engulf the nuclear body, forming an LE cell (*lower right in top micrograph*); the neutrophil's own nucleus is compressed at the rim of the cell, as is shown more clearly by a stain for DNA (*middle micrograph*). Sometimes a nuclear body is surrounded by "rosette" of neutrophils (*bottom*).

particular class of autoantibodies, the anti-idiotypic antibodies, help to self-regulate the immune system [see "The Immune System," by Niels Kaj Jerne; SCIENTIFIC AMERICAN, July, 1973]. These are antibodies to epitopes (called idiotypes) that are at or near the combining site of the IgG molecule and are specific for antibody formed by a single clone of plasma cells. These unique epitopes were first described in man by Henry G. Kunkel of Rockefeller University and in animals by Jacques Oudin of the Pasteur Institute. Anti-idiotypic antibodies have been identified in experimental animals and induced anti-idiotypic antibodies have been shown to enhance or suppress the animals' immune response.

Among the commonest autoantibodies in man are "rheumatoid factors," a group of antibodies directed against gamma globulin. Rheumatoid factors are present in low concentrations in normal individuals and in much higher concentrations in patients with various diseases; they are particularly pronounced in rheumatoid arthritis and SLE, where they may have an important role in the disease process. Other autoantibodies are directed against the material in the cell nucleus. They are present in a small percentage of normal individuals, with a higher incidence in older people and (as will appear below) a much higher incidence in SLE patients. (I should point out that the distinction between a "normal" and an "abnormal" antibody level depends on the sensitivity of an assay and becomes increasingly obscure as more precise immunologic probes become available.)

What is the biological function of autoantibodies and immune complexes normally present in the absence of foreign challenge? All in all, the available evidence suggests that they have a central role in preserving homeostasis. The events associated with SLE represent malfunctions of normal homeostatic mechanisms, and the malfunctions are expressed clinically as the SLE syndrome, or group of symptoms.



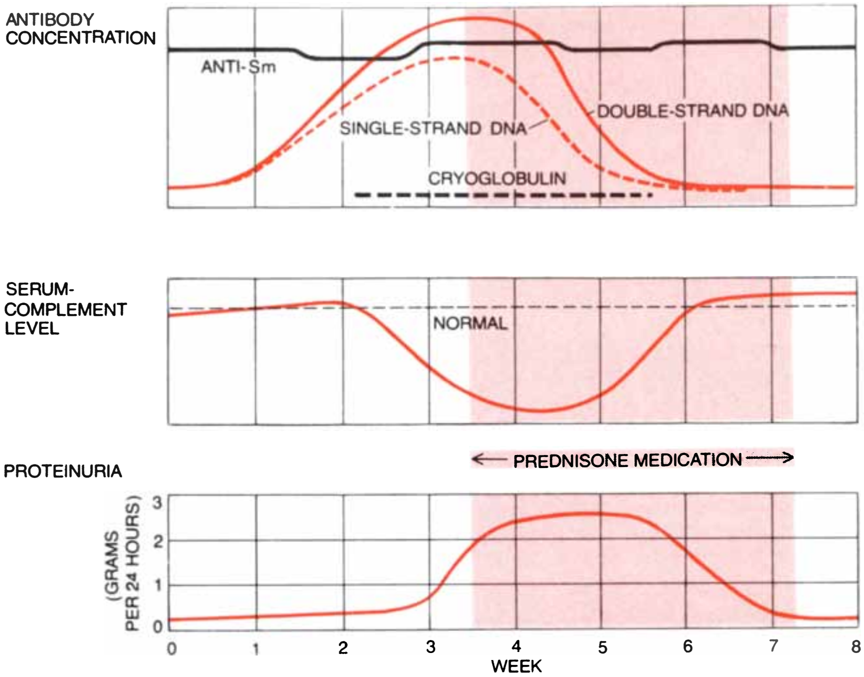
IMMUNE RESPONSE is effected by circulating antibodies and by sensitized cells. Stem cells (*1*) in the bone marrow differentiate, forming immature lymphocytic cells that migrate to lymphoid tissue associated with the intestine (*2a*) and to the thymus (*2b*), where they mature to become respectively *B* and *T* lymphocytes (*3*). A *B* cell is genetically programmed to make a few copies of a specific antibody, which act as receptors on the cell's surface. Different kinds of *T* cells, with somewhat similar receptors, have regulatory roles or become sensitized by antigen to function in cellular immunity. There are five classes of antibodies, of which IgM and IgG are the most important in SLE. Here an IgM-producing cell is "switched" to make IgG (*4*). An epitope (a small site on an antigen molecule) interacts with the combining site of an IgG receptor, causing the *B* cell to give rise to a clone of plasma cells (*5*) that synthesize large amounts of a specific antibody, which can be identified by its unique idiotype: an epitope in the antibody's combining site. The same idiotype may be present on the surface receptors of *T* cells that modulate the antibody response. The antibody can react with an epitope on the same antigen that originally stimulated the *B* cell to proliferate, thus forming an immune complex (*6*).

Lupus erythematosus was identified in the mid-19th century as a disease limited to the skin, but by the end of the century Moritz Kaposi in Vienna and William Osler at Johns Hopkins had described a number of systemic manifestations that were often associated with "facial lupus": arthritis, pneumonia and symptoms involving the central nervous system. Early in this century pathology of several abdominal organs was recognized as part of the clinical picture. In 1941 SLE was classified by Paul Klemperer, Abou D. Pollack and George Baehr of Mount Sinai Hospital in New York as a "collagen-vascular disease," a disease affecting connective tissue and blood vessels.

It was the discovery of the lupus erythematosus (LE) cell in 1948 by M. M. Hargraves of the Mayo Clinic that drew attention to SLE's immunological aspects. Halsted R. Holman and Kunkel found that the serum of SLE patients contained a gamma globulin with an affinity for nuclei and specifically for the nucleoproteins: the proteins complexed with DNA in the chromosomes. It turned out that a variety of antibodies that react with nuclei are present in patients' serum, one of which, an antibody to the DNA-protein complex, was shown to be responsible for the phenomenon of the LE cell. The observation that most SLE serums contained antibodies reactive with "self" antigens led to the view that SLE was an autoimmune disease.

It is difficult, however, to invoke autoimmune reactions to explain the typical clinical findings in SLE patients: fever, fatigue, skin rash, arthritis, vascular inflammation, pleurisy, pericarditis and glomerulonephritis: the kidney disease, characterized by inflammation of the renal glomeruli (which filter the blood), that is one of the most serious manifestations of SLE. Antinuclear antibodies, a common type of autoantibody, do not ordinarily penetrate the membrane of intact living cells. And one does not usually find deposits of antinuclear antibodies in the tissues of SLE patients, as one would expect to if such antibodies were responsible for the pathology.

The clinical and pathological manifestations of SLE resemble those of serum sickness much more than they do those of a typical autoimmune disease—and serum sickness results from the formation of immune complexes. Evidence suggesting a direct pathogenic role for antigen-antibody complexes has come from several different kinds of investigation. First, at about the time circulating antinuclear antibodies were found in SLE serum, deposits of gamma globulin were found in patients' glomeruli by Robert C. Mellors of the Cornell University Medical College and in small arteries of the liver and spleen and in LE cells by Jacinto J. Vazquez and Frank J. Dixon, who were then working at the University of Pittsburgh School of Medicine. The circulating antibodies were found to be directed primarily against nuclear antigens: DNA, in both its "native" double-strand form and its "denatured" single-strand form, and DNA complexed with the protein histone. Eng M. Tan, Kunkel and their colleagues showed that the serum level of DNA antigens and of antibodies to them fluctuated in several patients, suggesting an association between immune-complex formation and increasing disease activity. Subsequent studies have confirmed a striking correlation between flare-ups of disease activity and a reduction in the serum level of complement, which is fixed by antigen-antibody complexes and deposited in tissues or cleared by reticuloendothelial cells.



FLARE-UP OF DISEASE is traced for a typical SLE patient with limited kidney disease. The blood serum's content of antibodies against double- and single-strand DNA increases, and the complement level begins to decline, before the onset of proteinuria (increased urinary excretion of protein), a sign of kidney malfunction. Note the rapid decrease in antibodies associated with extreme complement depression; the antibodies have combined with antigen to form immune complexes, which fix complement (and which may be observed in the serum as cryoglobulin, a gel-like precipitate formed at low temperature). Another antinuclear antibody, anti-Sm, does not increase during a flare-up; apparently only certain antibodies participate in forming immune complexes capable of inducing tissue damage. Prednisone, an anti-inflammatory steroid drug, is effective in treating an acute episode; its mechanism of action is not understood.

Eventually, with Kunkel and others in his laboratory, I was able to provide direct visual evidence that antigen-antibody complexes can be important mediators of both vascular and kidney injury in SLE, depending on the site of deposition. We examined kidney biopsy samples taken from patients at various stages of disease. Light microscopy with conventional stains showed different degrees of tissue injury. Fluorescence microscopy of samples exposed to an antibody labeled with a fluorescent dye revealed, in most patients with SLE, deposits of immune complexes in the mesangium, a region of the glomerulus that clears the blood of large molecules.

When these deposits were limited to the mesangium, there was either minimal evidence or none of proteinuria (excess protein in the urine, the major laboratory sign of glomerulonephritis). In samples from patients with increased proteinuria, on the other hand, granular deposits of gamma globulin together with complement provided evidence for the presence of complexes, which were seen along the basement membrane, the membrane through which the blood is filtered. Continued deposition of these complexes gives rise to a thickening of the capillary walls. Observation of just such thickening in SLE patients' glomeruli had led Klemperer in 1935 to relate SLE to serum-sickness lesions in the kidneys of horses given repeated injections of diphtheria toxoid in order to prepare antiserum.

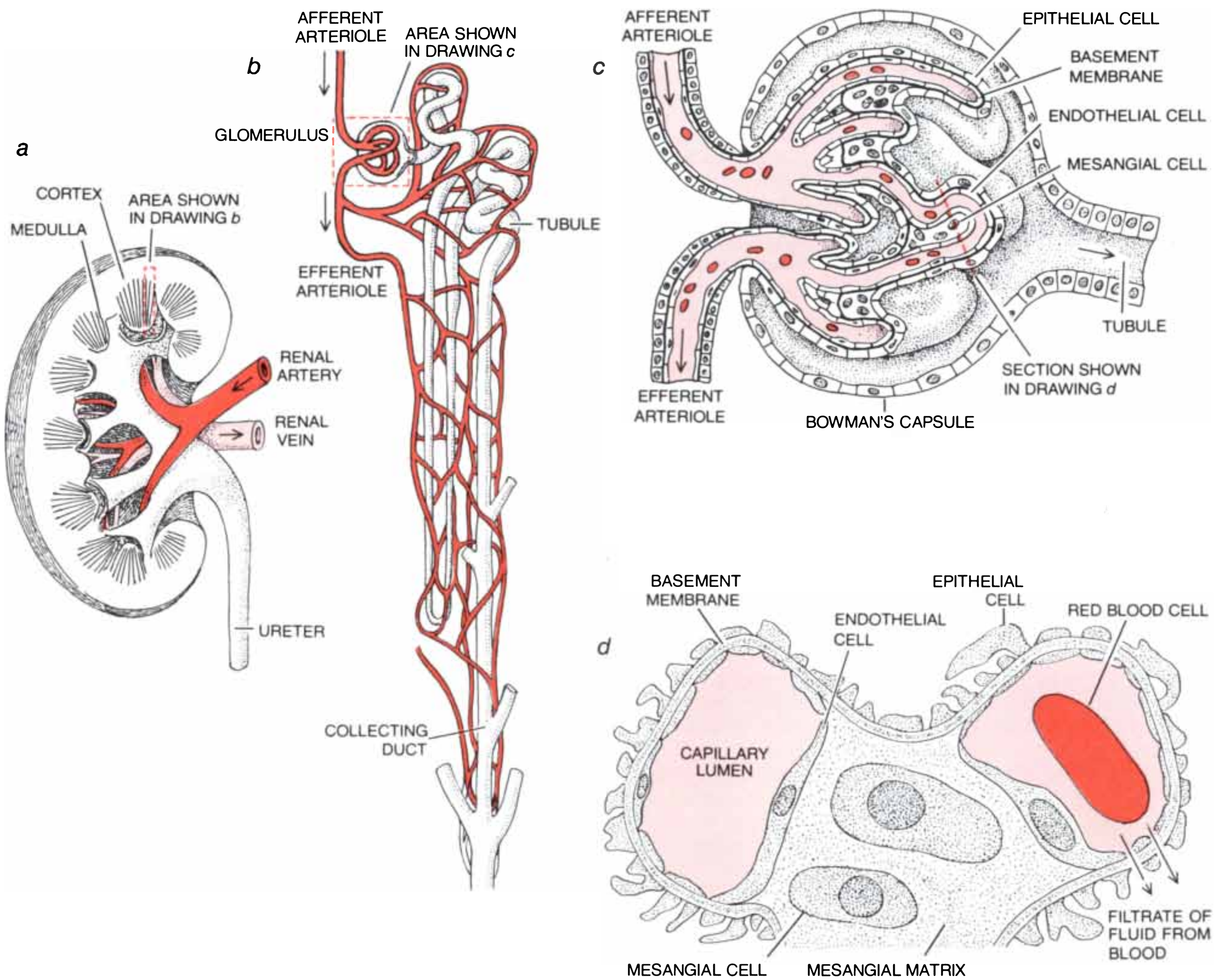
The role of complexes in SLE remained hypothetical until Frederick G. Germuth, Jr., who was then at Johns Hopkins, did animal experiments in the 1950's that established their importance and specifically related the extent of disease to the immune response to an antigen. Germuth and his colleagues showed that long-term administration of protein antigens to rabbits gave rise to a chronic disease similar to serum sickness in animals and to glomerular lesions like those in the kidneys of SLE patients. When the rabbits were immunized with a constant dose of a protein (bovine serum albumin), the severity of the glomerulonephritis depended on the individual rabbit's antibody response. In rabbits with a strong response (large amounts of antibody) extensive lattices of linked antigen-antibody complexes were formed; these were readily removed by the reticuloendothelial system and seemed not to be toxic. Rabbits with a lesser antibody response formed slightly smaller complexes, which had an affinity for the mesangium. Rabbits with a still weaker antibody response formed still smaller complexes, and developed a diffuse proliferative glomerulonephritis, the severe form of the disease. Lower antibody responses yet gave rise to yet smaller complexes and no signs of disease.

Earlier studies by Theodore P. Pincus and Charles L. Christian of the Columbia University College of Physicians and Surgeons had indicated that the intrinsic properties of antibodies against bovine serum albumin influenced the toxicity of complexes. Germuth, on the basis of subsequent investigation at the Saint Louis University School of Medicine, emphasized the role of such qualitative characteristics of the antibody as its avidity, or tendency to bind antigen. Dixon and his colleagues, who are now at the Scripps Clinic, did a series of critically important experiments in which they adjusted the antigen dose to maintain a state of moderate antigen excess, that is, a state in which there is a little more antigen than antibody in the complexes. Such complexes were particularly likely to produce chronic membranous glomerulonephritis. The onset of the nephritis was associated with a marked increase of deposits in the kidney, suggesting that the initial injury facilitated increased deposition and that a possible contributing factor was saturation of the reticuloendothelial system (so that it could no longer dispose of the complexes). Robert T. McCluskey and Baruj Benacerraf, who were then at the New York University School of Medicine, went on to produce both glomerulonephritis and vasculitis (inflammation of the blood vessels) in mice directly by administering soluble immune complexes that had been preformed in the laboratory.

In other words, immune complexes vary in their toxic potential, and this explains in part the wide range of clinical symptoms. The site of deposition and the initiation of an inflammatory response are influenced by the size of the complex, the number of binding sites on the antigens and the antibodies, the number of different antibodies and their avidity for the antigens. Charles G. Cochrane of the Scripps Clinic found, moreover, that agents tending to increase the permeability of blood vessels also had an effect on the toxicity of complexes. Treating rabbits with antagonists of such agents as histamine and serotonin reduced the deposition of complexes and the ensuing inflammation. Mart Mannik of the University of Washington School of Medicine showed that the ability of the reticuloendothelial system to remove complexes is affected by the ability of phagocytes to recognize a particular part of the antibody.

In some SLE patients with kidney symptoms the disease remains limited to a focal glomerulonephritis, in which only some of the glomeruli have immune-complex deposits and suffer injury; drugs that reduce inflammation or suppress the immune response seem to be particularly effective at this stage. In other patients the disease progresses to the diffuse, proliferative phase marked by massive deposits in most glomeruli and by significant proteinuria. Recently Giuseppe A. Andres of the State University of New York School of Medicine at Buffalo and McCluskey, who is now at the Harvard Medical School, have stressed the importance of injury to the kidney tubule (where fluids and dissolved substances are reabsorbed from the glomerular filtrate) as the result of complex deposition in tissue around the tubule.

In an effort to characterize the nature of complexes found in progressive forms of the disease, Kunkel, Peter H. Schur and I examined the kidneys of patients who had died as the result of SLE kidney disease; John B. Winfield and I examined the cryoglobulins (immune complexes precipitated at low temperature from the serum) of SLE patients.



GLOMERULONEPHRITIS, an inflammation of capillary loops in the glomeruli of the kidney, is a serious manifestation of SLE. The kidney (*a*), where toxic substances and fluids are removed from the blood for excretion as urine, is composed of some one million of the functional units called nephrons. Each nephron (*b*) consists of a tuft of looped capillaries, the glomerulus, where the blood is filtered, and a tubule through which the filtrate passes to a collecting duct; much of the fluid and some dissolved substances are reabsorbed from the tubule into a network of capillaries surrounding the tubule. The glomerulus (*c*) is a lobed tuft of interconnected capillaries. A section of the glomerulus (*d*) shows the lumen, or bore, of capillary branches wrapped in a basement membrane covered by epithelial cells; the membrane and the epithelial cells control filtration. Large molecules are removed from the blood and are deposited in the mesangium.

Several antibodies were eluted (washed out) from the glomeruli and similar ones were isolated from cryoglobulins. Antibodies from both sources were found to be more highly concentrated than they were in the serum as a whole, suggesting that they had been concentrated in antigen-antibody complexes. The antibodies were directed against double-strand DNA, single-strand DNA and ribonucleoprotein; complexes containing DNA were most prominent, and DNA antigen could be demonstrated in glomeruli containing anti-DNA antibodies.

The DNA immune complex was the first specific tissue-damaging agent identified in glomerulonephritis in man; other antigens have been localized in the kidneys of patients with "primary" glomerulonephritis—kidney disease similar to that of SLE but not associated with it. Because its many symptoms give notice of kidney involvement early SLE has served as a prototype for the study of immune-complex disease in general.

Winfield, Kunkel and I found that the anti-DNA antibodies eluted from glomeruli have a higher avidity than the same kind of antibodies in patients' serum. Moreover, the circulating anti-DNA antibodies in patients with kidney disease have lower avidity than those in patients without kidney disease. These findings suggest that in glomerulonephritis the high-avidity antibodies are preferentially deposited as complexes in the kidney. The result is that monitoring the serum for circulating antibodies alone may fail to detect a significant antibody because it is part of a complex. The ability to identify circulating complexes and so evaluate disease activity has been facilitated by several important advances, including the isolation of complement components and the elucidation of the complement pathways by Hans J. Müller-Eberhard of the Scripps Clinic, by Shaun Ruddy of the Medical College of Virginia and by others.

One way to get at complexes is by examining the cryoglobulins, whose interactions with complement had been studied by Christian. Vincent Agnello, Kunkel and I found that patients with particularly large amounts of cryoglobulins have prominent deposits of IgM in their kidneys, in contrast to most patients with SLE kidney disease, in whom IgG is the major immunoglobulin; IgM rheumatoid factor is also a major component of circulating cryoglobulins in SLE. These findings indicate that complexes of gamma globulin and antibodies against it contribute to kidney damage in at least one group of patients. Recently Morris Reichlin of the State University of New York School of Medicine at Buffalo has found still another immune-complex system involving antigens from the cytoplasm of cells, again suggesting that various complexes can contribute to glomerular damage.

The observation by Agnello, Robert J. Winchester and Kunkel that C1q (the complement protein that interacts with antibody) forms a precipitate with aggregates of immunoglobulin provided a way to detect circulating complexes. Agnello, Kunkel and I found such precipitates in the serum of SLE patients. Paul H. Lambert and Peter A. Miescher of the University of Geneva have developed another assay for complexes that utilizes C1q labeled with a radioactive isotope. And Argyrios N. Theofilophoulos and Dixon have developed an assay exploiting lymphoid cells (called Rajii cells) that have receptors for components of the complement fixed to complexes. These and other assays for immune complexes provide information about the course of SLE that cannot be obtained by monitoring only the level of antigen or antibody.

Whereas a clear relation has been demonstrated between immune complexes and both kidney disease and blood-vessel injury, it has been difficult to identify the mechanism of tissue injury in other manifestations of SLE. Immunofluorescence studies of skin biopsies show, for example, that about half of all SLE patients have immune-complex deposits in the skin just below the outer layer (the epidermis), but the deposits do not always cause inflammation. Tan and Kunkel observed granular deposits of gamma globulin both in inflammatory lesions and in areas of quite normal skin. Morris Ziff and James N. Gilliam of the University of Texas Health Sciences Center in Dallas and Reichlin reported that the incidence of such deposits in normal skin is increased in patients with kidney disease, who often have larger amounts of circulating immune complexes. The deposits may also be found in "discoid" SLE, where skin changes are the major symptom. There are many unanswered questions about these lesions and their associated deposits. What differentiates the effect of complexes in normal skin from the effect in inflamed skin: the properties of the deposits themselves or the patient's reaction to them? Are complexes deposited in the skin different from those deposited in glomeruli? If they are, why?

The basis of the neurological symptoms in SLE, which tend to be seen only some years after the onset of the disease and are being encountered more often as the life span of SLE patients lengthens, is still poorly understood. There is little evidence connecting deposits with any lesions in the brain or even connecting particular lesions with such symptoms as convulsions and psychoses. A variety of antibodies have been found in SLE patients' serum that react with nervous-system antigens, but they have not been linked to neurological disease.

What evidence there is for a direct effect of autoantibodies (as distinguished from complexes) in SLE primarily involves the interaction of circulating antibodies with cellular elements of the blood. Antibodies that react with antigens on the surface of red blood cells may be associated with hemolytic anemia, a state in which the red cells are damaged and the hemoglobin level is therefore reduced.

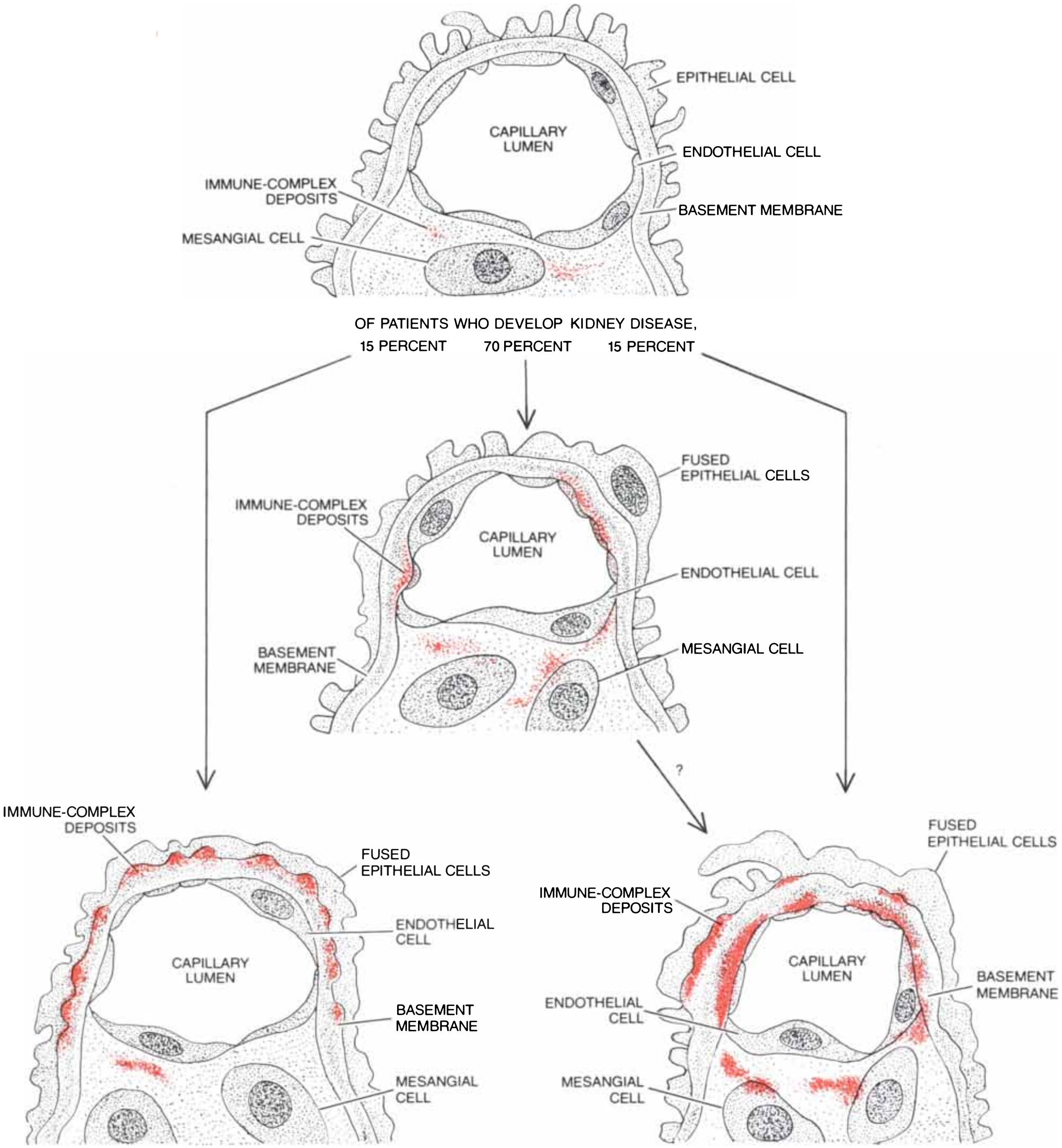
Antibodies against both *B* and *T* lymphocytes are found in the serum of SLE patients. Defects in *T*-cell function, including reduced suppressor-cell activity, have been observed; one population of antibodies against *T* cells cross-reacts with antigens similar to antigens in the mouse brain, but it is not yet known whether these antibodies alter cellular immune function or affect brain tissue in patients. Winfield, Winchester and Kunkel noted that the presence of certain antilymphocyte antibodies in SLE serum is associated with a reduced level of circulating lymphocytes. They also found a high concentration of such antibodies in cryoglobulins, suggesting that they may form immune complexes with surface antigens released from circulating lymphocytes.

Apart from the antibodies responsible for hemolytic anemia, however, autoantibodies have not been clearly shown to be directly responsible for tissue injury in SLE. There is now general agreement that it is immune complexes that cause tissue injury in blood vessels and glomeruli. And studies by Andres of the synovial membrane in joints, of the pleura and pericardium (respectively the membranes enclosing the lung and the heart) and of the alveolar septae (the walls of the air sacs of the lung) suggest that complexes may be responsible for many symptoms of SLE.

Up to this point I have been discussing the mechanisms of tissue injury in SLE. To say that SLE appears to be an immune-complex disease, however, is not to identify its cause or causes. That remains to be done; so far only a few etiological facts have been established. There is a genetic predisposition to the development of SLE; abnormalities of the patient's immune system are implicit in the SLE syndrome; a variety of factors have been associated with the disease, including certain drugs, estrogenic hormones, ultraviolet radiation and viral infection.

The most striking evidence for a genetic determination is the fact that if one of two identical twins develops SLE, the other twin also develops the disease in from 50 to 60 percent of the instances, whereas the concordance is only 2 or 3 percent in fraternal twins and in other first-degree relatives of SLE patients; the incidence in the general population is only about .2 percent. Several rare congenital abnormalities in the complement proteins have been associated with SLE. Genes controlling the synthesis of some complement proteins are linked with (on the same chromosome as) genes controlling the immune response and the production of histocompatibility antigens: a set of molecules that differ from individual to individual and are recognized in the rejection of grafts. The incidence of SLE is markedly increased in individuals with certain histocompatibility types. All of this is evidence for a genetic influence in SLE, but the mechanism whereby genetic factors affect the disease remains to be explained. The linkage between a genetic factor and the disease could be dependent on some auxiliary factor. For example, a defect in the complement system or the immune system might represent a predisposition to a viral infection that in turn gives rise to the disease.

Dysfunction of both the antibody and the cellular components of the immune system have been demonstrated in SLE. Patients have many different autoantibodies that have a particular propensity for immune-complex formation, one of which, the antibody against double-



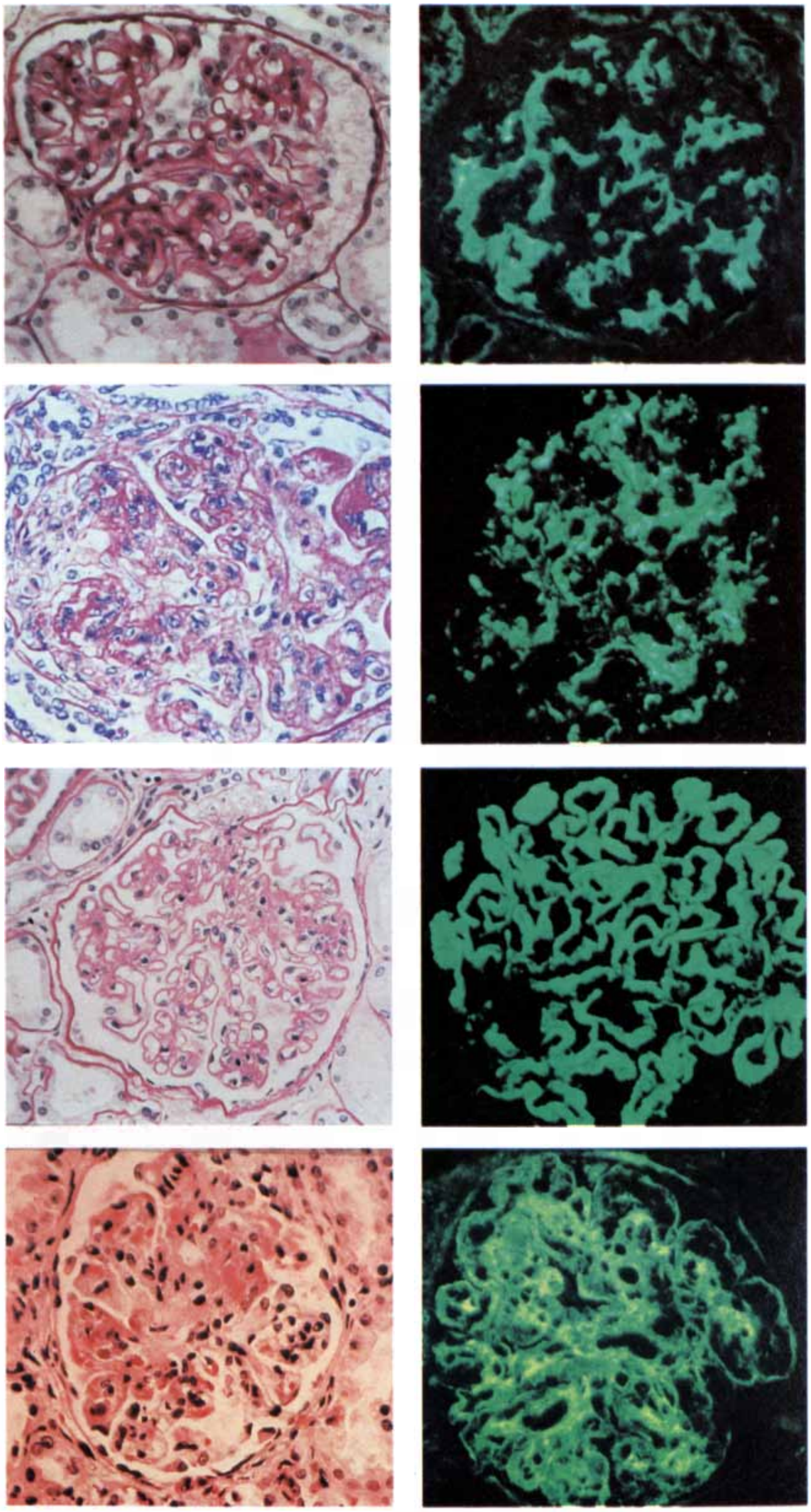
IMMUNE-COMPLEX DEPOSITION in the kidney is diagrammed at different stages of SLE. Even in a patient without kidney disease a structurally normal glomerulus has deposits of immune complexes in the mesangium (*top*). Two-thirds of SLE patients have kidney disease; of these about 70 percent develop focal proliferative glomerulonephritis and minimal or moderate proteinuria (*middle*), with spotty deposits along the basement membrane under the endothelial-cell lining. Some 15 percent develop membranous glomerulonephritis, with moderate or marked proteinuria (*bottom left*); continued deposition results in fairly homogeneous linear deposits outlining the basement membrane. Another 15 percent (and also some proportion of those with focal disease) develop diffuse proliferative glomerulonephritis, with massive deposits of immune complexes in most glomeruli (*bottom right*). Diffuse disease usually leads to kidney failure.

strand DNA, is present almost exclusively in SLE. One subpopulation of anti-DNA antibodies capable of forming complexes that fix complement has been closely linked to tissue destruction. The infrequent occurrence of these antibodies in other diseases and the fact that one cannot induce antibody formation by immunizing animals with double-strand DNA suggest that the immune system is normally resistant to recognizing DNA as a foreign antigen. Nevertheless, circulating *B* cells with receptors for DNA antigen were found in normal individuals, and more such cells were found in SLE patients, by Arthur D. Bankhurst and Ralph C. Williams, Jr., of the University of New Mexico School of Medicine. It would seem that clones of cells capable of producing antibodies against this particular self-antigen have somehow escaped, in SLE patients, from the system of checks and balances that normally modulates the immune system.

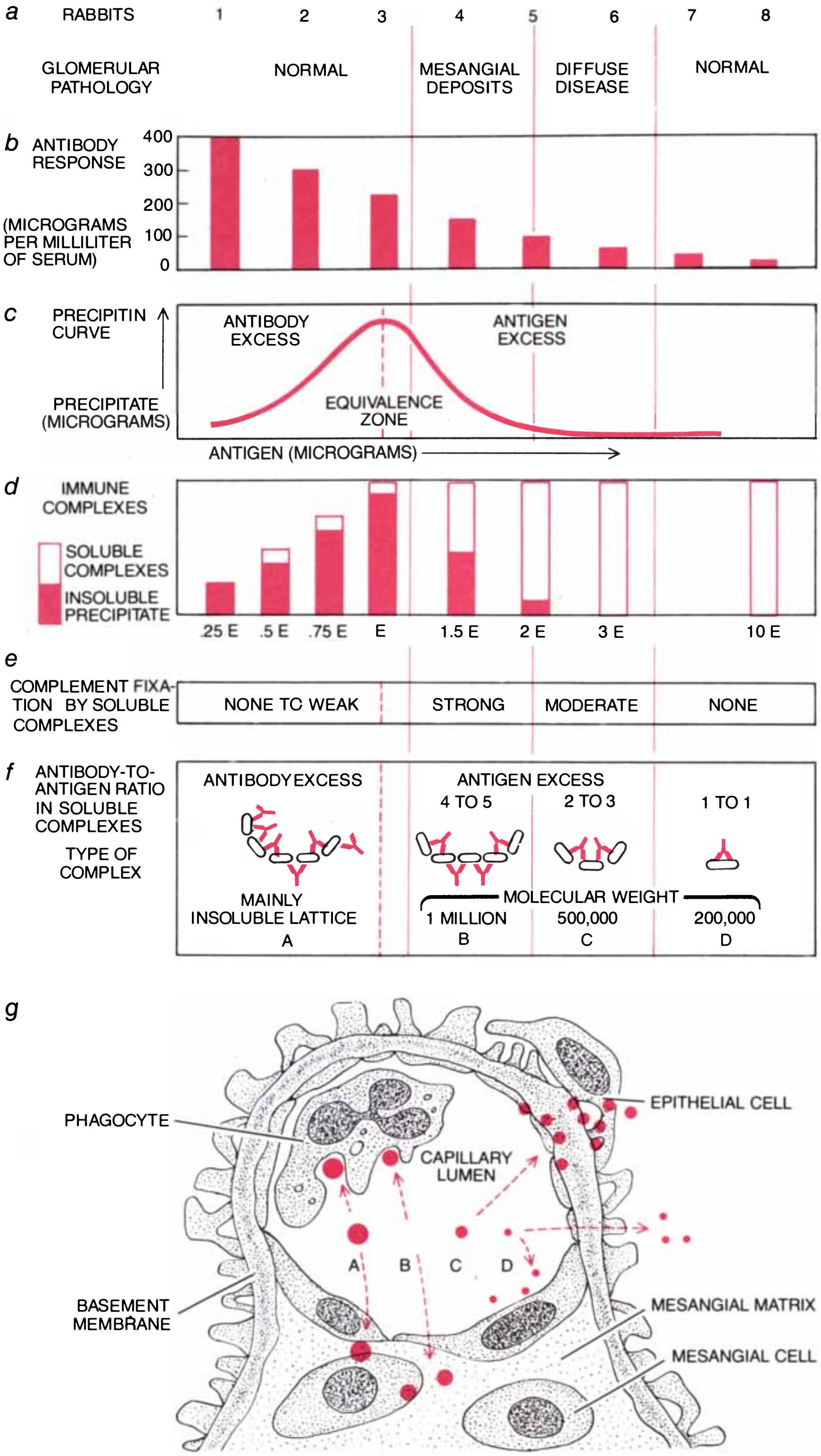
How might that happen? Several possible explanations are suggested by studies of cellular immune mechanisms. Both SLE patients and a strain of mice that develop a disease resembling aspects of SLE are deficient in suppressor-*T*-cell function. Suppressor-cell function decreases with age in mice with an SLE-like syndrome; the decrease is associated with enhanced production of anti-DNA antibodies and with a virulent form of glomerulonephritis. Alfred D. Steinberg, Anthony S. Fauci, M. Eric Gershwin and Thomas A. Waldmann of the National Institutes of Health have presented particularly convincing evidence for suppressor-cell defects and information on the functioning of *B* cells in both mice and human SLE patients.

In man the level of antinuclear autoantibodies has been shown to increase with age, perhaps as a result of reduced suppressor-cell function; no symptoms have been attributed to the increase. It has been suggested that certain antibodies against *T* cells demonstrable both in man and in mice contribute to suppressor-cell abnormalities. Peter Wernet and Kunkel have demonstrated antibodies that interact with *T* cells and inhibit proliferation of the *T* cells. Such inhibition, if it affects suppressor cells in the body, could tend to free antibody-producing *B* cells from suppression. Robert S. Schwartz of the Tufts University School of Medicine has recently reported suppressor-cell abnormalities in the families of SLE patients. It is possible, on the other hand, that some factor such as interaction with circulating immune complexes could cause helper *T* cells to proliferate and so induce more antibody production by *B* lymphocytes.

Data from long-term studies of "SLE dogs" by Schwartz and Robert M. Lewis suggest that both genetic and environmental factors are involved. These dogs



TISSUE SPECIMENS representing four stages of kidney disease show the structural changes revealed by histological staining (*left*) and the localization of immune complexes by fluorescence microscopy (*right*). The structurally and functionally normal glomerulus from a biopsy (*top left*) has irregular strandlike deposits of immune complexes in mesangial areas (*top right*). A glomerulus from a patient with focal kidney disease (*second from top*) manifests irregular thickening of the membrane and some fusion of capillary tufts; irregular granular deposits of complexes are bound in the capillary walls. In membranous disease (*third from top*) there is increased thickening of the membrane (*left*) and "wire loop" deposits are observed (*right*). In diffuse disease (*bottom*) severe structural damage (*left*) and a massive buildup of immune deposits (*right*) can combine to obliterate normal architecture of glomerulus and disrupt filtration.



SIZE OF COMPLEXES was related to antibody response and disease level in rabbits by Frederick G. Germuth, Jr., of the Johns Hopkins University School of Medicine; properties of the complexes were studied in the author's laboratory. Given a constant level of injected antigen, the pathology seen in eight rabbits (*a*) was related to the amount of antibody formed by each animal (*b*). A precipitin test (done in the laboratory by adding increasing amounts of antigen to a constant quantity of antibody) shows how the amount of precipitated (insoluble) immune complexes varies with the antibody-antigen ratio (*c*). The curve is interpreted by the bars (*d*): the amount of precipitate increases until the antigen is at equivalence (*E*) with antibody, then decreases. Complexes vary in ability to fix complement, a major determinant of toxicity (*e*). The size of the complex, which is related to the antibody-antigen ratio (*f*), affects the site of deposition (*g*). Very large complexes (*A, B*) tend to be disposed of by mesangial cells and circulating phagocytes; complexes with a molecular weight of about 500,000 (*C*) are deposited along the glomerular membrane and lead to significant kidney disease; very small complexes (*D*) either recirculate or are filtered into the urine without causing any damage to kidney tissue.

develop a syndrome with remarkable similarities to human SLE, including the characteristic rash and the presence of specific antibodies in the serum. After sustained inbreeding, however, some of the dogs show the serum abnormalities without any evidence of clinical disease, as is often the case in relatives of human SLE patients. Lawrence E. Shulman of Johns Hopkins has concluded from his analysis of SLE patients' families that there is clear evidence for both genetic and environmental influences.

If both a genetic predisposition and an auxiliary precipitating factor are prerequisites for the development of clinical SLE, what might such an auxiliary factor be? Several drugs have been associated with an increased incidence of antinuclear antibodies. In particular procainamide (given to treat cardiac arrhythmia) and hydralazine (for hypertension) induce antibodies against single-strand DNA, nucleoproteins and possibly other nuclear antigens; prolonged administration of procainamide gives rise to an SLE-like syndrome in about a third of the recipients. The drug-induced symptoms are self-limiting, ending when the drug is withdrawn, and rarely include kidney changes or antibodies against double-strand DNA.

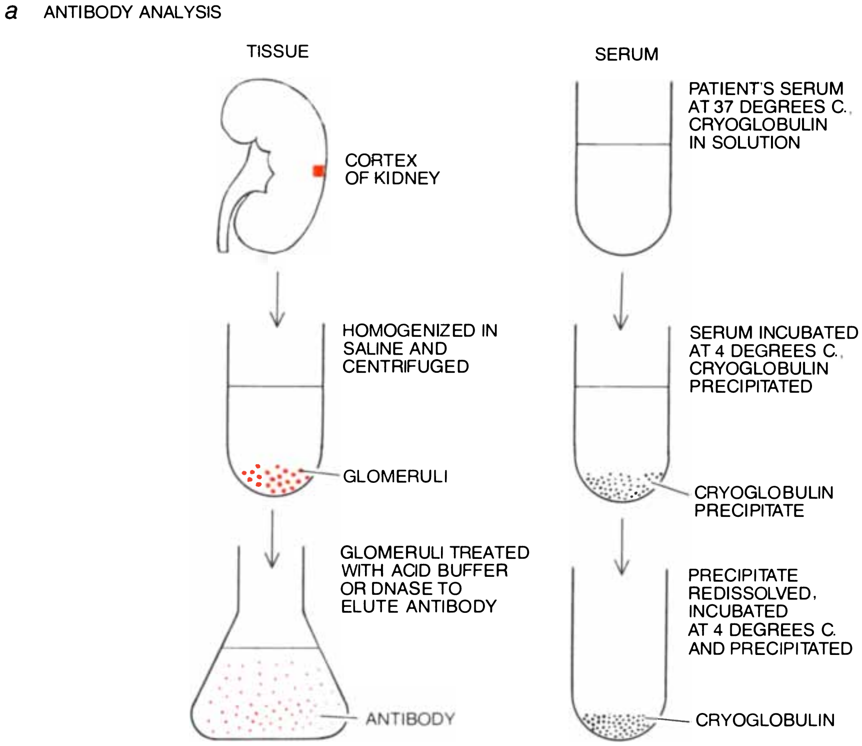
As an experimental model, however, drug SLE provides an opportunity to study mechanisms by which autoantibodies are induced. For example, procainamide and hydralazine are metabolized by acetylation, and Marcus M. Reidenberg and Dennis E. Drayer of Cornell found the drugs' effects to be most pronounced in "slow acetylators": people who have a reduced level, genetically determined, of the acetylating enzyme. Robert G. Lahita, Reidenberg, Drayer and I showed that patients given an acetylated analogue of procainamide instead of the usual form of the drug develop the autoantibodies more slowly and in smaller amounts. So far there is no evidence implicating any drug or other chemical agent as a factor precipitating SLE itself, but the possibility needs to be investigated.

An association with estrogenic hormones is suggested by several observations. SLE occurs primarily in women during their reproductive years, when the estrogens are at peak levels; taking estrogenic oral contraceptives sometimes exacerbates the disease; men rarely develop SLE, but the disease is seen more often in men with the rare congenital condition called Klinefelter's syndrome, who have two *X* chromosomes and a *Y* chromosome and have abnormal estrogen metabolism. Norman Talal of the University of California School of Medicine at San Francisco and his colleagues have shown that the severity of mouse SLE is reduced in female mice deprived of estrogen (by removal of the gonads) or given male hormones, and that the disease is exacerbated in male mice that are castrated or given estrogens. Recent studies by Lahita, Kunkel and Jack Fishman have demonstrated abnormally high levels of certain estrogen metabolites in both male and female SLE patients.

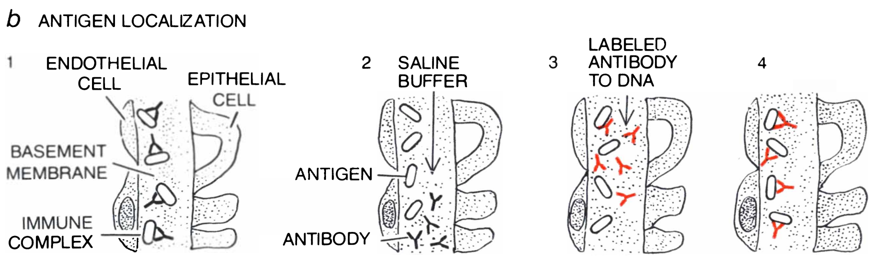
The sensitivity to sunlight generally manifested by SLE patients implicates ultraviolet radiation as a possible factor. Skin lesions tend to appear on parts of the body exposed to the sun; occasionally the systemic symptoms are exacerbated. Ultraviolet irradiation can alter the strands of the DNA double helix. Some of the abnormal structures, such as thymidine dimers (in which two of the DNA's constituent nucleotide bases are linked), have been shown to serve as immunogenic epitopes in experimental animals. Ultraviolet irradiation might enhance the immune response and disrupt the natural tolerance of DNA. These hypotheses are not proved but are reasonable extensions of the available data.

The role of viral infection in SLE has been the subject of considerable speculation and investigation, with major interest centering on the role of RNA viruses of the *C* type. Viral proteins have reportedly been observed in SLE tissues, but their significance is not yet clear. Viral infection might well be not a precipitating factor but rather the result of an immune deficiency or of suppression of the immune response by drugs given to treat SLE; viral genetic material might be incorporated in human tissues without actually producing new virus particles and without affecting the development of SLE. Circumstantial evidence for an infectious causation has been presented by Raphael J. DeHoratius and Ronald P. Messner, who were then working at the University of New Mexico School of Medicine; they reported an increased incidence of certain antilymphocyte antibodies in household contacts of SLE patients who were not blood relatives. Studies by Schur, Kunkel and me and by Talal of antibodies that react with RNA have also raised the possibility that viral antigens may be involved; it is hard, however, to exclude the role of cellular rather than viral antigens.

It is clear that the mechanisms of tissue injury in SLE have so far been more clearly established than any causative factor. The most one can say about etiology is that the disease probably arises from an interplay of genetics and a factor or factors such as drugs, hormones and viruses. Further progress toward understanding the cause of SLE will require better understanding of normal immune functions.



ANTIBODY AGAINST:	GLOMERULAR ELUATE	CRYOGLOBULIN
DOUBLE-STRAND DNA	6 OUT OF 9	21 OUT OF 29
SINGLE-STRAND DNA	8 OUT OF 9	16 OUT OF 28
NUCLEAR RIBONUCLEOPROTEIN	3 OUT OF 9	3 OUT OF 13



NATURE OF COMPLEXES associated with severe kidney disease was established (*a*) by examining autopsy specimens of glomeruli and of cryoglobulin: a serum fraction, mainly containing immune complexes, that forms a gel-like precipitate at low temperature. Certain antibodies were found to be present in the glomeruli and cryoglobulin in much higher concentration than in the serum, indicating that they had tended to form immune complexes. The table shows, for each of three antibodies, the proportion of cases in which this preferential concentration was more than fiftyfold. Further evidence of immune-complex localization in glomeruli was obtained (*b*) by dissociating antibody from complexes in kidney sections, adding labeled anti-DNA antibody and noting its binding to DNA antigen in same sites as antibody deposits.

The increasing ability of physicians to diagnose SLE at an early stage, the clarification of mechanisms of tissue injury and the enhanced ability to assess disease activity through blood testing have led to improved methods of treatment that mitigate symptoms, particularly for kidney manifestations. More than 90 percent of the patients are alive five years after developing SLE, compared with 50 or 60 percent 25 years ago. Better diagnosis has also made it clear that the disease is much more prevalent than it was thought to be.

Studies of the immunology of SLE have made basic contributions to our knowledge of how immune agents injure tissue. Evidently in SLE something has gone awry in the immunological ecosystem; the fragile symbiosis of its various intercommunicating elements—*B* cells, *T* cells, antibodies, complement components and others—has been disturbed. Investigation of the immune system's disordered functioning in SLE should help to clarify the "language" whereby these elements normally communicate with one another. It should also lead to a more precise understanding of the immunological aberrations in other diseases, such as rheumatoid arthritis, multiple sclerosis and chronic glomerulonephritis.

Gamma-Ray-Line Astronomy

Astrophysical processes emit fine-tuned electromagnetic radiation at energies even greater than those of X rays. The resulting spectral lines can now be detected by instruments on balloons and satellites

by Marvin Leventhal and Crawford J. MacCallum

The nature of the universe has been deduced almost entirely from the photons, or quanta of electromagnetic energy, that arrive in the vicinity of the earth. Until half a century ago astronomers could detect only photons with energies of between 1.5 and 3.5 electron volts: the photons of visible light. Then they began to extend the photon energy range downward into parts of the infrared and radio regions of the electromagnetic spectrum and upward into the near ultraviolet. With the advent of rockets, high-altitude balloons and artificial satellites they were able to extend it much farther upward to the energy range of photons that cannot penetrate the earth's atmosphere: the photons of the far ultraviolet, X rays and gamma rays.

Gamma rays are the most energetic form of electromagnetic radiation; the energy of their photons is measured in millions of electron volts (MeV), and in principle it has no upper limit. Gamma-ray photons from space were first detected some two decades ago. The early detectors simply recorded the arrival of the photons without being able to analyze their energies, as the photons of light are analyzed into spectral lines by a spectrograph. Now, however, instruments have been developed that can detect gamma-ray spectral lines. They are beginning to yield information on the high-energy processes and objects that command the attention of modern astronomers, such as supernovas, neutron stars and phenomena at the center of galaxies.

Whereas the lines in the optical spectrum arise from transitions between the energy levels of electrons in atoms, lines in the gamma-ray spectrum arise from transitions between the energy levels of atomic nuclei. Consider the nucleus of neon 22, which consists of 10 protons and 12 neutrons. The nucleus is usually in its ground state, or lowest energy level, but it can also occupy excited energy levels. If the neon-22 nucleus is in its lowest excited state, it will decay to the ground state in a few trillionths of a second. The decay will be accompanied by the emission of a gamma-ray photon with an energy of precisely 1.2746 MeV. That amount of energy corresponds to the difference between the two states. The gamma ray is thus the signature of a specific nuclear transition.

The transition can occur as a consequence of several different events. In one event a nucleus of sodium 22, which is radioactive, decays to an excited state of neon 22 with a half-life of 2.6 years. Such a decay is thought to be common in the aftermath of a nova. In another event an energetic subatomic particle collides with a neon-22 nucleus in its ground state and excites it to a higher energy level. In a third event the capture of a neutron by a neon-21 nucleus leaves an excited nucleus of neon 22. In each case the subsequent—indeed, the almost immediate—decay of the excited nucleus to its ground state is accompanied by the emission of a photon of characteristic gamma-ray energy.

Gamma-ray lines can also result from processes that do not involve atomic nuclei. One such process is the encounter between an electron and a positron, or antielectron. The two have identical properties except for their electric charge, which is negative for the electron and positive for the positron. When an electron and a positron meet, they annihilate each other. Their mass is liberated as electromagnetic energy, which radiates away from the site of the annihilation, usually in the form of two gamma-ray photons, each with an energy of .511 MeV. Electron-positron annihilation appears to be a particularly common process near neutron stars and at the center of our galaxy.

A second non-nuclear process that gives rise to gamma-ray lines is electron cyclotron emission. Electrons moving in a magnetic field are subjected to a circular acceleration that makes them radiate photons at a characteristic energy. The energy is proportional to the strength of the field, and so if the field is sufficiently intense and uniform, as it may be at the magnetic poles of a neutron star, the electrons radiate gamma rays.

The first gamma-ray lines unambiguously detected from space were produced by several of these mechanisms. They emanated from a giant solar flare that occurred in August, 1972, and they were detected by Edward L. Chupp and his colleagues at the University of New Hampshire, who had placed a small gamma-ray detector aboard the satellite *OSO-7* (the seventh Orbiting Solar Observatory of the National Aeronautics and Space Administration). The lines included one at .511 MeV, which originated with the annihilation in the solar atmosphere of positrons produced in the solar flare. Another line, at 2.2 MeV, originated with the capture of neutrons by protons and the emission of a gamma ray from each resulting deuterium nucleus. Finally, two feeble lines at 4.4 and 6.1 MeV originated with the decay to their ground states of carbon-12 and oxygen-16 nuclei that had been raised to excited energy levels in the solar atmosphere by collisions with protons accelerated in the flare.

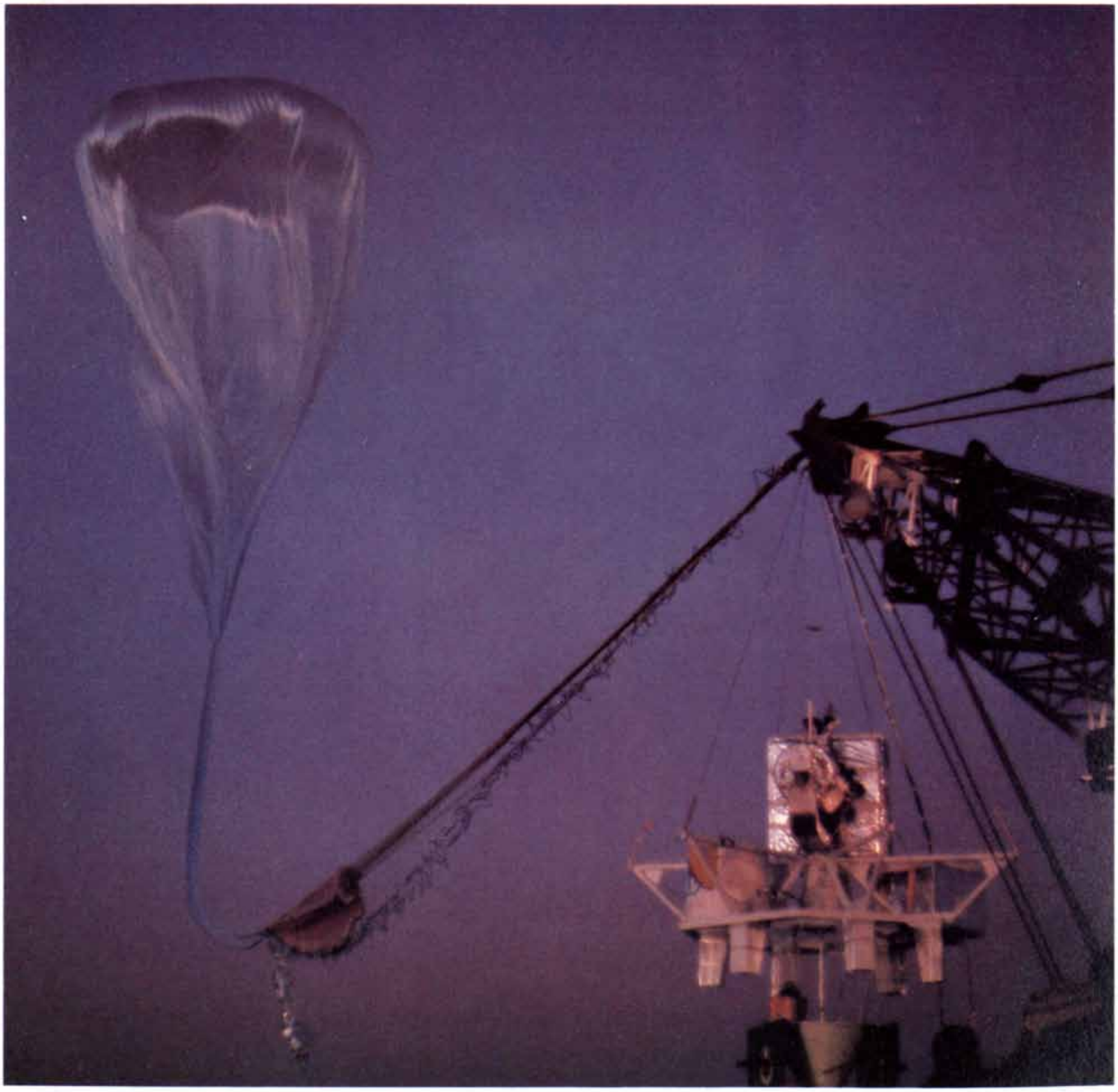
In the eight years that have passed since those first observations were made several research groups, including our own at Bell Laboratories and the Sandia Laboratories, have been flying gamma-ray telescopes, mostly on balloons but sometimes in satellites, in attempts to raise the instruments above nearly all of the earth's atmosphere and detect gamma-ray lines of astrophysical origin. The field is still in its infancy. Here we shall review some of its prospects, describe a gamma-ray-line telescope and give some of the early results.

Perhaps the most important prospect is that of making observations that bear on the cosmic synthesis of the chemical elements. It is hypothesized that the big bang created principally hydrogen and helium, which are converted into heavier elements by nuclear fusion in the interior of stars. For nuclei heavier than that of iron 56, however, fusion does not release energy. A mechanism other than fusion in stars must therefore be invoked to account for the creation of these elements. Specifically, it is hypothesized that the intermediate and heaviest elements are produced largely in the explosion of stars.

The details are as follows. A star is a sphere of gas in which the heat produced by nuclear fusion counters the tendency of the gas to fall inward. Eventually, however, the star exhausts its nuclear fuel. A well-accepted hypothesis maintains that if the mass of the star is comparable to the mass of the sun, the star contracts gradually until it has approximately the radius of the earth. It is now a white dwarf; it has exhausted its nuclear fuel and is cooling. If the mass is substantially larger, the star does not contract gradually but collapses catastrophically and explodes. It is estimated that once every 10 to 300 years one star out of the hundreds of billions in a typical galaxy reaches this end point in stellar evolution, and the light of the explosion can outshine the entire galaxy for a period of days or even weeks. It is this event that is called a supernova. It leaves an expanding scatter of debris, of which the best-known example in our own galaxy is the Crab Nebula in Taurus.

The theoretical astrophysicists W. David Arnett, Donald D. Clayton and Stirling A. Colgate were among the first to devise computer models of the nuclear reactions that occur in those few explosive seconds, when the processes of stellar nucleosynthesis, which up to then may have lasted for billions of years, take entirely new paths. Such models suggest that the explosion tends to make heavier nuclei out of lighter ones. In particular it does seem possible that supernovas have made most of the intermediate and heavy elements in the relative cosmic abundances that are observed.

It would of course be well to put these theories to a test, and in principle it can be done. Spectral lines at light and X-ray energies reveal which atoms are present in the debris of a supernova explosion. The atoms emitting the radiation, how-

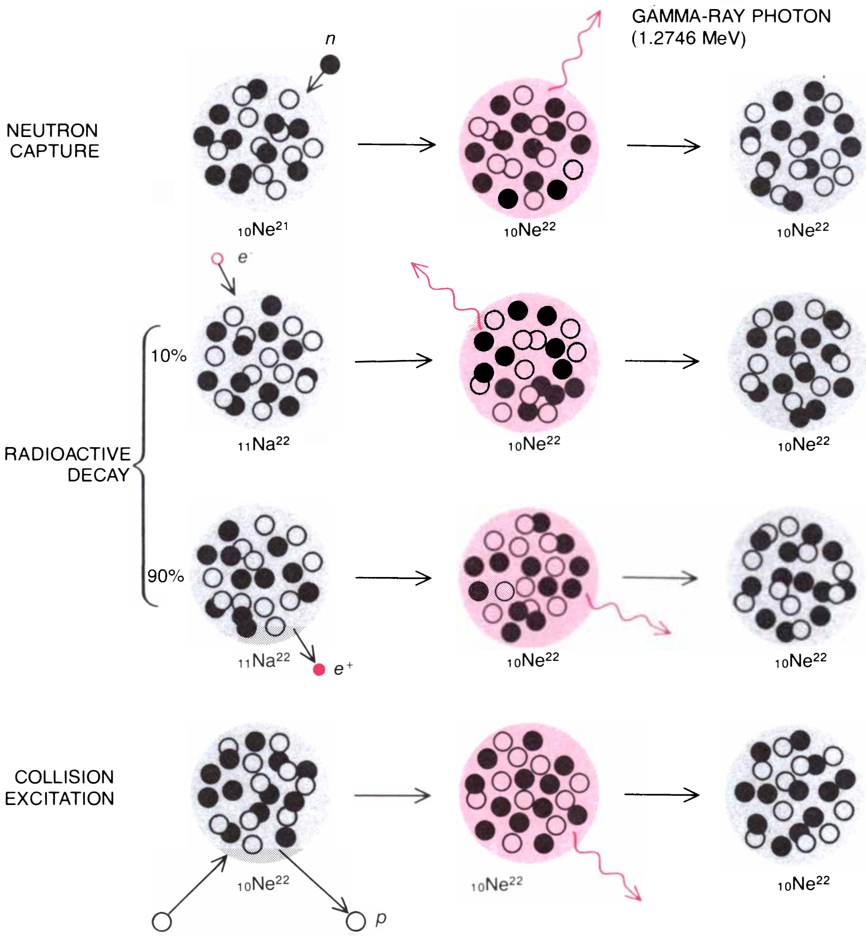


BALLOON LAUNCH at dawn on October 9, 1975, sent the authors' gamma-ray-line telescope aloft on a test flight above Alamogordo, N.M. The balloon is attached to a parachute, which is attached to the crane at the right, from which a payload is suspended. The base of the payload includes a set of crash pads. The cone under the pads is ballast. The box at the top of the payload contains electronics. In front of the box is a drum of liquid nitrogen for use as a coolant. The telescope is between the two and hence cannot be seen. At the end of the flight the parachute with its payload is freed from the balloon, and the balloon is torn open so that it too returns to the earth.

ever, may have been part of the star before it exploded, and some of them may even have been swept up from interstellar space. On the other hand, many of the nuclei synthesized in the explosion should be radioactive, and the decay of such nuclei is accompanied by the emission of gamma rays at signature energies. (A few of the signature emissions are at somewhat lower energies and correspond to high-energy X rays.) The detection of even a single gamma-ray line from the site of a supernova explosion would therefore represent a major advance in our understanding of how the present universe was made.

A second type of stellar explosion is the nova. Novas are much less violent than supernovas but much more frequent; they flare up perhaps 25 times per year per galaxy. They too may make an important contribution to nucleosynthesis. In the currently accepted hypothesis a nova explosion occurs in a binary stellar system where one star is an ordinary one and the other is a white dwarf, a star that as we have seen has exhausted its nuclear fuel without exploding. Although a white dwarf has no energy source of its own, it is thought that when such a star is a member of a binary system, it slowly accretes matter, largely hydrogen, from its companion star. Eventually the temperature and pressure at the base of this accreting gaseous envelope reach explosive values and the envelope blows off.

The calculations done for nucleosynthesis in nova explosions by groups led by Clayton and by James W. Truran, Jr., suggest a large production of sodium 22, almost all of which decays to the excited state of neon 22 with a half-life of 2.6 years. Several months after the explosion the most prominent gamma-ray line from the site of the explosion should therefore be a line at 1.2746 MeV, which represents the decay of the freshly synthesized neon nuclei to their ground state. Several months should be adequate for astronomers to get a balloon-borne gamma-ray telescope aloft. Since novas are frequent and one of them is therefore likely to be relatively close to the solar system, the 1.2746-MeV line may well be the first gamma-ray line detected from an exploding star.



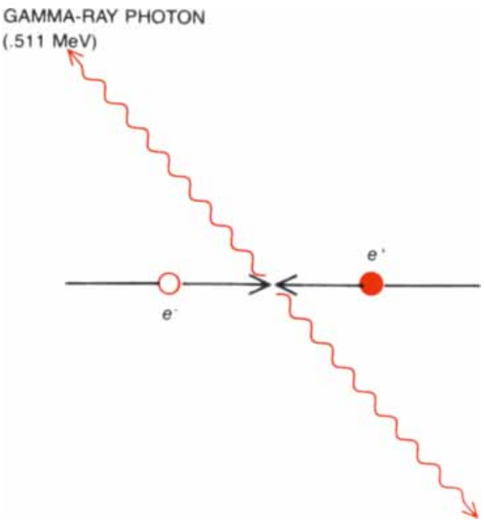
ATOMIC NUCLEUS EMITS A GAMMA RAY when it decays from an excited energy level to its ground state, or level of least possible energy. Here the nucleus of neon 22 is used to diagram three ways the emission can occur. In neutron capture (*top line*) neon 21 becomes neon 22 in an excited state by the addition of a neutron (*n*). In radioactive decay (*second and third lines*) sodium 22 becomes neon 22 in an excited state by the addition of an electron (e^-), usually the capture of an electron from its orbit in the atom, or the emission of a positron (e^+), which is the antiparticle of an electron. The latter possibility is the commoner of the two, but either converts a proton into a neutron. In collision excitation (*bottom line*) the energy imparted to neon 22 by an energetic particle (here a proton, *p*) raises it to an excited state. In each case the decay of the excited nucleus then releases a photon (a quantum of electromagnetic radiation) with a signature, or characteristic, energy of precisely 1.2746 million electron volts (MeV).

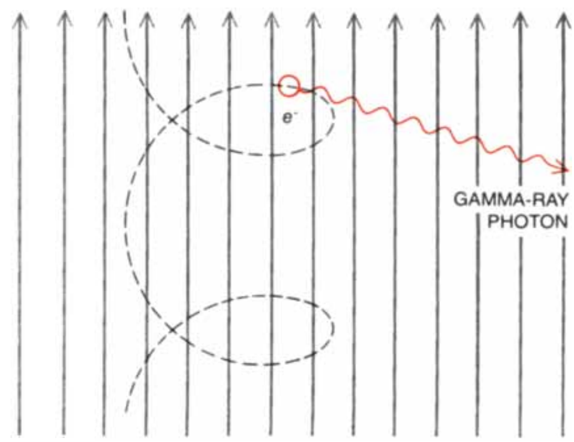
The same supernova explosions that may contribute to making the intermediate and heavy elements sometimes leave at the site of the explosion a stellar remnant that has approximately the mass of the sun compressed into a radius of approximately 10 kilometers. The density at the center of the remnant is calculated to be some 10^{15} grams per cubic centimeter, a value that exceeds the density inside the atomic nucleus. Indeed, the density is so great that the electrons of the atoms are thought to have been squeezed into the protons, canceling the protons' positive charge and making them neutrons. The remnant is known as a neutron star.

The youth of such an object is different from its old age, as far as either is known from inferences made by astrophysicists on the earth. Young neutron stars that are spinning rapidly are radio pulsars: sources of pulsating electromagnetic radiation at radio frequencies, first detected in the 1960's. The source of the energy for the pulsar emissions appears to be a steady loss in the neutron star's rotational kinetic energy, because in every known pulsar the period between successive pulses is observed to be slowly increasing.

The actual mechanism by which the rotational energy is converted into the observed radiation remains a mystery. There are many competing theories. In general, however, the neutron star is posited to have a magnetic field locked into it because the field of the parent star is thought to be conserved throughout the events that create the stellar remnant. Since the remnant is compressed, so too is the magnetic field, which thereby attains an enormous strength: it is estimated to be 10^{13} gauss. (The strength of the earth's magnetic field is a fraction of a gauss.) The remnant star is spinning, and so the magnetic field spins too. The result is that the magnetic field induces an electric field at the surface of the star. The electric field pulls charged particles from the surface and accelerates them to high energies. The trajectories of the particles are then bent by the magnetic field. The bending of the trajectories makes the particles radiate photons.

The pulsing of the radiation observed from the earth is explained by assuming that the two poles of the magnetic field are not aligned with the neutron star's axis of rotation. Hence as the neutron star rotates, each pole of the field, where the field strength is particularly intense, alternately appears and disappears from





TWO NON-NUCLEAR PROCESSES that generate gamma-ray lines are diagrammed. In electron-positron annihilation (*left*) a positron meets an electron, its antiparticle, and the mass of both particles is converted into electromagnetic energy, which takes the form of two gamma-ray photons, each bearing .511 MeV. In electron cyclotron emission (*right*) an intense and uniform magnetic field (represented by dense and parallel field lines) adds a circular motion to the trajectory of an electron. The frequency of the circling corresponds to the frequency (and hence the energy) of each photon that the electron emits. If the frequency is high enough, the photons are gamma rays.

any given direction of view. The radiation from it is thus detected in pulses, in what is called the lighthouse effect.

Why is all of this important for gamma-ray-line astronomy? One reason was first pointed out by Peter A. Sturrock of Stanford University. In most of the accepted theoretical models large fluxes of positrons are generated in the magnetic field of the neutron star. They are produced in a curious way. The magnetic field is so strong that when it bends the trajectory of a high-energy charged particle, the particle emits gamma-ray photons. If the energy of such a photon is more than 1.022 MeV, then when the photon encounters the magnetic field it can give rise to an electron-positron pair. (The energy of the photon must be more than 1.022 MeV because the mass of an electron or a positron corresponds to .511 MeV by application of the formula $E = mc^2$.) When a positron created in this way encounters another electron, that pair is in turn annihilated, giving rise to gamma-ray photons that contribute a line at .511 MeV. The detection of that sharply tuned radiation should enable astronomers to place limits on the production of positrons by pulsars and hence to apply an important constraint to allowable pulsar models.

In a million years or so even the most rapidly spinning neutron star slows down. Its life as a radio pulsar is over but its old age can still be interesting. Hundreds of pointlike X-ray sources have been discovered in our galaxy. Most are thought to be old neutron stars that are accreting matter from companion stars in a binary system. As the accreting matter swirls into the gravitational field of the neutron star it is accelerated to high energies and forms a hot accretion disk. The collisions among the particles in the disk generate the broad spectrum of X rays that is detected from the earth.

The accreting matter is thought to be highly ionized and so to consist mostly of energetic protons and electrons. It might therefore be possible to detect gamma-ray lines also, because energetic protons that impinge on the surface of the neutron star would excite its nuclei, and the decay of the nuclei to their ground state would generate gamma-ray photons. From the gamma-ray signature energies the nature of the surface of an old neutron star could be inferred.

A well-established cornerstone of relativity theory holds that a photon loses energy as it climbs out of a gravitational field, and that its wavelength is accordingly increased; in the case of light the photon gets redder. The shift is proportional to the change in gravitational potential energy, which for escape from the surface of a star is proportional to the mass of the star divided by its radius. Since the mass of a neutron star is large and its radius is small, its gravitational potential is immense, and so is the loss in energy of the photons that escape from its surface. It is therefore expected that all the gamma-ray lines would be shifted to lower energies (and longer wavelengths) by large amounts. The surface red-shift parameter is designated Z_s. For any given line it is the difference between the observed energy per photon and the energy per photon that would be detected if the radiation had been emitted on the earth, expressed as a fraction of the observed energy per photon.

A second line of argument now comes in. It was developed independently by Kenneth Brecher and by Richard Bowers. Several known neutron stars have orbits in a binary system that cause them to be eclipsed periodically by their companion star. For each of these neutron stars the interruptions in the X-radiation detected on the earth reveal the orbital period and the eclipse time. Meanwhile the orbital motion of the companion star brings it alternately toward the earth and away from it, so that its light emission is Doppler-shifted. In addition the intensity of the light is modulated because the companion star changes shape as it responds to tidal forces in the neutron star's gravitational field. Taken together, these various kinds of evidence yield the parameters of the neutron star's orbit, and when the orbit is known, the mass of the star can be calculated.

What next affects the argument is that neutron stars may in fact consist largely of subatomic particles more exotic than neutrons. Many different possibilities have been detailed in the scientific literature, including even quarks, and several theorists have pointed out that different compositions lead to different radii (R) for any given mass (M). For any neutron star a determination of M and a measurement of Z_s, which is proportional to M/R, will yield a determination of R. That will severely constrain the proposals concerning the star's composition. Gamma-ray-line astronomy therefore leads directly to the physics of superdense matter.

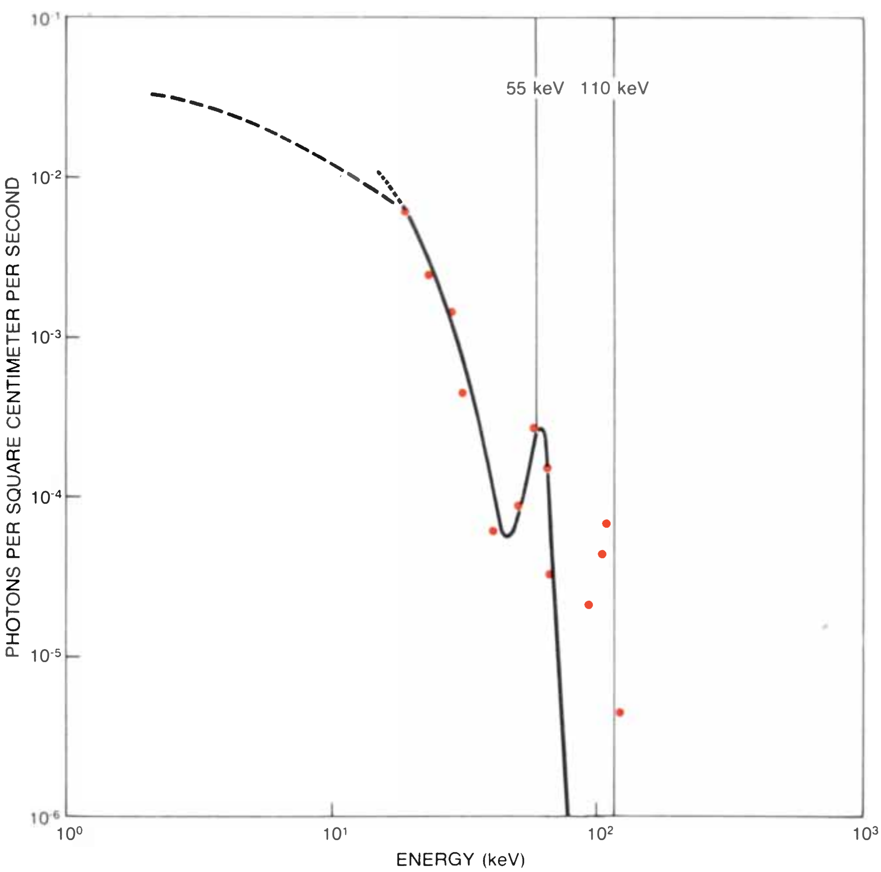
In this connection it may be significant that a new class of astronomical objects has recently been identified. The objects emit enormous fluxes of gamma radiation for periods of seconds or minutes and then the emission stops. Not much is known about them yet. Indeed, it is not yet known whether they are in our galaxy or outside it. The most credible speculation is that the gamma-ray bursts are caused by energetic events on or near the surface of neutron stars. The bursts could occur, for example, if the accretion of matter onto a neutron star from its companion in a binary star system occurs episodically.

On June 10, 1974, a group led by Allan S. Jacobson of the Jet Propulsion Laboratory of the California Institute of Technology commanded a wide-field balloon-borne gamma-ray-line telescope to point in the direction opposite to that of the center of the galaxy and detected a burst of gamma rays that lasted for 20 minutes. The group interprets the burst as being of astrophysical origin. Some linelike features seem to lie in the spectrum of the burst. It has been proposed that they represent several common mechanisms that generate gamma radiation, with the qualification that the lines are red-shifted by the same proportion to lower energies. For example, a linelike feature near .4 MeV in the spectrum of the burst is interpreted as being a .511-MeV positron-annihilation line that has been red-shifted by a Z of .28. One would be hard pressed to name for that region of the electromagnetic spectrum another common nuclear transition that might account for the line. On the other hand, a value of .28 for the surface red-shift parameter of a neutron star agrees well with theoretical calculations.

An independent observation by our own group later revealed a possible line at .4 MeV when our balloon-borne instrument was pointed in the same direction. The instrument's field of view was large, but it included the Crab Nebula and the neutron star at its center. More recently Bonnard J. Teegarden and Thomas L. Cline of the NASA Goddard Space Flight Center have reported the first measurement of the spectrum of a gamma-ray burst by a high-resolution satellite-borne instrument. They too found some evidence for gravitationally red-shifted gamma-ray lines. It seems quite possible that the first gravitational red-shift measurements from a neutron star are now in hand.

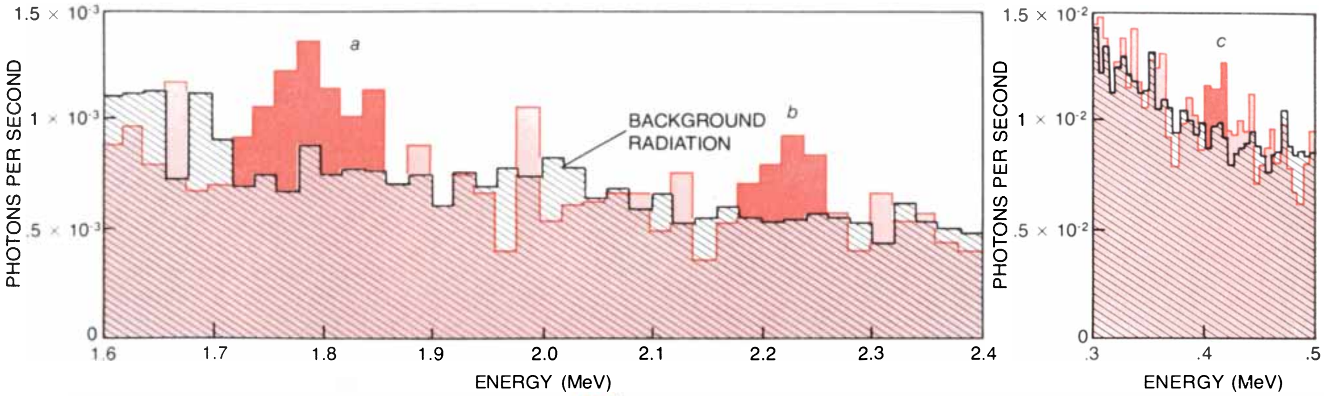
The electrons impinging on the surface of a neutron star present the possibility of a further observation because they have a circular cyclotron motion in the magnetic field of the star and therefore radiate photons. Where the field is huge and uniform near the two magnetic poles of the star the cyclotron emission might consist of gamma-ray lines. It is a possibility first proposed by the Russian astrophysicist Rashid A. Sunyaev and his colleagues. The detection of such lines would amount to a measurement of the field strength. It would also be proof that the enormous magnetic fields are really there.

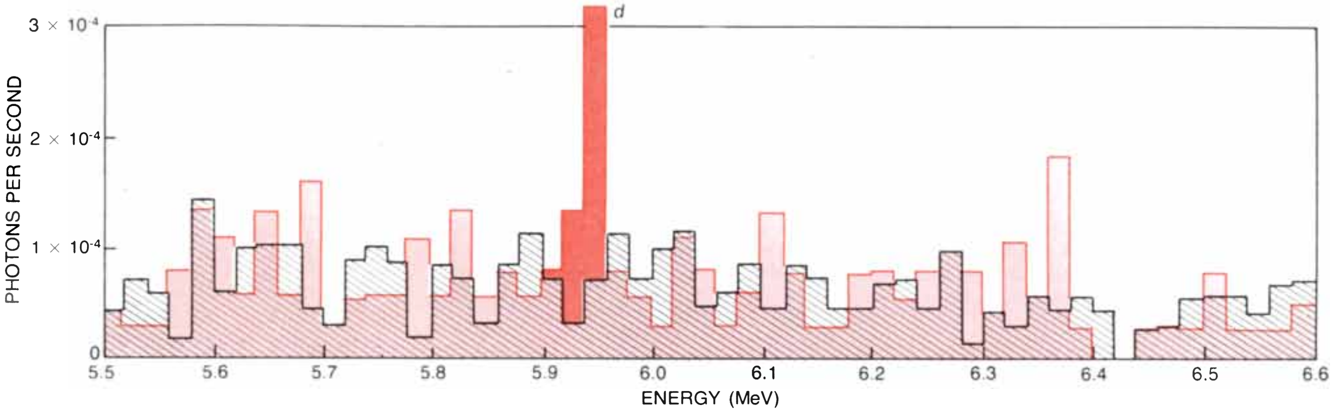
In May, 1976, a group from the Max Planck Institute for Physics and Astrophysics at Garching in West Germany, working in collaboration with a group from the University of Tübingen, with both groups led by Joachim Trümper, sent a balloon aloft above Palestine, Tex. It bore an instrument with which they obtained a spectrum of the low-energy gamma radiation emitted by the slowly spinning X-ray pulsar Hercules X-1. The photons at very low gamma-ray energies formed a continuous spectrum, not a line one, which the investigators interpreted as being the upper end of the X-ray spectrum emitted by an accretion disk surrounding the pulsar. A linelike feature appeared, however, at an energy of 55 keV (55,000 electron volts). A second linelike feature lay at 110 keV. Both of them arrived at the detector in pulses at intervals of 1.24 seconds, which is the period of the pulsar. It is believed they are an electron-cyclotron line, together with a harmonic at twice the cyclotron frequency. The existence of the 55-keV feature was confirmed on a second balloon flight and again last year by an instrument aboard the NASA satellite *HEAO-1* (the first High Energy Astronomical Observatory). If it is really a cyclotron line, the observation by Trümper's groups makes it possible for the first time to measure a neutron star's magnetic field. On this basis the strength of the field of Hercules X-1 is approximately 5.3×10^{12} gauss.



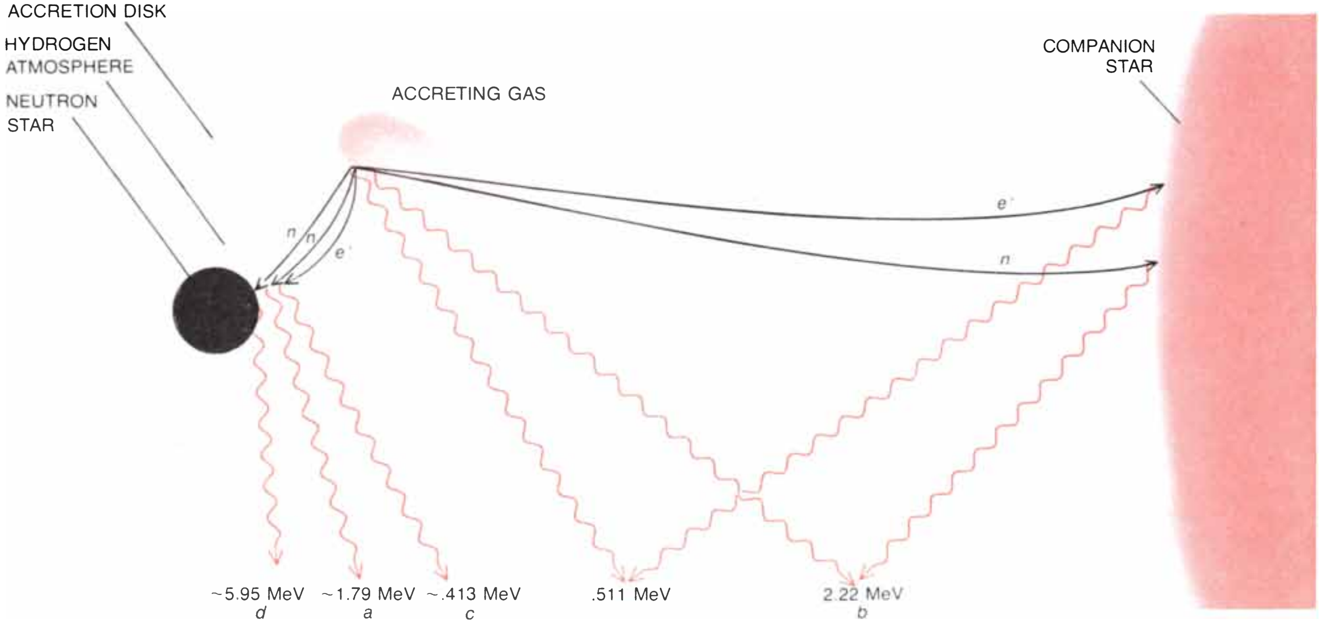
GAMMA-RAY SPECTRUM OF HERCULES X-1, a prominent X-ray source, includes two gamma-ray peaks, or linelike features, at 55 keV (55,000 electron volts) and 110 keV. The line at 55 keV is interpreted as the cyclotron radiation emitted by electrons in the intense magnetic field of a neutron star: the superdense remnant of a stellar explosion. The line at 110 keV is interpreted as a harmonic at twice the cyclotron-emission energy. The data were recorded on May 3, 1976, by a balloon-borne instrument launched near Palestine, Tex., by a German group led by Joachim Trümper. The rest of the spectrum, including data at the left (*broken line*), obtained by a satellite-borne instrument, is thought to be the upper end of a continuous spectrum of X rays emitted by atoms colliding in a disk of matter that surrounds the neutron star.

A diffuse emission of gamma-ray lines permeates our galaxy and therefore is expected to emanate from the plane of the Milky Way. Its observation will require telescopes more sensitive than those in service today. It arises because cosmic-ray particles (mostly energetic nuclei) collide with the dust grains, atoms and molecules in interstellar space and raise their nuclei to excited energy levels. The decay of such nuclei then gives rise to the gamma rays. Since the dust grain, atom or molecule struck by an energetic particle recoils, the nucleus is in motion when the gamma-ray pho-

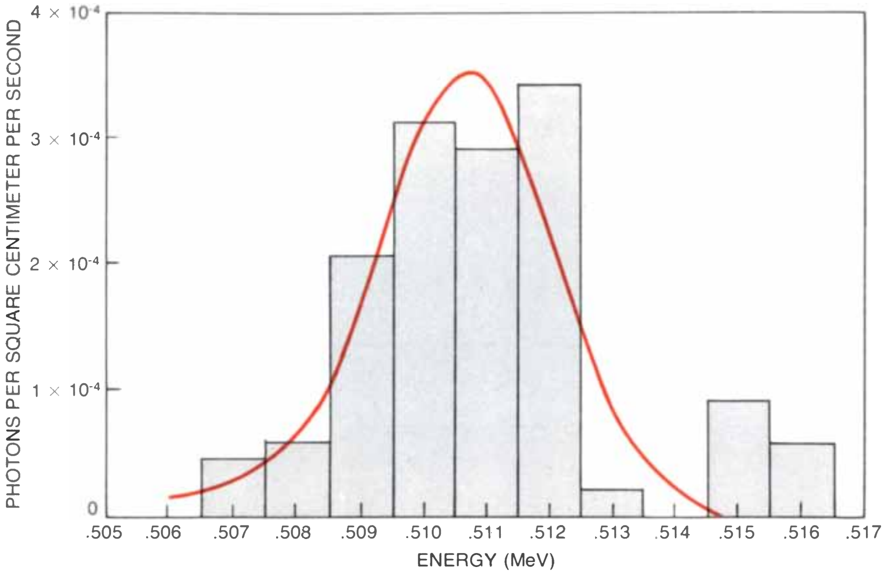




GAMMA-RAY BURST from a direction in the sky opposite to the direction of the center of our galaxy occurred on June 10, 1974, and lasted for 20 minutes. Its spectrum is shown here in three different energy regions. Four linelike features rose prominently above the background radiation when the telescope pointed at that part of the sky. Two of them, labeled *a* and *b*, may represent the creation of gamma-ray photons by the reaction that joins a neutron and a proton to form a deuterium nucleus. The offset of feature *a* to an energy less than that of *b* is explained by assuming that the photons contributing to *a* were produced near the surface of a neutron star and lost energy escaping from the star's powerful gravitational field. Feature *c* is attributed to electron-positron annihilation, with the energy of the photons also being less than its normal value: .511 MeV. Feature *d* is attributed to the capture of neutrons by the nuclei of iron 56, converting each of those nuclei into an excited state of iron 57, which decays to its ground state with the emission of a photon. Iron 56 is the stablest atomic nucleus, and so in many theoretical models it forms the surface of a neutron star. Again the energy of the photons is less than normal. The data were collected by a group led by Allan S. Jacobson at the Jet Propulsion Laboratory of the California Institute of Technology.



EPISODIC ACCRETION OF GAS onto a neutron star from its companion in a two-star system may have caused the gamma-ray burst whose spectrum is shown at the top. In this hypothesis the collisions of arriving gas with the gas in a disk that surrounds the neutron star give rise to neutrons and positrons. Subsequent events, including the capture of neutrons by hydrogen and by iron 56 and the mutual annihilation of positrons and electrons, then create the photons of which the burst consists. The photons emitted at and near the surface of the neutron star lose energy when they escape from the gravitational field of the star. Letters *a* through *d* identify the gamma-ray lines that appear in charts at the top. The illustration is adapted from one by Richard E. Lingenfelter, J. C. Higdon and Reuven Ramaty.



GAMMA RADIATION FROM GALACTIC CENTER (or at least a region in the direction of the galactic center) includes a linelike feature at .511 MeV. The strength of the line corresponds to the annihilation of 10^{43} positrons per second. The mechanism by which the positrons are produced is not yet known. The chart summarizes data amassed by the authors' gamma-ray-line telescope during a balloon flight above Alice Springs in Australia in November, 1977.

ton is emitted. The result is a Doppler shift in the energy of the photon. It is a shift to a higher value if the nucleus is moving toward the solar system and to a lower value if the nucleus is moving away. Hence for a group of nuclei that are moving in random directions the gamma-ray lines are broadened.

The broadening is proportional to the velocities of the nuclei. Thus it is hypothesized by Reuven Ramaty, Benzion Kozlovsky, Richard E. Lingenfelter and others that each gamma-ray line can be a combination of three distinct contributions. A sharp component of any given line should originate with the emission of photons by the nuclei of atoms in the interstellar dust grains, which are likely to lose their recoil velocity, and then emit the radiation, while they are still inside the grain. The sharp component of the lines may therefore offer a good opportunity to determine the composition and distribution of the interstellar dust.

A somewhat broader contribution to any given line should originate with the emission of photons by the nuclei of atoms that are not confined in dust grains. The broader component hence offers a good opportunity to determine the composition and distribution of the interstellar gas. To be sure, other kinds of astronomy can serve the same purpose. In the case of optical astronomy, however, interstellar dust limits observations toward the center of the galaxy to a distance of only from 10,000 to 13,000 light-years in a galaxy whose radius is more than 30,000 light-years. The interstellar medium is highly transparent to gamma rays. Moreover, light and X-ray photons are emitted by the electrons in atoms. Thus the pattern of the emission lines is complicated by the chemical state of the matter, such as the degree to which it is ionized. The fact that gamma-ray photons are emitted from the nuclei of atoms ensures that their spectral pattern is independent of the chemical state of the matter.

Finally, a broad contribution to any given line should originate with the emission of photons by the cosmic-ray nuclei themselves, whose velocities are a substantial fraction of the speed of light. The broad component therefore offers an opportunity to study cosmic rays, whose collisions with interstellar matter are thought to be responsible for heated regions in the interstellar gas. It also is suspected that cosmic-ray collisions splitting the nuclei of interstellar atoms contribute to the creation of light elements such as lithium and beryllium. (One can postulate fusion reactions that create those elements in the interior of stars, but other reactions consume them.) The problem with detecting the cosmic rays themselves as they arrive at the earth is that the rays interact with magnetic fields in the solar system, and so the sample arriving at the earth may not be representative of the cosmic rays in interstellar space. The detection of their characteristic gamma-ray-line emission would eliminate the problem.

Still other objects of importance to gamma-ray-line astronomy are entire galaxies or their central regions. Quasars, for example, are thought to be young galaxies that are immensely distant from the earth and yet are remarkably luminous. Two classes of models are invoked in explanation. The models in the first class propose that large numbers of supernova explosions occur at the center of quasars. The models in the second propose that stars at the center of quasars are torn apart by the gravitational field of a huge black hole and that the stellar matter forms a hot accretion disk around it. In the second case one would hope to observe the gamma-ray-line emission from collisions in the disk, with the lines greatly red-shifted and also broadened, each from a line to a spectral smear, by the black hole's gravitational field.

Even the center of our own galaxy is a mysterious region, partly because optical telescopes cannot reveal its structure. In November, 1977, and again in April, 1979, we flew our instrument on balloons over Alice Springs in Australia in an effort to detect gamma-ray lines from the direction of the galactic center. In earlier work a group at Rice University led by Robert C. Haymes had found tantalizing evidence for line emission near .5 MeV. In our 1977 flight we detected a line precisely at .511 MeV. It lay in the midst of a gamma-ray continuum. We confirmed the line's existence in 1979. It corresponds to the annihilation of 10^{43} positrons per second at or near the galactic center. The existence of such a line from the galactic center was postulated two decades ago by Philip Morrison of the Massachusetts Institute of Technology, and later by Floyd W. Stecker of the Goddard Space Flight Center and by one of us (Leventhal).

The scientific literature abounds with suggestions as to how the positrons are made. In what is perhaps the leading contender among the various hypotheses the positrons are said to be a by-product of supernovas. The idea is that the decay of radioactive nuclei produced in the stellar explosions releases the positrons, which speed away from the nuclei with energies of some hundreds of thousands of electron volts. Thousands of years may pass before a positron is annihilated in an encounter with an electron. Hence the line observed from the earth may represent supernova explosions occurring over long periods of time. The fact that stars are plentiful in the galactic center counts in favor of the hypothesis.

A second possibility is that cosmic rays collide with interstellar matter and split its nuclei into radioactive nuclei that emit positrons as they decay. The collisions can also yield pi mesons, whose decay gives rise to positrons. Recent calculations indicate, however, that the current values of the flux of cosmic rays and of the density of interstellar matter fail by two orders of magnitude to account for the strength of the .511-MeV line we detected. That strength is great: it is as if the luminosity of the sun were multiplied by 10,000 and concentrated at a single electromagnetic frequency.

A third possibility is that the galactic center harbors a large number of radio pulsars beaming large fluxes of positrons into interstellar space. Some additional possibilities are at present less credible. One hypothesis proposes the existence of a large black hole at the center of the galaxy, with an accretion disk around it. Another proposes that small black holes were created in great numbers in the early universe and that by quantum-mechanical processes they are, so to speak, evaporating: they are radiating their energy largely in the form of electron-positron pairs. Perhaps the least plausible hypothesis is that in the center of our galaxy is a large bulk of antimatter.

The important point, however, is that the mechanism by which positrons are generated at the center of the galaxy can be determined by observations that can be made in the coming decade, perhaps even in the next five years. It will be determined by means of detailed maps of the emitting region. If the region is a point source, then a black hole is the likely emitter. If peaks appear near radio pulsars, then those are the likely emitters. In addition many of the hypotheses require the existence of lines besides the one at .511 MeV, so that the detection of such lines will limit the possibilities.

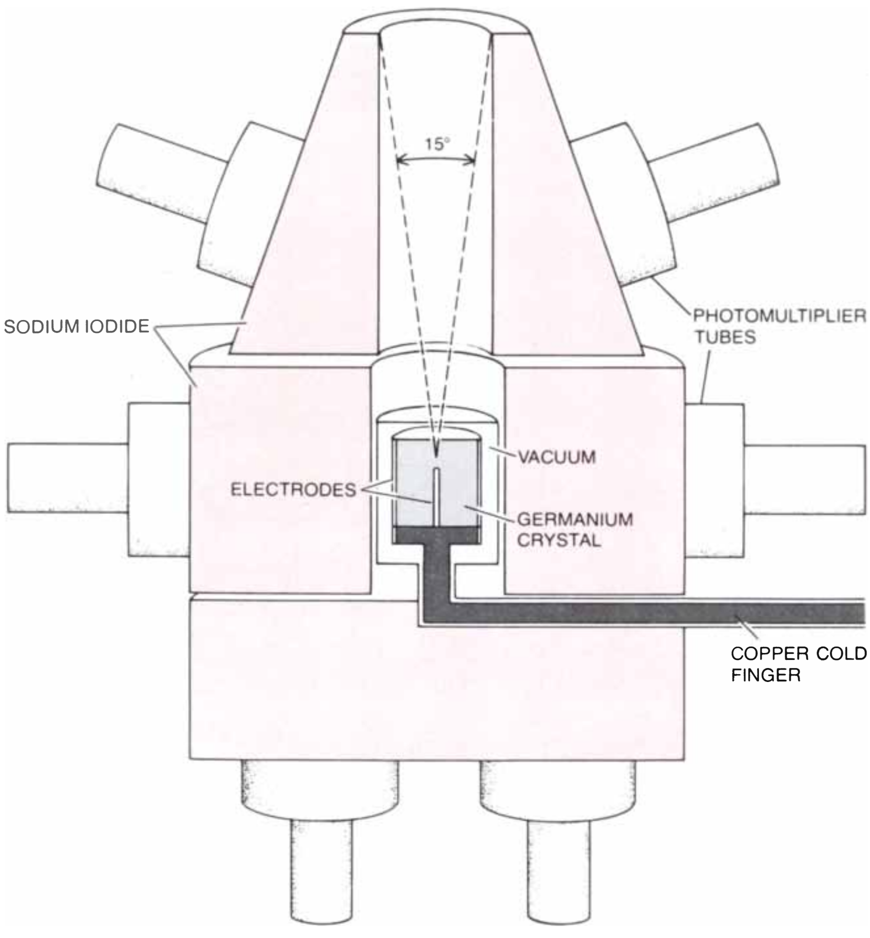
The crucial fact about a gamma-ray-line telescope is that photons with gamma-ray energies cannot be made to reflect from mirrors. They penetrate matter and tend to lose some of their energy in being scattered, or deflected, in encounters with electrons. This kind of scattering dictates the design of the telescope. At the center of the telescope flown by us is a germanium crystal with a volume of 130 cubic centimeters. Inside it an entering gamma-ray photon typically scatters off several electrons in succession. The recoil of the electrons generates a tiny current whose amplitude is proportional to the energy of the gamma ray. The gamma-ray photons can therefore be detected one at a time.

Surrounding the central detector is 400 pounds of sodium iodide with tellurium atoms introduced as a dopant. This too is a material that is sensitive to gamma rays; the collision of a gamma-ray photon with an electron causes the electron to recoil and then collide with many dopant atoms, and the atoms in turn emit pulses of light in the form of photons with optical energies.

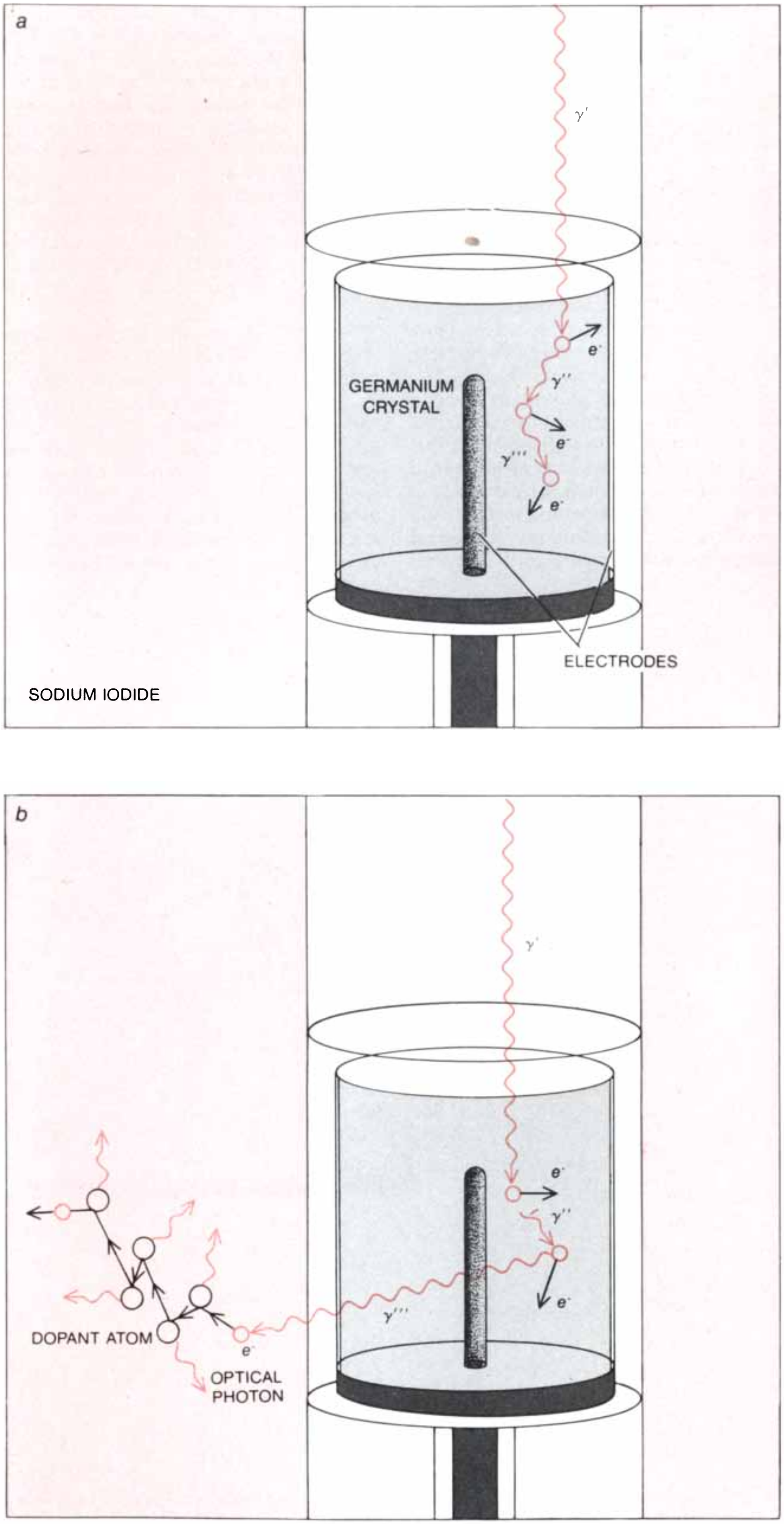
The energy resolution of the germanium crystal at 1 MeV is approximately 2 keV. This means that 1-MeV photons can be detected as a spectral line that has a width of 2 keV. The resolution of the sodium iodide detector is poorer by a factor of approximately 30. The sodium iodide, however, serves three important functions. First, it leaves open a channel to the central detector that defines the telescope's field of view. Second, it shields the central detector from a strong background of gamma-ray photons generated in the earth's atmosphere by cosmic rays arriving from space. Third, the pulses of light are amplified by photomultiplier tubes and are then used to veto the detection of those gamma-ray photons that scatter out of the germanium crystal and into the sodium iodide. Such a photon leaves some of its energy in both materials, so that the current in the central detector is not proportional to the true energy of the ray.

In the course of a balloon flight lasting for one to two days our telescope is commanded by a ground-based telemetry station to point in a certain direction. If the telescope is tracking an emitter of gamma radiation, each square centimeter of the central detector's surface may receive a single gamma-ray photon of astrophysical origin once every several minutes. The chances are better than even, however, that the photon will scatter out of the central crystal. The detection of the photon will therefore be vetoed by the almost simultaneous production of light in the surrounding shield. Slowly, then, the central detector accumulates a spectrum of gamma-ray counts plotted against photon energy. At the end of the flight the telescope is cut free from the balloon at an altitude of some 40 kilometers and is returned to the ground by parachute.

It appears to us that balloon-borne experiments will continue to be important in gamma-ray-line astronomy. For one thing, large fluxes of energetic particles are present in radiation belts at the altitudes to which satellites are launched. The particles generate a strong gamma-ray background. Moreover, balloon-borne experiments are less expensive and quicker to do than satellite-borne experiments. From the



GAMMA-RAY-LINE TELESCOPE flown by the authors and their colleagues consists of a detector inside a detector. At the center is a germanium crystal, in which an arriving gamma-ray photon is scattered by electrons. The recoil of the electrons gives rise to a pulse of electric current. The germanium is cooled by a copper "cold finger," which in turn is cooled by liquid nitrogen; otherwise a current that flows spontaneously would swamp the current generated by a gamma ray. Surrounding the central crystal are blocks of sodium iodide doped with a small impurity of tellurium atoms. Here an arriving gamma-ray photon is absorbed by an electron, which collides with atoms of the dopant. Light emitted by the atoms is amplified by photomultiplier tubes. Details of the detection of a gamma ray are in the illustration on the next page.



GAMMA-RAY PHOTONS ARE DETECTED one at a time by the authors' gamma-ray-line telescope. In *a* an arriving photon (γ') is scattered repeatedly but remains in the central germanium crystal. The motion of the electrons off which it scatters gives rise to an electric current that accurately represents the energy of the photon. In *b* an arriving photon is scattered into the surrounding sodium iodide, where it causes emission of light. The photon leaves only part of its energy in the central crystal; hence matters are arranged so that its detection is vetoed by a signal resulting from the flash of light in the sodium iodide. Event *b* is commoner than *a*.

moment it is proposed to the moment it starts to generate data a satellite-borne experiment takes from five to 10 years and costs tens of millions of dollars. A balloon-borne experiment may take less than two years and cost less than $1 million. Finally, balloons are an excellent testing ground for future satellite instruments. On the other hand, satellites stay aloft for long periods of time and accumulate a large body of data.

A network of satellites has now been established that is capable of precisely determining the direction of gamma-ray bursts by triangulation. On March 5, 1979, the most energetic gamma-ray burst so far was detected by the network and localized in the center of the debris of a supernova explosion in the Large Cloud of Magellan, a satellite galaxy of our own. A spectrum of the burst was obtained by E. P. Mazets and his colleagues at the A. F. Ioffe Physical-Technical Institute in Leningrad. It includes a linelike feature at .4 MeV: the energy of a red-shifted positron-annihilation line. The location of the burst and the detection of the line strongly suggest that what underlies gamma-ray bursts are indeed neutron stars.

Last fall NASA launched the satellite *HEAO-3*. It carries a gamma-ray telescope consisting of four germanium detectors in a cesium iodide shield. The instrument, which was built by a group at the Jet Propulsion Laboratory led by Jacobson, is designed to operate for a year and scan most of the sky for sharp gamma-ray lines with a sensitivity as great as 10^{-4} photons per second per square centimeter of detector surface. The instrument is functioning properly, but it is too early for any results to have been reported.

NASA is now planning to launch the Gamma-Ray Observatory satellite in the mid-1980's. The satellite will carry five gamma-ray-sensitive instruments, of which three will be broad-spectrum telescopes and two, a germanium-based telescope designed by Laurence E. Peterson's group at the University of California at San Diego and a sodium iodide–based telescope designed by James D. Kurfess' group at the Naval Research Laboratory, will search for gamma-ray lines. Peterson's group includes workers at Bell Laboratories, the Goddard Space Flight Center, the Jet Propulsion Laboratory, Sandia Laboratories and two French institutions: the Saclay Nuclear Research Center and the Center for Cosmic Ray Studies at Toulouse. Kurfess' group includes workers at Northwestern University, Rice University and the Royal Naval College in Britain. The instruments should scan most of the sky during the projected two-year life of the venture and achieve line sensitivities of 10^{-5} photons per second per square centimeter. With the launching of that observatory gamma-ray-line astronomy will have come of age.

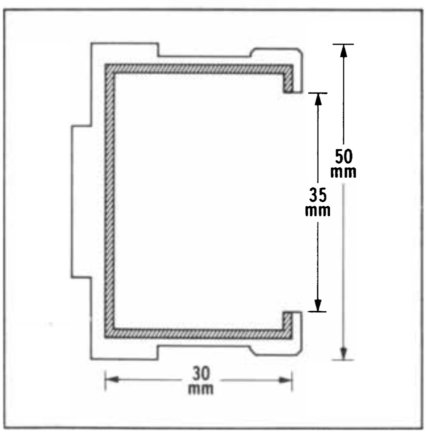
How to make a useful black hole.

No matter how black the blackest ink that can be laid down on a magazine page, it cannot come even close to the blackness of this thing. Anybody who knows how to make a blacker black ought to publish a paper on how to do it. Until that happens we know of nothing in the world that's blacker. Not only is it black in the visible and the UV—as black as the blackest known alternative in the visible—but it is blacker than anything else we have ever heard of out to 2.5 μm in the infrared, where few things are really very black.

If you need something that black, we tell you how to make it yourself. We sell the principal component, which happens to be the whitest known white. Needs for the whitest of whites and the blackest of blacks both occur in spectrophotometry and radiometry, among other endeavors. They serve as standards against which samples are measured. When measuring reflectance of a sample with considerable transmittance in the infrared, the departure from blackness in the backing can introduce significant error. The light trap pictured above shows the following reflectances:

λ in μm	0.35	0.5	1.0	1.5	2.0	2.5
reflectance	.0004	.0004	.0006	.00075	.001	.001

Extreme diffusivity, nonselectivity, stability, and ruggedness are, of course, the basis of success in achieving useful blackest blacks and whitest whites. KODAK White Reflectance Coating is applied by the KODAK Laboratory Sprayer for high-viscosity use to an aluminum disk. Result is a spectral reflectance from 0.2 μm to 2.5 μm higher than that of the old whitest white, magnesium oxide smoke. To make the new blackest black, we apply the KODAK White Reflectance Coating to the inside of a baffled light-trap and overspray when dry with ultraflat black enamel. The light-trap might be fabricated like this:



But you'll have to do it, not we. KODAK White Reflectance Coating is sold by lab supply houses handling KODAK Laboratory Chemicals as CAT No. 6080, the sprayer as CAT No. 13270.

Our interest in standards for optical radiation measurements runs deeper than promoting the sale of these two products, as you can see when you look at the title page of the new "Optical Radiation Measurements," Vol. 1, recently published by Academic Press, 111 Fifth Avenue, New York, N.Y. 10003.

1880 Kodak 1980

A 100-year start on tomorrow

SCIENCE AND THE CITIZEN

Identity Crisis

The neutrino is the nearest thing to nothing, being distinguished from the other elementary particles of matter mainly by what it does not have and what it does not do. The traditional theory of the neutrino describes it as a particle without mass and without electric charge. It interacts with matter only through the weakest of the basic forces of nature, and those interactions are so rare that a stream of neutrinos can pass through the earth with negligible attenuation. The earth moves in a thick soup of neutrinos (there are estimated to be about 100 per cubic centimeter throughout the volume of the universe), but they are such ephemeral particles that they almost never leave any trace of their passage.

This description of the neutrino may soon have to be revised in a small but crucial detail. Serious consideration is being given to the possibility that at least some types of neutrino have mass. The quantity of mass per particle would be minuscule, but the actual value is in some ways less important than the mere fact that it is different from zero. Since neutrinos are so numerous, even a small mass could drastically alter the large-scale structure of the universe.

The recent speculations on the neutrino mass have to do with the search for the bizarre phenomenon called neutrino oscillations. For almost 20 years it has been well established that there are at least two flavors, or kinds, of neutrino; one flavor can appear only in association with an electron and the other is always created together with a muon. More recently a third neutrino flavor has been added to accompany the newly discovered tau particle, which is a massive sibling of the electron and the muon. The flavors have seemed immutable, so that an electron-type neutrino could never be transformed into a muon-type neutrino. It is just such transformations that would be allowed by neutrino oscillations; indeed, a single neutrino would change its identity continuously and repeatedly as it moved along its trajectory.

In order to understand the mechanism of the hypothetical oscillations it is necessary to describe the neutrino as a wave. The intensity of the wave at any point gives the probability of finding the particle at that point. In the traditional theory each flavor of neutrino has its own independent wave, and there can be no mixing of the waves associated with the different neutrino types. The possibility of oscillations is introduced when a single neutrino is represented as a superposition of waves of all three types. The flavor of the neutrino at any point is then determined by the relative amplitudes of the three components. If all three components have the same amplitude, the neutrino is equally likely to take on any of the three flavors.

Even when a neutrino is represented as a superposition of waves, the identity of a given neutrino would still remain stable indefinitely as long as all three components moved with the same velocity. The relative phases of the three waves would then be the same at all points. Oscillations would arise only if the components moved with different speeds, so that the phases changed continuously and first one component dominated the amplitude and then another. The phenomenon is similar to the "beats" that can be heard when two sound waves of slightly different frequency are superposed.

According to the special theory of relativity, a particle whose mass is exactly zero must always move with the speed of light, whereas a massive particle can never quite attain the speed of light. It follows that neutrinos cannot oscillate if all of them are perfectly massless. Conversely, the observation of neutrino oscillations would imply that at least one neutrino type has a nonzero mass. Nothing about the oscillations would fix the actual value of the mass, but from the frequency or wavelength one could deduce the mass difference between the types.

The possibility of neutrino oscillations was first proposed in 1963 (only a year after the existence of distinct neutrino types was established) by the Japanese physicists M. Nakagawa, H. Okonogi, S. Sakata and A. Toyoda. Five years later a similar idea was proposed independently by Bruno Pontecorvo, an Italian-born physicist who lives in the U.S.S.R. For a decade the hypothesis remained an intriguing one, but no concerted effort to search for the oscillations was undertaken. Skepticism about the existence of oscillations was only one impediment; even if they did exist, it seemed doubtful that a practical experiment could detect them. A recent revival of interest stems largely from work proceeding on "grand unified" theories of the interactions between elementary particles. In these theories, which are still speculative, it becomes not only plausible but also likely that neutrinos have a small mass.

An observation that might represent evidence for neutrino oscillations has been reported by Frederick Reines, Henry W. Sobel and Elaine Pasierb of the University of California at Irvine. It was Reines, together with Clyde L. Cowan, Jr., who first detected the neutrino in the 1950's; the definitive experiment employed a nuclear reactor at the Savannah River Plant in South Carolina. In the search for neutrino oscillations Reines and his colleagues returned to the Savannah River installation. The recent experiment relied on a detector whose neutrino-sensitive medium is 268 kilograms of exceptionally pure deuterium oxide, or heavy water. Instruments for detecting the products of neutrino interactions are placed in the heavy water, and the entire apparatus is sheathed in lead and cadmium to screen out particles other than neutrinos.

The idea underlying the experiment is to simultaneously monitor two separate kinds of neutrino interaction. One process, called a neutral-current event, can be initiated by any flavor of neutrino; the other process, a charged-current event, can be instigated only by an electron-type neutrino (or, to be more precise, by an electron-type antineutrino). In a nuclear reactor neutrinos are generated by the beta decay of unstable nuclei among the fission products of uranium, and they are exclusively antineutrinos of the electron type. Hence in the absence of oscillations all the neutrinos reaching the detector from the reactor should be electron-type antineutrinos and the two measured processes should both indicate the same flux of particles. Oscillations would change the flavor of some of the neutrinos, with the result that the rate of charged-current events would be reduced. The neutral current events, however, being induced equally by all three neutrino types, would be unaffected by the oscillations.

The experiment has lasted for some three years, and data are still being recorded. Reines and his colleagues interpret these data as showing that the ratio of charged-current events to neutral-current events is less than half the expected ratio. If oscillations are the explanation of the discrepancy, they imply that the electron-type neutrino and the muon-type neutrino differ in mass (or in the energy equivalent of mass) by roughly one electron volt. An analysis of some older data, recorded at a different distance from the reactor core, is consistent with this result.

Reines emphasizes that there are a great many uncertainties associated with the experiment, and that it cannot stand as a discovery of neutrino oscillations unless it is supported by independent observations. The main difficulties can be traced to the abundance of neutrinos and to the rarity of their interactions. The reactor at Savannah River emits some 10^{13} neutrinos per square centimeter per second, but they are so reluctant to interact with matter that

only about 80 events per day resulting from the reactor flux were detected. This small signal had to be extracted from an extraneous background of about 400 events per day attributed to neutrinos from other sources.

A number of other recent experiments also bear on the question of the neutrino mass, and in sum they suggest the issue is far from being settled. Since last fall a group of physicists from the California Institute of Technology, the Laue-Langevin Institute in Grenoble and the Technical University of Munich has been measuring the spectrum of neutrinos emitted by a research reactor at the Laue-Langevin Institute. The characteristic signature by which the neutrino interactions are detected is somewhat different in this experiment, but the measurement should be sensitive to any vacillations in neutrino type. Felix H. Boehm of Cal Tech reports that no evidence of oscillations has been perceived.

Three quite different experiments that in retrospect seem to have a close connection to the reactor studies have been carried out at the European Organization for Nuclear Research (CERN) in Geneva. Called beam-dump experiments, they measured the relative numbers of electron-type and muon-type neutrinos presumably created in the decay of short-lived particles called *D* mesons. Earlier investigations of the *D* mesons had shown that the two neutrino types are produced with equal likelihood in these decays, but the beam-dump experiment registered twice as many muon-type neutrinos. This result came as a surprise when it was reported a year ago, and it remains somewhat enigmatic today. Neutrino oscillations are a possible explanation, but there are also several others.

Another puzzling result has recently been announced by E. T. Tretyakov and his colleagues at the Institute of Theoretical and Experimental Physics in Moscow. Tretyakov's experiment did not concern neutrino oscillations but looked instead at another phenomenon that could reveal evidence of a neutrino mass. The experiment measured the energy spectrum of neutrinos emitted in the decay of tritium, the radioactive isotope of hydrogen. If the electron-type neutrino had mass, this spectrum would have a deficiency at the highest energies observed; Tretyakov has reported finding such a deficiency. The result implies that the electron-type neutrino has a mass on the order of 10 electron volts. The details of the work, however, are not yet widely known among physicists outside the U.S.S.R.

Another search for oscillations has been carried out by Lawrence R. Sulak of Harvard University and his colleagues from Harvard and the Brookhaven National Laboratory. The neutrinos were obtained as secondary decay products from the beam of the proton synchrotron at Brookhaven, which was run at exceptionally low energy in order to ensure that only muon-type neutrinos would be created. Data were recorded for several months ending last November, but the analysis of the data is not yet complete. Sulak estimates that the experiment is 10 times more sensitive than earlier ones.

Ten electron volts represents a very small mass; the lightest particle known to have a nonzero mass is the electron, and its mass is 511,000 electron volts. Even the smallest measurable neutrino mass, however, would have far-reaching consequences not only for the physics of elementary particles but also for astrophysics and cosmology. For example, the oscillations (if they exist) might resolve a long-standing controversy over the nuclear chemistry of the sun. For more than a decade Raymond Davis, Jr., and his colleagues at Brookhaven have been operating a mammoth neutrino detector deep in a mine shaft in order to measure the flux of neutrinos from the sun. The observed flux has consistently fallen short of expectations by a factor of three. The detector in the Davis experiment is sensitive only to electron-type antineutrinos, and so oscillations could offer an alternative explanation. If the emitted electron-type antineutrinos oscillated to yield equal proportions of the three neutrino types before they reached the earth, only one-third of the expected number would be registered by the detector.

Since neutrinos are so abundant, the discovery of even a small neutrino mass could more than double the total mass of the universe. The doubling would in turn substantially alter the gravitational interactions of matter in galaxies. Sky surveys have suggested that most galaxies are members of clusters; curiously, however, the apparent mass of the galaxies, estimated from the number of stars and the amount of gas and dust present, seems 10 times too small to hold the clusters together. An all-pervading sea of neutrinos with a mass of about 10 electron volts might account for the cohesion of the clusters.

On a still larger scale a neutrino mass would influence the overall evolution of the universe. Estimates of the visible mass of galaxies and intergalactic matter suggest that the universe has an "open" structure, meaning that the expansion observed now will continue indefinitely. A large enough neutrino mass could eventually halt the expansion and then reverse it in a great collapse.

The consequences of a neutrino mass will remain idle speculation until it is definitely established that the mass is not zero. The result of Sulak's search at Brookhaven should be announced within a few months, and it might alleviate some of the confusion. Several further experiments now in preparation are aimed at settling the question.

Reines and his colleagues are installing a new detector at the Savannah River reactor. The detector and all its shielding will be on wheels so that measurements can be made at various distances from the reactor core. If the wavelength of the neutrino oscillations is only a few meters, as Reines' earlier results suggest, it should be possible to observe peaks and troughs in the components of the neutrinos simply by moving the apparatus. Similarly, the group now working at the Laue-Langevin Institute will move to a more powerful reactor this summer and will make measurements at varying distances, and the CERN beam-dump experiments will also be repeated with more sophisticated detectors in new positions.

At Brookhaven a detector for a new search is being assembled by a group from the University of Pennsylvania, Brown University, Brookhaven and Japan. The working medium will be 150 tons of a liquid that scintillates when a charged particle passes through it. A still more ambitious plan for the Brookhaven neutrino beam has been proposed by Sulak and workers at the University of Michigan. Seven detectors would be buried in the path of the underground beam at various distances of up to a kilometer from the source. Each detector would be a vat holding some 200 tons of ultrapure water.

Social Invention

A "new social invention" developed on the Lower East Side of Manhattan in New York City gives promise of integrating the fragmented social, medical and income-maintenance services made available to families in trouble by governmental and voluntary agencies, and doing so in behalf of a particular objective that has long been underemphasized: strengthening the family and keeping it intact. The Lower East Side Family Union is a social-welfare agency formed in 1972 at the behest of five neighborhood community-service settlements. Its major goal is to provide rehabilitation for "hard-pressed families for whom disintegration is an immediate threat"; it aims to reduce the frequency with which children in "the most problem-ridden families in a poor, multi-ethnic New York City neighborhood" are placed in institutions or foster homes, according to Bertram M. Beck of the Community Service Society of New York. His report on the Union has been published by the Foundation for Child Development.

Beck traces some of the historic trends in social welfare that led to the perceived need for a new kind of agency, notably an excessive addiction to the foster placement of children. The early child-welfare movement set great store on the benefits of a change of environment; relatively little attention was given to the relation of children to their natural parents. The availability since the 1930's of substantial public funds "made possible a vast expansion of foster placement.... Once funds were available, foster care was increasingly prescribed as the remedy of choice.... The choice of treatment is more apt to be dictated by the services available than by the nature of the problem." In New York in particular, according to Beck, "voluntary, sectarian child-care services exercised substantial political influence and thwarted the development of public child care."

Other influences have included psychoanalytic theory, which led to "heavy emphasis on counseling and insight development and minimal practical assistance" and a "crazy quilt" of programs and agencies, each focused on "a particular service authorized by law and bound by regulation" and ignoring "the fact that the family itself is a system, and no one individual or problem can be looked at or successfully dealt with entirely independent of other individuals and problems." The fragmentation is institutionalized, since "organizations seek their own autonomy and will not allow themselves to be coordinated in any way that interferes with their own aggrandizement."

The Family Union attempts to remedy the obstacles to direct family support that stem from these influences by integrating services at the neighborhood level. It is supported by funds from foundations and from the State of New York and by a loan of personnel from the New York City Department of Social Services; it is independent of the various organizations whose services it works to integrate. Some 400 families a year are referred to the Union—most of them by public child-welfare agencies, others by schools, the police, the clergy, another social agency or a family that has been helped. The Union concentrates on some 200 multiproblem families that are at high risk of disintegration. A referred family is assigned to one of three teams, largely made up of social workers and homemakers drawn from the neighborhood and trained on the job.

After assessing the family's situation a Union social worker sets up a conference involving family members and representatives of organizations that can provide the necessary services. A written contract is drafted, specifying the responsibilities of each agency and of the family itself, and is signed by the participants. In one typical contract cited by Beck one agency agrees to provide a visiting nurse "about once every three weeks for the next three months"; a hospital social worker undertakes to provide a part-time housekeeper and to coordinate clinic appointments; a settlement house accepts two children in its Headstart program and agrees to provide counseling; another settlement promises after-school day care for older children; the children's mother agrees to make an appointment for a medical examination for herself and to see that the children show up at the Headstart program. The Union staff member follows up on the contract and sees that it is carried out.

Much emphasis is put on the role of the homemaker, who is provided by the Union and works closely with the social worker, helping parents to deal with daily problems and substituting for them in emergencies. When temporary placement of children outside the home is essential, the preferred solution is an informal arrangement with neighbors or relatives or, if necessary, formal foster care in the neighborhood. Of the 200-odd families taken on each year, only some 5 percent have children who need foster care, in most cases for less than six months.

Beck suggests that the Union could be replicated in other communities "if there were an inducement in the form of money" to support its particular functions. This might be done by foundations or by a public agency; possibly a government agency could require that social-welfare organizations participate in a similar system as a condition of receiving certain public funds. Although the Union's work is specifically aimed at preventing unnecessary foster care, Beck points out that a similar approach could be used to deliver services to the aged, the mentally ill or retarded or any other category of people who need integrated services and are in no position to secure them on their own.

Reverse English

One approach to the study of the brain holds that the functions of the human brain most worth investigating are the ones the brain performs almost instantly and without conscious effort. The best example is the understanding and the utterance of language. Some people, however, seem to have linguistic abilities that other people do not. A special case is Andrew Levine, because Andrew Levine can talk backward.

Levine is professor of philosophy at the University of Wisconsin. As someone speaks to him, even quite rapidly, he can almost simultaneously utter the same sequence of words with the sounds in each word reversed. He does not reverse each word in strict accord with its spelling. His utterance of the reverse of "peace," for example, sounds like "seep," and not like an effort to say "ecaep." Moreover, where two words have the same spelling but different sounds (such as "bow" in its different meanings) his utterances of the reverse are different in sound as well.

Levine's backward speech is being studied by Nelson Cowan, Lewis A. Leavitt and Dominic W. Massaro of the University of Wisconsin, in collaboration with Raymond D. Kent of the Boys Town Institute. The investigators have determined that Levine's reversal of English is not the precise equivalent of playing a recording backward. In the word "ironic," for example, the long *i* at the beginning is a single phoneme: one of the elemental sounds of which a spoken language consists. To utter the long *i*, however, the mouth of the speaker must move from one position to another. If Levine reversed the utterance of "ironic" in the way a reversed recording does, the two positions of his mouth would reverse in sequence as well, and Levine would say something sounding like "kinarea." Actually he says "kinari." The word is reversed but the phoneme is intact. Levine's ability therefore tends to confirm that it is not merely linguistic theory that analyzes words into trains of certain basic sounds. The human brain does so too.

Other evidence suggesting that this is the case has come from studies of infants. In one experimental technique, employed by Leavitt in collaboration with Philip A. Morse, the infant's heart rate is monitored as he listens to the repetition of a computer-synthesized sound. Typically his heart rate slows slightly at first but quickens as he loses interest. Then he hears a sound he regards as different and his heart rate slows again.

An example of such an experiment involves the sounds of *b* and *p*. The difference between the two is that in the case of the *p* there is a greater lag between the movement of the lips and the beginning of activity by the vocal cords. The computer can thus be programmed to synthesize not only the sounds of *b* and *p* in combination with a subsequent vowel (for example *ba* and *pa*) but also intermediate sounds that do not occur in human speech. An infant then hears the repetition of one computer-synthesized sound succeeded by the repetition of another. The monitoring of his heart rate as the experiment proceeds can establish whether the infant regards the new sound as novel and hence can show whether he distinguishes it from the first. In this manner it has been established that infants and adults tend to categorize sounds in the same way. They make the same distinctions and they fail to make certain distinctions. What has not yet been established is whether the infant categorizes phonemically. The sounds of *b* plus a vowel and *p* plus a vowel are combinations of phonemes; the sounds of *b* and *p* cannot be synthesized alone.

The investigators are trying to learn how it is that Levine is able to talk backward in the first place. They have found that his memory for speech sounds and letter sequences is exceptional. One thing that hampers the research is the lack of any earlier work on the phenomenon. Levine may be only the third backward talker on record. Although one of his predecessors became a professional entertainer with the stage name Professor Backward, Levine can talk backward better.

Eighth-Century Ark

The place where Noah's ark came to rest, as it is recorded in the King James Version, is "upon the mountains of Ararat," evidently a region at the headwaters of the Tigris within the boundaries of ancient Urartu. (Both Ararat and Urartu can be translated as "highlands.") Traditionally this has come to mean twin peaks in eastern Turkey, Big Ararat (at 16,950 feet the highest mountain in the country) and Little Ararat (12,880 feet). Perhaps since the start of the Christian Era, if not before, pious people have hoped that the timbers of the ark could be found somewhere on those slopes; indeed, some have believed they had been found. A film, *In Search of Noah's Ark,* widely exhibited and televised in the U.S. and Canada as recently as 1977, was inspired by such thoughts.

It is less well known that Fernand Navarra, a retired French industrialist, has visited Big Ararat twice, first in 1955 and again in 1969. He collected a section of worked timber and other fragments of wood from exposures in a glacier on the northern slope of the mountain at an altitude of some 14,000 feet. In an account of his discovery titled *Noah's Ark: I Touched It* (1974) Navarra reported that the wood was oak. The density and color of the wood led forestry experts in Spain to state that it is about 5,000 years old (an age in reasonable accord with the chronology worked out by Archbishop Ussher, which dates the Creation to 4004 B.C.).

Reviewing the matter in a recent issue of *Antiquity,* R. E. Taylor of the University of California at Riverside and Rainer Berger of the University of California at Los Angeles note that over the past 15 years several laboratories have measured the age of Navarra's wood by the carbon-14 method. One date determined (at the National Physical Laboratory in Britain) was A.D. 770–90, a second (at the University of California at Riverside) was A.D. 730–60, a third (at the University of California at Los Angeles) was A.D. 730, a fourth (at the University of Pennsylvania) was A.D. 640, a fifth (at Geochron Laboratories, Inc.) was A.D. 620–40 and a sixth (also at Geochron) was A.D. 270. A seventh date determined (at Teledyne Isotopes) was lost through the misplacement of the records.

Pointing out that five of the seven determinations are statistically identical, Taylor and Berger note that in modern times crosses and other objects of religious veneration have been raised near the summit of Big Ararat. "It is tempting to suggest," they conclude, "that one or two centuries before the Islamic conquest of Asia Minor a cenotaph or memorial was erected by Armenian or Byzantine clerics to commemorate what they believed to be... the final resting place of the Ark.... Perhaps this cenotaph was actually built in the form of a boat."

The Zapping of a Cold America

In spite of recent public concern about the possible health hazards associated with the use of various devices that generate microwave radiation, a leading expert on the interaction of electromagnetic waves with matter has come forward with a proposal that would deliberately promote rather than limit the presence of microwaves in the human environment. In a report published in *Science* R. V. Pound, professor of physics at Harvard University, draws attention to "the possible great social benefit that could be derived from the use of microwave radiant energy to provide directly the warmth needed by human beings and other living creatures." One reason for proposing such an application at this time, he writes, "is the view that such use could contribute importantly toward alleviating the developing world energy crisis." By delivering heat energy in the form of microwaves directly to the occupants of an enclosed space rather than to the enclosure or to the space itself, he explains, "body warmth and personal comfort could be maintained at far lower environmental temperatures than those that have become conventional, especially in the United States, in the departed era of cheap energy."

According to Pound, microwaves with a wavelength on the order of a centimeter interact with the human body mainly by raising the temperature of water molecules in the skin. The warming effect of this absorptive process, he says, "should seem little different from the warmth conveyed by the radiant heat from a warm stove." Moreover, microwaves, unlike infrared radiation, can be efficiently reflected from metallic walls; in the process neither the walls nor the air need be significantly heated. "In this way the room would become a resonator supporting a large number of modes of high numerical index that tend to overlap in frequency. Thereby, the room would be filled relatively uniformly with radiant energy.... An occupant would then experience uniform warming irrespective of his location in the room."

Pound estimates that the microwave power needed to keep people comfortable under such circumstances would be about 60 watts, a level that is not incompatible with the current radiation-protection guide used in Western countries, "even though the guide was set under the assumption that radiation could only be deleterious." Nevertheless, he adds, "experiments are needed to establish suitable parameters in practice."

Addressing himself to the question of whether there are important nonthermal effects that must be considered in the interaction of microwaves with biological tissue, Pound notes that "microwave radiation contains only photons whose quantum energy is very much smaller than the mean energy of the pervasive thermal agitation and vibration of individual atoms and molecules of all matter at temperatures supportive of life. Microwaves form a part of the spectrum of radiations emitted by all objects at temperatures above about 1 K, and are present in the radiation from the sun. These photons, individually, have none of the penetrating and disruptive qualities possessed by gamma rays, X rays or energetic particles encountered in nuclear reactions." In any event, he continues, even advocates of the view that there are dangerous nonthermal effects of microwaves accept the argument that the shallow penetration of microwaves with wavelengths shorter than about three centimeters makes them unsuitable for testing for such effects.

Pound acknowledges that the association of his proposed form of heating with microwave ovens for cooking might give rise to some fears. He points out, however, that the microwaves produced in ovens have a wavelength of 12 centimeters or more and that they penetrate more than 10 times deeper than three-centimeter waves in tissue with an appreciable water content. Furthermore, he maintains, "cooking is a matter of degree. All forms of domestic heating have their counterparts in cooking, when the level of the heat is made sufficient. One may associate, for example, forced hot air (or a dry sauna) with baking, a damp sauna with steaming, and radiant heat with broiling. The open fire itself is also used for broiling, but that does not prevent a fireplace from serving as a pleasant source of warmth. The important factor in all cases is the limitation of the exposure to an appropriate level."

Megalithic Monuments

These assemblages of massive stones, of which Stonehenge is one, are found by the thousands in Europe. How old they are was long uncertain, but they have now been dated to the Neolithic period

by Glyn Daniel

Among the most dramatic remains of the ancient cultural landscape of Europe are its many prehistoric stone monuments. These megaliths have long aroused the interest and curiosity both of the general public and of antiquarians and archaeologists, with their interest in correctly describing the nature, purpose, context and age of the structures. To give only two examples, the stone rows of the Carnac region in southern Brittany, where more than 3,000 menhirs stand in parallel lines that extend for nearly four miles, and the Grand Menhir Brisé at nearby Locmariaquer, now broken but originally 22 meters long, are among the most remarkable relics of prehistoric France.

Without doubt the most famous of all megalithic monuments is Stonehenge, on the Wiltshire plain of southern Britain. Visited by thousands yearly, it is second only to the Tower of London as a tourist attraction. It has a larger literature than any other archaeological site in the world, including the pyramids of Egypt and the great statues of Easter Island, as well as mythical sites such as Atlantis. The number of books on Stonehenge and on other megalithic monuments that have poured from the presses in the past decade or so is a measure of the continued interest in these antiquities.

It is also, alas, an all too clear demonstration of the imagination, wishful thinking and credulousness of many authors, and the abysmal ignorance of many alleged archaeologists who can only be described, if uncharitably, as fantasy buffs. This is no new phenomenon. As long ago as 1911 G. Elliot Smith's book *The Ancient Egyptians* brought all these ancient European monuments from the banks of the Nile. Such exercises of the imagination continue. As recently as 1977 Euan MacKie in *The Megalith Builders* declared that they were the work of wise men from predynastic Egypt and Sumeria. There are others, among them Erich von Däniken, who see the megalith builders as voyagers from space. Now there is also a widespread belief that these monuments were built with an astronomical purpose, and such words as "astro-archaeology" and "archaeoastronomy" are freely bandied about. Let us take a sober and balanced look at these structures in the context of our existing detailed knowledge of ancient Europe.

The Study of Megaliths

In 1849 in a book called *Cyclops Christianus* an Oxford don, Algernon Herbert, coined the word megalith (from the Greek *megas*, great, and *lithos*, stone). The word caught on. Although in 1872 James Fergusson, a Scottish architectural scholar, titled his book *Rude Stone Monuments in All Countries*, he too spoke of megaliths. So did T. E. Peet in his *Rough Stone Monuments*, published 40 years later.

These pioneer works established the proper study of megaliths, but they introduced one fundamental confusion. As travelers such as Fergusson journeyed outside Europe they found great stone monuments in Algeria, Palestine, Ethiopia and the Sudan, the Caucasus, Persia, Baluchistan, Kashmir and central and southern India. (In India the megalithic monuments of the Deccan, many of them with "portholes" resembling those of megalithic tombs in Europe, first interested Fergusson in undertaking his comparative researches.) The list does not end there. Megaliths are found in Assam, in Sumatra and on some Pacific islands such as Malekula in the New Hebrides. (The stone figures of Easter Island, although they are certainly large, are not megaliths in the generally accepted sense.) In Japan megalithic tombs were built from the second century B.C. until the seventh century A.D., when the emperor Kotoku forbade them as a waste of labor. The pre-Columbian civilizations of the New World also practiced megalithic construction.

The result of applying the term megalith to all these monuments in different countries, from different periods and in different cultural contexts gave birth to an absurdity: the idea that the structures were genetically connected, that they were the work of a megalithic race or a megalithic people. This notion has long been abandoned, and it is widely realized that the megalithic structures in different parts of the world are similar because they are made of similar materials in similar ways. The parallels are particularly striking in megalithic chambers or rooms that incorporate the basic elements of what is known as trabeate architecture (from the Latin *trabes*, beam). This type of construction is like building a house of cards or of children's blocks: slabs of stone are set upright (orthostats) and other slabs are laid across the uprights as capstones. The architectural possibilities are limited, and so it is not surprising that a megalithic chamber in France or Ireland, dating from the third millennium B.C., should resemble a megalithic chamber in southern India dating from the end of the first millennium B.C.

Another confusion in the minds of many is one between megalithic architecture and cyclopean architecture. The latter also makes use of large stones, but a cyclopean structure is built of stones that are carefully fitted together, even if irregular in shape, and generally set in layers. Cyclopean architecture is found in both the New World (for example the Inca structures of Peru) and the Old (for example the citadels of Mycenae and Tiryns in Greece or the nuraghi, or stone towers, of Sardinia).

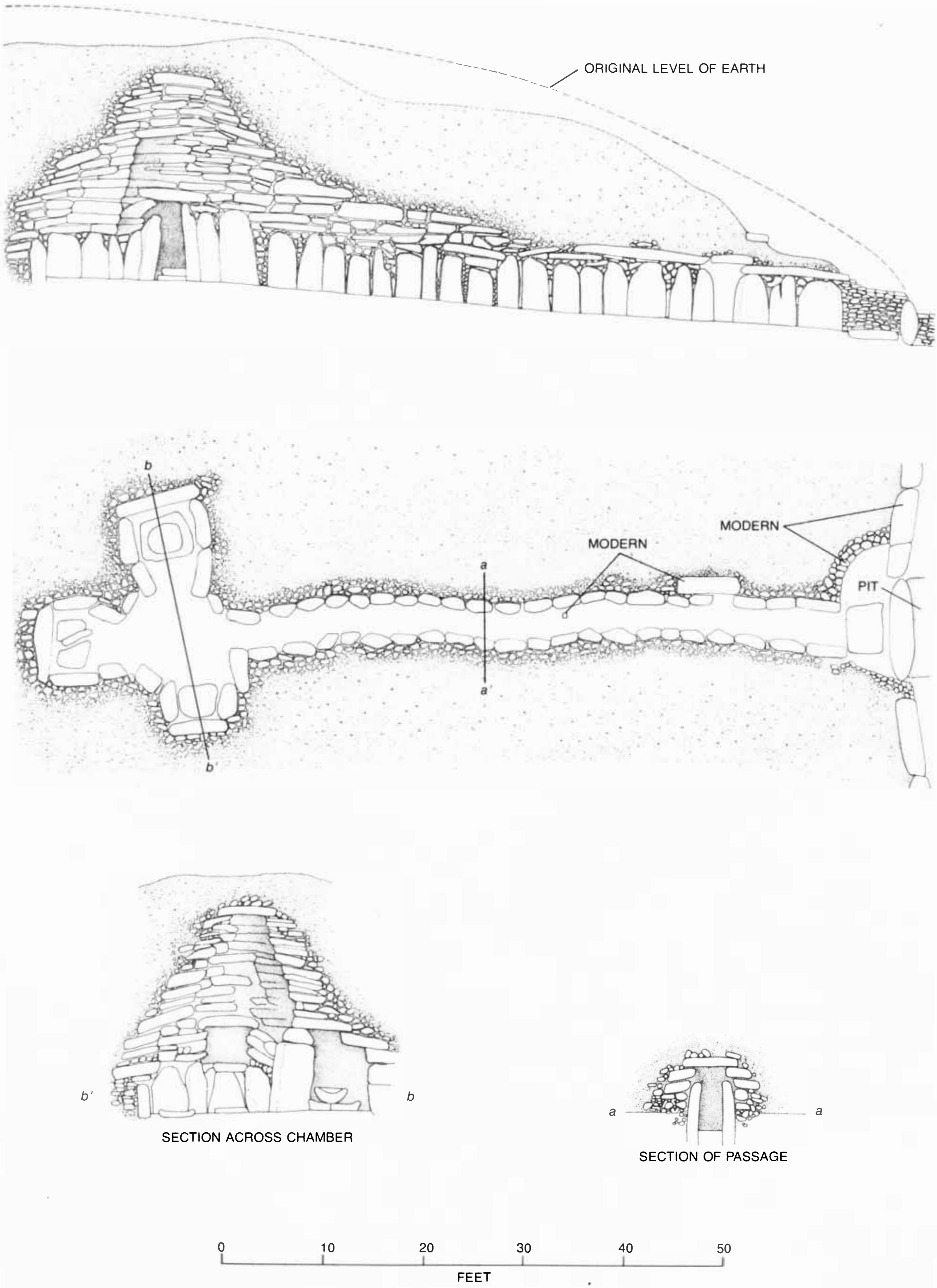
The European Megaliths

The megalithic structures of Europe fall into four main categories. First is the menhir, or single standing stone. The word comes from the Welsh *maen*, a stone, and *hir*, long. Brittany is rich in menhirs that range in length from one meter to six meters. A notable exception is the Grand Menhir Brisé, a much larger horizontal stone. No one knows whether this great stone ever stood erect; the earliest records describe it as it is now.

A special kind of menhir, known as a statue menhir, is sculptured so that it bears the representation of a person, hu-

GRAND ARRAY AT CARNAC in southern Brittany includes three separate avenues of menhirs, or standing stones. Seen in this aerial photograph is the Kermario display, consisting of 10 roughly parallel lines of menhirs, hewn from local granite, extending some 4,000 feet.

MOST FAMOUS MONUMENT of the megalithic tradition, Stonehenge is a ring surrounded by a bank and a ditch, situated on the Wiltshire plain of southern Britain. The ring was built in phases, starting in about 2800 B.C., and was completed sometime after 1100 B.C.



GREATEST BURIAL CHAMBER of the megalithic tradition is Newgrange in County Meath in Eire, shown in elevation at the top and in plan and sections at the bottom. The narrow passage that leads to the burial chamber at the center of the great mound is some 60 feet long; the corbeled vault of the chamber is 20 feet high (*see photograph on opposite page*). Newgrange was built in about 2500 B.C.

man or divine. These are found in southern France and northern Italy, with outlying examples both in Spain and on the Channel Islands. These interesting uprights, the earliest monumental sculpture in the round in human history, can be dated to the end of the third millennium B.C. and the beginning of the second. They are not, however, necessarily connected with the undecorated menhirs.

The second category of megalithic structures is made up of grouped standing stones. The stones are set either in rows, as those at Carnac are, or in what used to be called stone circles. The careful surveys of Alexander Thom, retired professor of engineering at the University of Oxford, have shown that many of the latter monuments are not strictly circular: many are in the form of an ellipse or a flattened ellipse. As a result the term now coming into favor as a description of these megalithic enclosures is stone ring. Some stand alone; some surround a burial mound. Some are associated with rows of stones; some, like the famous monuments of Stonehenge and Avebury in southern Britain and Stenness and Brodgar in Orkney, are surrounded by ditches and banks.

The third category, the burial chamber, constitutes the commonest form of European stone monument. There must be 40,000 to 50,000 of these chambers surviving, and originally there were perhaps twice as many. The largest number of them are found in Spain and Portugal, France, Britain and Ireland, Scandinavia and northern Germany. Some are completely buried under large mounds of earth and stone, some bear traces of partially destroyed mounds and others are entirely freestanding. It was thought in the 19th century that all freestanding chambers were the ruins of chambered mounds, but this is not now thought to be so. Some are the remains of denuded burial mounds but others, such as the portal chambers of Ireland and the great stone galleries of western France, were probably always as they are today.

Many burial chambers exhibit a roofing technique more elaborate than the simple capstone; stones were placed on top of the upright orthostats in such a way as to overlap until they formed a corbeled roof, or vault. Among the most famous surviving corbeled vaults are Maes Howe in Orkney and Newgrange in County Meath in Eire. Maes Howe was broken into by the Vikings, but Newgrange, with its vault rising 20 feet above the ground, has remained intact for 4,500 years. It is one of the wonders of prehistoric Europe.

Many of the stones in these chambers are very large. The capstone of the Mount Browne chamber in County Carlow in Eire is estimated to weigh some 100 tons. The great megalithic chamber at Bagneux, near Saumur in western France, is some 20 meters long by five meters wide, and its roof of four capstones gives three meters of head space to someone standing inside. (It once housed a café!) The capstones are about 60 centimeters thick; the largest of them is estimated to weigh 86 tons.

LOOKING UP to the top of the corbeled vault at Newgrange, this photograph reveals its stepped construction. Each successive layer of slabs was placed closer together until capstones could bridge the narrowed gap. A similar corbeled vault was built at Maes Howe in Orkney.

The fourth category, the megalithic temple, is a limited one. Most of the examples are found in the Mediterranean, on the island of Malta and its neighbor Gozo. The application of the term temple to these great megalithic monuments may conjure up in the minds of some a building comparable in appearance to the pillared structures of dynastic Egypt and Classical Greece. The Maltese monuments are very different. Their solid walls consist of very large slabs of stone, their floor plan includes projecting apses and they were probably roofed with wood beams and thatch. They are certainly among the most impressive prehistoric architectural monuments in the entire Mediterranean area.

That the Maltese structures were temples there is little doubt. They hold no burials and no traces of domestic occupation but do contain many cult objects, including figurines of a female deity and stones decorated with spirals and other designs. There is also no doubt that they date from the fourth millennium B.C. They are thus the earliest example of stone architecture in the Mediterranean and, together with the megalithic chambers of Spain and France, in the entire world.

In medieval times the common folk, and learned men too, thought the great stone monuments must have been the work of giants of long ago. Indeed, the first recorded excavation of a megalithic chamber in Sweden was for the specific purpose of testing this supposition. It was proved wrong; the remains found in the chamber were not those of giants. Later antiquarians sought to explain the megaliths in terms of written history. The structures were variously ascribed to the Romans, the Anglo-Saxons, the Danes, the Goths and the Huns. For example, Herbert, the coiner of the word megalith, argued that because Stonehenge and other British megaliths were not mentioned by Roman historians and visitors to Britain the structures must be post-Roman.

Who and Why

Gradually it came to be accepted that the megaliths were pre-Roman. The pre-Roman inhabitants of northwestern Europe, namely the ancient Gauls and Britons, had been described by Caesar and others as having a class of wise men called Druids. It is not surprising, then, that the megaliths came to be ascribed to the Druids. Some even took the freestanding chambers for Druidic altars.

Over the past two centuries the growth of archaeological excavation has revealed that megalithic chambers were primarily burial places, used collectively by a community or a family over a period of time. Some large chambers have yielded the remains of 200

1:1

Scarlet Lady (foreground) and Royal Gramma seen behind

3:1

Scarlet Lady (Hippolysmata grabhami)—Indo-Pacific and Caribbean

The Polaroid CU-70 photographic system uses all the technical advantages of the world's finest instant camera, Polaroid's SX-70 Sonar, and all the brilliance, detail and speed of the world's fastest developing color film, Time-Zero Supercolor SX-70.

In brief, the SX-70 Sonar Land camera is perhaps the most ingenious and easiest-to-use picture-taking system ever invented. It incorporates a computer, electronic eye, motor drive, film developing mechanism and revolutionary sonar focusing.

You just push a button. And in 1.5 seconds, the camera hands you a precisely focused, properly exposed, developing color photograph.

To take really close close-ups at 1:1, 2:1, or 3:1, just select the appropriate color-coded attachments from the CU-70 kit, frame your subject and shoot. (The kit's aluminum carrying case conveniently holds everything you need including camera and extra film.)

Close up or far away, there's no focusing, complicated exposure settings to calculate or lighting to adjust. And, since the photographs develop on-the-spot, you are certain that what you want to record is recorded.

Use the CU-70 wherever you need wide-range recording capability for instant documentation: quality control records, medical/biological specimens; insurance evidence and forensic photos; and collections of all kinds such as stamps, coins, objets d'art.

For more information about the new CU-70 photographic system and new Time-Zero Supercolor SX-70 film, contact a Polaroid photographic dealer. For the name of the CU-70 dealer nearest you, call Polaroid toll-free from continental U.S.: 800-225-1618. (In Massachusetts call collect: 617-547-5177.) He'll be glad to show you close up what our system can do for you.

Polaroid

Instant SX-70 pictures from 3X to Infinity.

DOLMEN, OR "TABLE STONE," is the name given to this form of megalithic monument in Brittany. It consists of a single capstone resting on three or more upright supports and is the simplest form of burial chamber. This monument, situated in Pembrokeshire in Wales, is known as the Carreg Samson cromlech, a Welsh word equivalent to dolmen. Some chambers are freestanding; others are mounded over.

LARGER CHAMBER, also in Pembrokeshire, now has only one capstone. This freestanding burial chamber, at Pentre-Ifan, was first recorded by George Owen in his survey of Pembrokeshire published in 1603. Perhaps 50,000 such burial chambers survive in Europe.

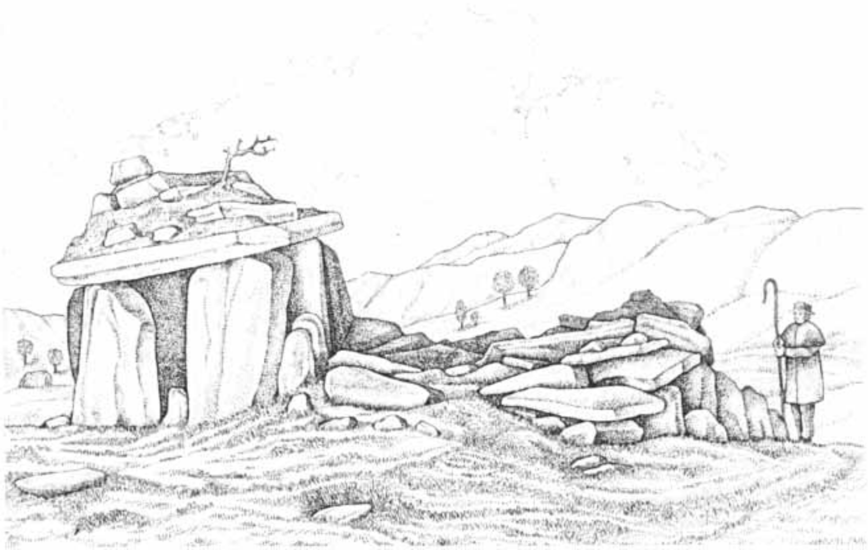
individuals or more. Not all megalithic chambers, however, were necessarily tombs. Some, like the freestanding chambers of Ireland and the Loire valley of France, may have been temples, although scarcely Druidic ones. Indeed, the line between a tomb and a temple is not a hard and fast one, as we realize when we look at Christian churches and cathedrals.

The grave goods found in many megalithic chambers are unspectacular: undecorated pots and polished stone axes. In some areas, however, they are rich and remarkable. In Scandinavia heavily decorated pots with oculi, or eye ornaments, are found. In Spain many tombs yield large numbers of schist plaques covered with designs, some of them anthropomorphic. Tombs in Brittany have yielded superbly fashioned polished axes and rings of jade, garnet and other precious materials. These date from the third millennium B.C., long antedating the Chinese and Maya achievements in jade.

The purpose of the menhirs is more difficult to evaluate. Sometimes burials are found at their foot, but they are not tombs. Moreover, such burials cannot even provide a date for the monument, let alone establish its original purpose. Perhaps these single standing stones were intended as territorial markers. Perhaps alternatively they were meant as memorials for the dead, a prehistoric version of the Greek cenotaph, or empty tomb.

The purpose of the menhir alignments and the stone rings is clearly neither funerary nor domestic. Perhaps these monuments were places of assembly where tribes or communities met from time to time for both secular and religious purposes. I see them functioning like the Breton *pardons* of the present day: opportunities for priests to conduct sacred rites, for crops and animals to be blessed, for friends and relatives to meet and enjoy a social occasion, a market and a fair, particularly a hiring fair.

Does such a description fit Stonehenge, the best-known stone ring of them all? First it should be remembered that Stonehenge is a complex monument of several periods and many architectural features, including two circles, two horseshoes and carefully shaped "trilithons" that are a far cry from simple orthostats and capstones. The stones themselves are in part sandstone from the Marlborough Downs and in part "foreign" blue stone from the Preseli hills of Pembrokeshire in southern Wales. There are some 80 of the blue stones, weighing up to four tons each; the distance from Preseli to Stonehenge is 135 miles as the crow flies. The sandstone units, better known as sarsen stones, are also large and were carefully dressed. The largest single component of the sarsen trilithons measures 29 feet eight inches, and the average height of the stones in the sarsen circle is 13 feet six inches. One of the most interesting and architecturally sophisticated devices used by the builders of Stonehenge is what is known in Greek architecture as entasis. This consists of shaping an upright so that the effect of perspective is canceled; when the stones are seen at close range, they do not appear to taper upward but give the optical illusion of being straight-sided.



EROSION OR VANDALISM had stripped away the mound that once concealed this chambered tomb by the time the artist recorded the monument in the 1840's. Only one capstone still bears some of the former covering of rubble and earth. The engraving on which this sketch of Bryn Celli Ddu in Anglesey in Wales is based appeared in *Archaeologia Cambrensis,* 1847.

Measuring the Year

There is no doubt in anyone's mind that Stonehenge was built with its central axis pointing to the midsummer sunrise. Many years ago the British astronomer Sir Norman Lockyer argued that by calculating the exact orientation of Stonehenge he could assign an exact date to the construction of the monument. His conclusion was 1680 B.C. ($\pm$200 years). Since then R. J. C. Atkinson of University College, Cardiff, who has done more than anyone by excavation and research to contribute to our modern knowledge of Stonehenge, has pointed out that, owing to errors in Lockyer's original reasoning, his date should be altered slightly. The new reading is 1840 B.C. ($\pm$200 years).

The astronomical purpose of Stonehenge and other stone rings was not seriously argued until after World War II, when Gerald S. Hawkins of Boston University proposed in his book *Stonehenge Decoded* (1966) that the monument was a giant calculator for the prediction of eclipses, both lunar and solar. Five years later Thom, in his book *Megalithic Lunar Observatories,* postulated that many megalithic monuments served for observation of the movements and phases of the moon. Thom's surveys had already led him to argue for the existence of a megalithic "yard" measuring 2.72 feet and to suggest that the builders of stone rings had a knowledge of Pythagorean geometry 2,000 years or more before the Greeks. These are extravagant and unconvincing claims; what the builders of megaliths had was a practical knowledge of laying out right-angled triangles.

Many people, no doubt bored by the prosaic account of megaliths to be got from archaeological research, jumped on the Hawkins-Thom bandwagon, accepting the builders of megaliths not only as experts in Pythagorean geometry and possessors of accurate units of mensuration but also as skilled astronomers who studied eclipses, the movements of the moon and the positions of the stars. To me this is a kind of refined academic version of astronaut archaeology. The archaeoastronomy buffs, although they very properly eschew wise men from outer space, very improperly insist on the presence in ancient Europe of wise men with an apparently religious passion for astronomy. It seems to me that the case for interpreting megalithic monuments as astronomical observatories has never been proved. The interpretations appear to be subjective and imposed by the observer. Already new surveys are showing the inaccuracy of some of the earlier observations and undermining the hopes of those who believe the builders of megaliths were slaves of an astronomical cult.

The entire study of megaliths in Europe has been revolutionized not by surveyors with their eye on the moon and the stars but by advances in prehistoric dating. The first of these was the development by Willard F. Libby and his

REVERSING SPIRALS decorate a low wall in a Maltese megalithic temple at Hal Tarxien, near Valetta. The numerous megalithic temples of Malta and nearby Gozo, once believed to have been inspired by Minoan structures, are now known to be indigenous in origin.

SPIRAL DECORATION is predominant among the motifs executed in low relief on this megalithic column at the entrance to Newgrange. The decoration, if it is as old as the burial chamber itself, is at least 1,500 years later than the earliest Maltese megalithic temples.

colleagues of carbon-14 dating. The first carbon-14 dates were published 30 years ago, and since then thousands of them have been determined by laboratories all over the world. A more recent geochronological technique, thermoluminescence dating, has confirmed many carbon-14 findings.

The Question of When

Before the carbon-14 revolution—and it has been no less than revolutionary for the field of prehistoric archaeology—the ages of various ancient works of man had either been guessed at or calculated in a regrettably uncertain way by correlations between the dated civilizations of Egypt and Mesopotamia and undated barbarian Europe. Thus in the period between the two world wars it was customary to assign the earliest European megaliths to the Neolithic period, say between 2500 and 2000 B.C., and the great monuments such as Newgrange and Stonehenge to the Bronze Age, between 1800 and 1500 B.C.

Now, thanks to the carbon-14 revolution, we can confidently state how old the megaliths of Europe actually are. The Maltese temples date from 4000 to 2000 B.C. The megalithic chamber tombs of Spain and Portugal date from 3800 to 2000 B.C. The British and Irish tombs date from just after 3800 to 2000 B.C. and the Scandinavian tombs from before 3000 B.C. to, say, 1800 B.C. In the Irish sequence it is good to have a definite date for Newgrange, showing that it was put up in about 2500 B.C. In the British sequence Atkinson has set down the chronological details of Stonehenge: the first phase was from 2800 to 2200 B.C.; the second phase, including the arrival of the blue stones from Wales, was from 2100 to 2000 B.C.; the third phase, which includes three subphases, was from 2000 to 1100 B.C., when the final phase began.

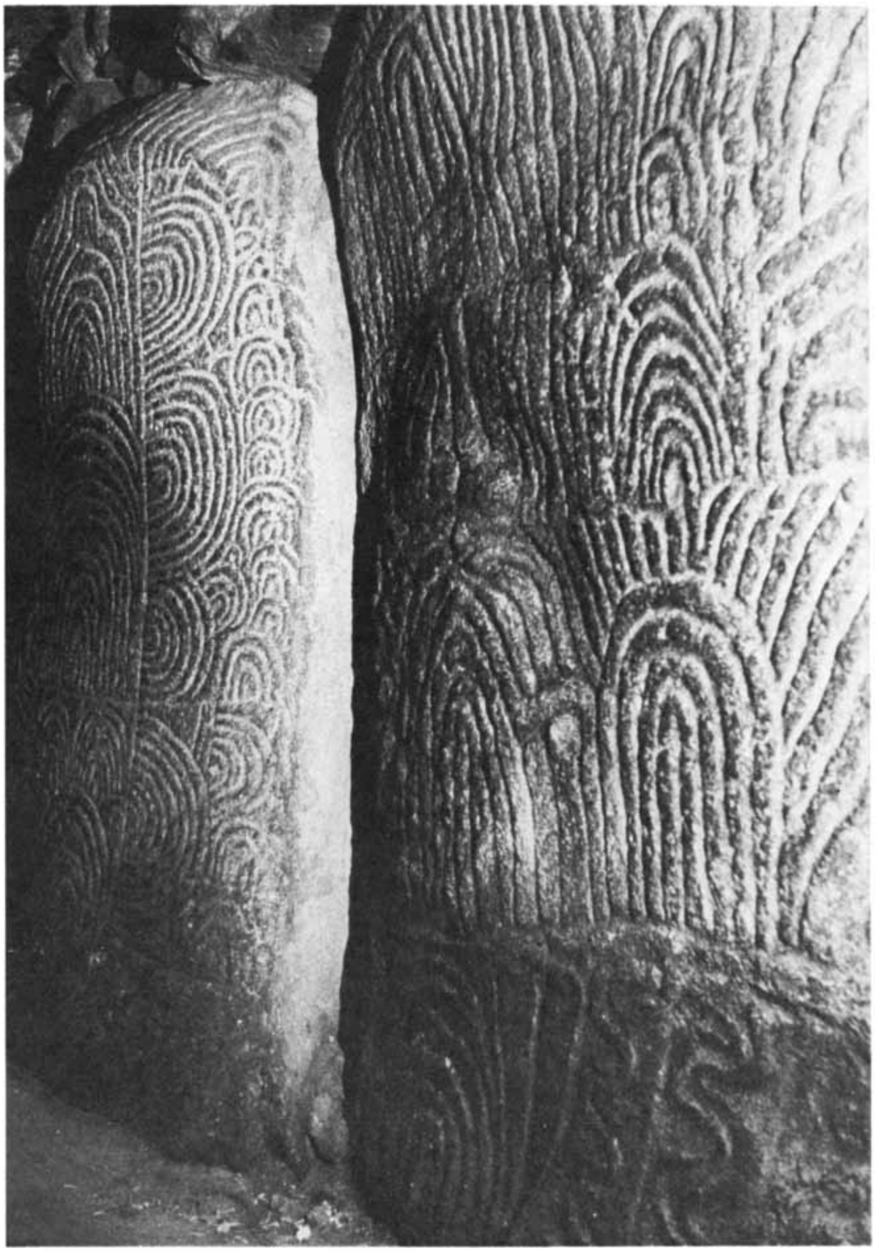
The dates I have given here, for Stonehenge and for megalithic monuments in general, are what are called calibrated carbon-14 dates. This is to say that they have been adjusted to the corrections based on the study of the rings of the bristlecone pine as displayed in the variation curves plotted by Hans E. Suess of the University of California at San Diego, by R. Malcolm Clark of Monash University in Australia and by others.

How can one explain the origins of these megalithic monuments now that they are accurately dated? There is only one tenable explanation of the menhirs of Brittany and elsewhere in northwestern Europe. They are a local invention; perhaps they represent the translation into stone of an earlier practice of setting up wood posts as cenotaphs or territorial markers or even totem poles.

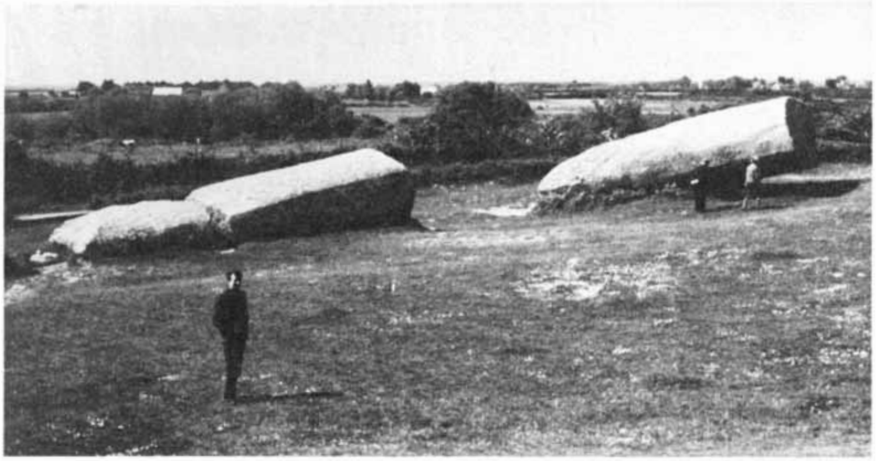
There is now only one tenable explanation of the Maltese megalithic temples. It used to be argued that they were derived by diffusion from Minoan Crete or from Mycenae or from even farther afield. Carbon-14 chronology now shows them to be earlier than any Minoan, Mycenaean, Egyptian or Sumerian context. The Maltese temples were an indigenous development. Possibly they are aboveground versions of subterranean rock-cut temple tombs. In any event they appear to have no antecedent anywhere, and no structures that can be confidently derived from the Maltese temples have ever been found outside Malta and Gozo.

We now come to the origins of the megalithic chambers and the stone rings. These must not be looked on as isolated phenomena. Instead we must seek to explain them in the context of the Neolithic societies that created them. Elliot Smith's idea that the European megaliths were derived from the mastabas, or stone tombs, of Egypt has no basis in fact and no suggestion of probability. We now know that the megaliths of Europe are older than the mastabas and pyramids of Egypt.

Still another hypothesis, voiced by V. Gordon Childe and others and tenaciously held by most archaeologists during the second quarter of this century, suggested that the megaliths of Europe were built by people who originated in the eastern Mediterranean, particularly in Crete and the islands of the Aegean. Even in the days before carbon-14 dating it was becoming clear that the idea of, for example, deriving great chamber



SEMICIRCULAR MOTIF predominates in the low-relief sculpturing of these uprights at the megalithic burial chamber of Gavrinis in France, roughly contemporaneous with Newgrange. Whether the patterns were decorative or whether they had religious meaning is not known.



LARGEST MENHIR in western Europe is this broken specimen at Locmariaquer in Brittany. Three of its four fragments are seen here; the combined length of the four is 22 meters. Whether the menhir ever stood upright is not known; earliest records show it in its present position.

CHRISTIANIZED MENHIR in the Yorkshire village of Rudston, 25 feet high, is the tallest in Britain. The megalithic monument was not thrown down; the graveyard grew up around it.

tombs such as Newgrange and Maes Howe from the domed subterranean tombs of Mycenae was chronologically impossible. As a second line of defense the hypothesis was adjusted backward in time to the vaulted tombs of the Messara in Crete and the rock-cut tombs of the Cyclades. Now both carbon-14 and thermoluminescence determinations show that the European megalithic chambers antedate any of the collective tombs in the eastern Mediterranean.

The New Conclusion

We are therefore forced to conclude that chamber tombs originated independently in at least seven areas of Europe: southern Spain, Portugal, Brittany, northern France, northern Germany, Scandinavia, southern Britain and Scotland. Does this statement appear to be a victory for the "independent evolution" school of archaeological theory over the "diffusionist" school that has held the field for so long?

It is not as simple as that. The Neolithic societies that began to put up megaliths in some areas of Europe in about 4000 B.C. were those that were already building comparable nonlithic or nonmegalithic structures, to wit houses and graves. Many of us think at present that an explanation of the appearance of megalithic chamber tombs in many different parts of Europe must take into consideration three successive phases. The first phase postulates an early European Neolithic tradition of building houses of wood or stone. The second phase involves the transformation of these domestic structures into tombs, still constructed of wood, turf or nonmegalithic stone. The third phase involves the translation of these widespread Neolithic traditions into megalithic architecture in separate areas of Europe that were without doubt interconnected.

The origin of stone rings may be explained in much the same way, although here the phenomenon does not involve seven separate but interconnected areas. The rings are found only in Britain and Ireland. Just as the chamber tomb translates a simpler precursor, so the stone rings cannot be dissociated from the wood-ring monuments now commonly referred to as woodhenges.

I see the origin of stone rings this way. First there were circular clearings in the forests that covered Neolithic Europe in the fifth and fourth millenniums B.C. We can postulate that sacred and secular gatherings took place in these clearings. Next, owing to the agency of man's domestic animals and man himself, the forests disappeared, whereupon artificial clearings were created by setting posts in a ring as a stage for similar gatherings. The third phase was the translation of the wood rings into stone rings. Then finally, as the tour de force of a succession of what one can only call cathedral architects, Stonehenge was built in the middle of the third millennium B.C. and flourished as a temple cum meeting place cum stadium for more than 1,000 years. This brings us back to my earlier question: Does Stonehenge fit the description of megalithic rings in general as sacred and secular meeting places? The answer seems to be emphatically in the affirmative.

Megaliths and Religion

One feels there must have been an impelling faith to inspire the Neolithic people who labored mightily to construct the megaliths of Europe. Tombs, temples, cenotaphs, meeting places—whatever the monuments were—appear to be manifestations of some powerful religious belief including a belief in an afterlife. Was the religion connected with the annual round of the sun, as the orientation of Stonehenge and other stone rings suggests? Was it connected with a mother goddess, as seems to be hinted by the figures from the Maltese temples and those on pottery and on the schist plaques from the tombs in Spain and Portugal? We may never know. Without written sources it is virtually impossible to reconstruct the characteristics, religious and social, of prehistoric societies. We do not know why the hunters of the Upper Paleolithic created paintings, engravings and sculptures in southern France and northern Spain, and we must admit that we also do not know the social and religious ethos of the megalith builders.

Of the thousands of megalithic tombs in Europe a few are decorated with designs carved in low relief or incised on the stone surface. Some in central Germany seem to be representations of tapestry, perhaps the hangings that may have lined the walls of houses. Others, particularly in the Paris basin, are clear representations of a female in frontal view; the eyes, nose and breasts of the deity are emphasized. The majority of the megalithic designs in the tombs of Brittany and Ireland, however, are geometric patterns: spirals, zigzags, lozenges, concentric circles. Perhaps the finest are at Gavrinis in Brittany and at Newgrange. Some, like the spirals on the great stone at the entrance to Newgrange, seem simply decorative. Others are to our eyes a bewildering confusion of what may be signs and symbols. It is important to remember that some of these geometric decorations are buried in the encompassing mounds of chamber tombs and were never meant to be seen; they were deliberately hidden at the time of construction. Why? Are they messages or are they sacred symbols? Whatever their meaning, it is set down in a notation we can never decipher.

In this connection it was discovered only 15 years ago that a representation

of a dagger had been engraved on one of the uprights at Stonehenge, as had representations of several axes of Early Bronze Age style. Some fancied the dagger to be a Mycenaean one. This kind of art is quite different from that of the chamber tombs, and of course the representations may have been executed late in the history of Stonehenge. To say that they are Bronze Age graffiti, however, is not to say that Stonehenge was a gigantic Bronze Age lavatory.

The practice of building great stone monuments in Europe died out by 1000 B.C., but the general population need not have simultaneously forgotten the nature and significance of the structures. It is by no means impossible that the folk of the first millennium B.C. continued to congregate and worship at the stone rings, and it is more than possible that the Druidic priesthood of the pre-Roman Celts of Gaul and Britain used them as temples. There is, however, no archaeological evidence of it.

At first Christianity strongly disapproved of people who worshiped stones, but gradually there came a new tolerance, which was generous enough for certain menhirs to be Christianized. Indeed, in Spain and Brittany a few megalithic monuments have been incorporated into functioning modern Christian churches. I take this to be a sign that the older faith of the builders survived in some shape or form until at least the Middle Ages of western Europe.

The Builders' Skills

It used to be asked: Who are the megalith builders of ancient Europe and where are their houses and settlements? We now think these questions are the wrong way around. It is the Neolithic villagers of Europe in the fourth and third millenniums B.C. who in certain areas built their tombs and temples in enduring stone. There has been much spoken and written recently about a revolution in our picture of the prehistoric past. Archaeologists are described as having thought of Neolithic peoples as savages, and so the new view of the past that shows them having great mathematical, geometrical and astronomical skills and knowledge dramatically changes our image of prehistoric man.

This thesis makes sense only to those who have never understood the archaeological record and who want to sensationalize prehistory. I have never had any doubt that the Neolithic peoples of Europe were good technicians and skilled engineers. They quarried large stones, transported them for considerable distances and erected them with consummate skill and artistry. We can gain some information about their probable techniques by studying the methods used today in areas such as Assam, where megalithic monuments are still being built. We can measure their accomplishments by noting the difficulties present-day farmers and building contractors encounter in trying to break up the prehistoric structures. For example, late in the past century the capstone of a megalithic chamber near Saumur was moved to be used as a bridge across a local stream. The movers built a number of enormous rollers more than a meter in circumference; each roller was made by lashing the trunks of four oak trees together. Even with the rollers in place 18 pairs of oxen were needed to move the load.

We must never deny the greatness of the megalith builders' achievements. Nor should we deny that from time to time, pausing from their labors at the harvest or at the construction of a monument, they looked as we all do to the sun, the moon and the stars. It is 6,000 years since the first megaliths were built in Malta and Brittany: 4,000 years before the beginning of the Christian Era and 1,000 years before the literate civilizations of Egypt and Mesopotamia. This is a sobering thought when we contemplate, with pride and pleasure, our megalithic patrimony.

AVEBURY ENCLOSURE, first recognized as a megalithic circle by John Aubrey in the 17th century, caused him to remark that it "does as much exceed in greatness the so renowned Stoneheng as a Cathedral doeth a parish Church." Few of the 190-odd menhirs that made up the circle and its avenues are visible in this aerial photograph, but the bank and ditch, formed by the quarrying of 200,000 tons of chalk, enclose an area of more than 28 acres. The conical mound visible at upper right, called Silbury Hill, is also manmade.

Owning a Toyota is always a good feeling.

But this price sticker could make owning one feel even better.

Because just $5,679* buys a lot of car—the Celica ST Sport Coupe.

$5,679! That's $710 less than the Datsun 200-SX.

$916 less than the Honda Prelude.

At $5,679, the Celica ST Sport Coupe is also $1,816 less than a Mazda RX-7.

And over $2,000 less than a Volkswagen Scirocco!

With a sticker price as low as $5,679, you may ask how we do it.

Do we give you less?

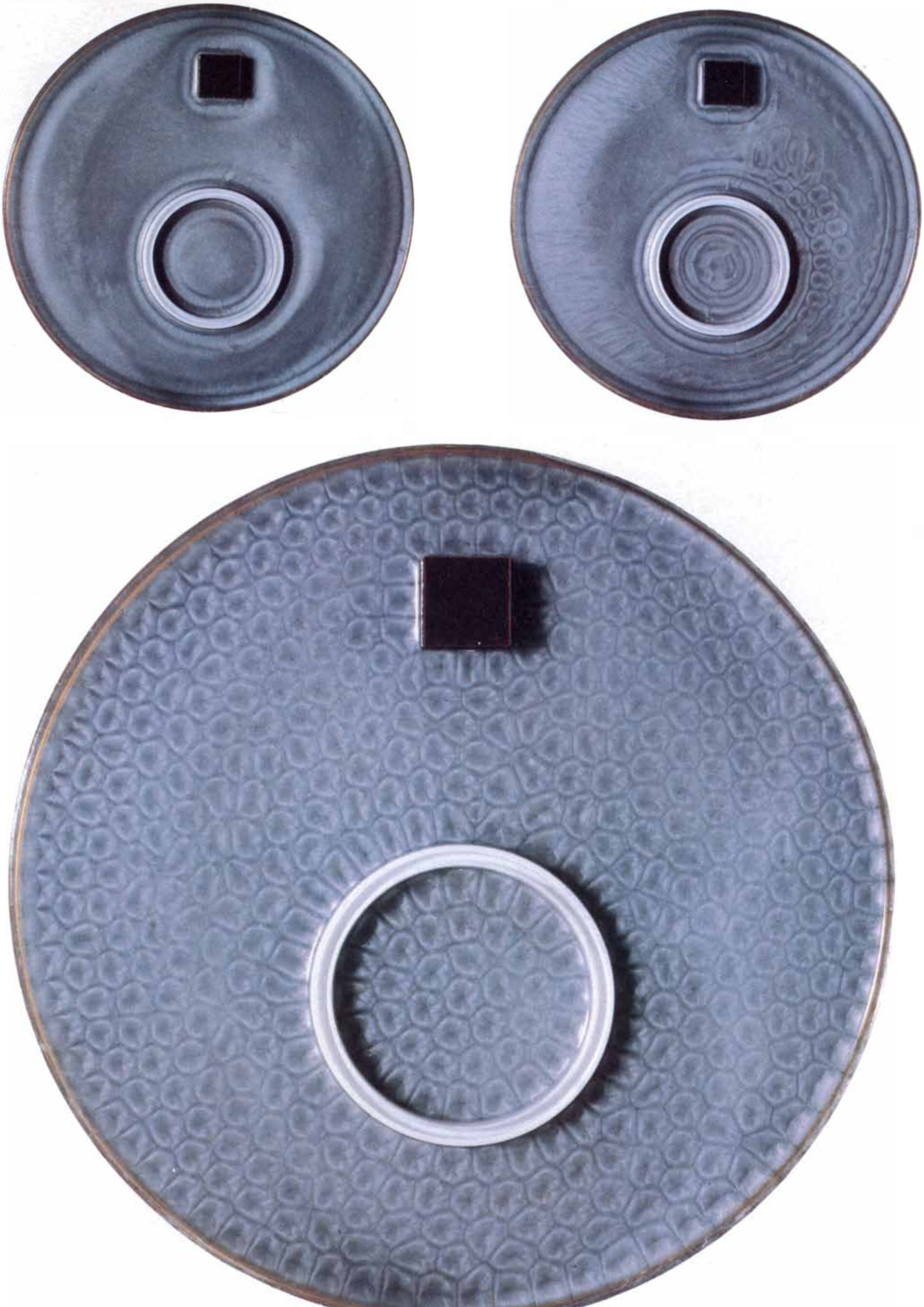
Or do the others just charge you more?

Take a look at the sticker (not this one—you'll strain your eyes!). For $5,679,* we give you a lot. Our standard equipment column lists a 2.2 liter overhead cam engine... power assisted front disc brakes... steel belted radial ply tires...electric tachometer, and lots more.

The 1980 Celica ST Sport Coupe. $5,679!

Like we said, it's a lot of car for less.

CELICA ST SPORT COUPE. A LOT OF CAR FOR LESS THAN A LOT OF CARS.



Convection

The spontaneous upwelling of a heated fluid can be understood only by untangling the intricate relations among temperature, viscosity, surface tension and other characteristics of the fluid

by Manuel G. Velarde and Christiane Normand

Convection should be familiar to anyone who has observed the roiling of a heated broth, felt the draft drawn up the flue of a fireplace or noticed the shimmering of air currents over a paved road on a sunny day. The same mechanism of convective flow is responsible for the great ocean currents and for the global circulation of the atmosphere; it gives rise to motions on a still larger scale in the sun's photosphere. Certain kinds of clouds are formed when warm, moist air ascends in a convective plume, and it is a disruption of normal convective transport that periodically leaves cities such as Los Angeles and Madrid smogbound under a temperature inversion. Other examples are less familiar or less easily observed. Convection has an important influence on the drying of paint films and on the dispersal of gases and particles in the lungs. Convection in the earth's mantle is apparently the motive force for the slow migration of the continents.

The most elementary kinds of convection would seem to have a simple explanation, which can be summed up in the expression "Heat rises." In the simplest cases convective flow begins when a fluid (either a gas or a liquid) is heated from below. In response to the heating the bottom layer of the fluid expands and thereby becomes less dense than the overlying layers. The warmer and lighter bottom layer then tends to rise and the cooler top layer tends to sink. This much at least was known by the 18th century. It may therefore come as a surprise that the formulation of a detailed and quantitative account of convection has proved a lasting challenge to the ingenuity of theorists. Indeed, even the simplest system undergoing vigorous convective motion cannot yet be given an exact mathematical description.

The nature of the theoretical difficulties can be suggested by considering again the case of a fluid layer heated from below. The force that drives the convective flow in such a fluid is the buoyancy of the heated layer, and the magnitude of this force is determined by the difference in temperature between the top and the bottom of the layer. The complexity of the matter becomes apparent when one recognizes that the temperature distribution is greatly altered by the convective flow itself, which carries heat from the bottom of the layer to the top. Thus the force that drives the flow is subject to modification by the flow.

Although exact solutions to problems such as this one are still lacking, substantial progress toward a general theory of convection has been made in the past two decades. The advances have come about in large part through the adaptation of ideas and mathematical techniques developed in other realms of physical science, most notably the study of phase transitions, of ferromagnetic materials and of superconductors. These methods lead to an analysis of the stability of various modes of motion in the fluid and hence to predictions of which modes are the most likely to be observed. The results are only approximate, but in some instances the approximation is a close one. It offers some hope of explaining what goes on in a simmering broth.

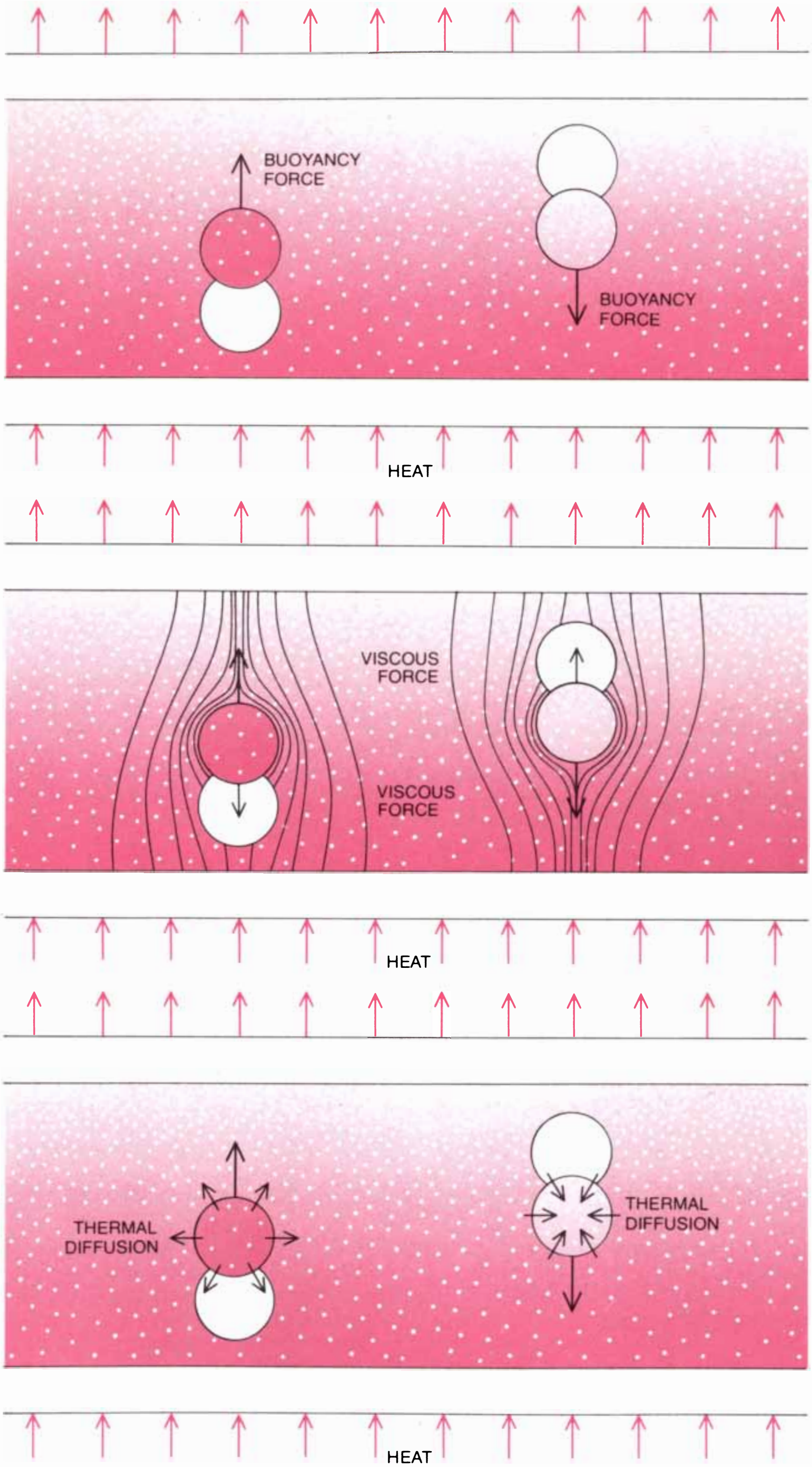
CONVECTION CELLS with a characteristic polygonal geometry arise spontaneously when a thin layer of fluid is heated from below. The evolution of the pattern over the course of a few hours is traced in the photographs at the top of the opposite page; the bottom photograph shows the fully developed pattern. Initially the cells are long "rolls" that follow the boundaries of the layer; the rolls give way to polygons, which tend toward a regular hexagonal form and are independent of the boundaries. In each cell the fluid rises in the center and sinks at the periphery. The circulation is driven largely by forces associated with surface tension, and in most fluids it assumes this form only when the upper surface is free. The photographs were made in the laboratory of one of the authors (Velarde) at the Autonomous University of Madrid. The fluid is a silicone oil in which flakes of aluminum have been suspended to make the flow visible.

The kind of convective transport we shall discuss here is called natural or free convection, meaning that the flow is a response to forces acting within the body of the fluid. The force is most often gravity, but there are circumstances where some other agency, such as surface tension or an electromagnetic field, plays a significant or even a primary role. Natural convection is defined in contradistinction to forced convection, where the fluid motion is induced by a force imposed externally, for example by a pump or a fan.

One of the earliest descriptions of natural convection was written in the 1790's by Benjamin Thompson, Count Rumford. He introduced the idea to account for the transport of heat in an apple pie. There had been earlier proposals of a convective mechanism for atmospheric circulation, and a number of anecdotal reports were added to the literature throughout the 19th century. It was not until about 1900, however, that systematic investigations were undertaken. The most influential experimental work in that era was done by the French investigator Henri Bénard. He studied a convective system that was more complicated than he knew, and its true nature was only recently recognized. Bénard's observations and their modern interpretation will be taken up below.

The preeminent theorist of convection in the early 20th century was John William Strutt, Lord Rayleigh. Among his last works was an article on convection, published in 1916, that attempted to explain Bénard's results. It is now known that Lord Rayleigh's theory does not apply to the system examined by Bénard; nevertheless, Lord Rayleigh's work is the starting point for almost all modern theories of convection.

Lord Rayleigh's theory can be explained in the context of a model experiment that employs a fluid with somewhat simpler properties than any real gas or liquid. A thin layer of the fluid is confined between two flat, rigid, horizontal plates and completely fills the space between them, so that there is no free surface. What is meant by a thin layer is



IMBALANCE OF FORCES is needed to initiate a convective flow. The forces can be analyzed in a hypothetical experiment where a thin layer of fluid between two rigid plates is heated from below, generating a gradient in temperature and density. If a parcel of warm fluid from near the bottom of the layer is displaced upward slightly, it enters a region of greater average density and is therefore subject to an upward buoyant force. Similarly, if a parcel of cool fluid from near the top of the layer is displaced downward, it becomes heavier than its surroundings and tends to sink. The buoyancy force is opposed by viscous drag and by heat diffusion, which tends to equalize the temperature of a displaced parcel and its environment. The relative importance of these influences is measured by the dimensionless ratio called the Rayleigh number. Convection begins when buoyancy overcomes the dissipative effects of viscous drag and heat diffusion, or in other words when the Rayleigh number exceeds a critical value.

that the horizontal extent of the layer is much greater than its depth (which is equal to the separation between the plates). The assumption that the layer is thin is intended to minimize the influence of the boundaries at the edges of the plates, which do not explicitly enter into the theoretical description. Ideally the layer would be of infinite horizontal extent; in practice a layer several centimeters across and a few millimeters deep fulfills the requirements.

The apparatus is to be heated from below in such a way that the bottom of the layer has a constant and uniform temperature. Similarly, heat is withdrawn at the top of the layer in such a way that the temperature there is also constant and uniform but is lower than at the bottom. It follows, of course, that the temperature difference between the top and the bottom is also constant and uniform. Furthermore, the temperature gradient—the change in temperature corresponding to a given change in vertical position—is required to be linear, meaning that a graph of temperature v. height is a straight line.

Just a few more simplifying assumptions must be added to this list. One is that gravity is the only force acting within the fluid. Since practical experiments are of modest scale, the gravitational field will be essentially uniform throughout the volume. The fluid must also be incompressible; for liquids, if the depth is not great, that is a realistic assumption. Finally, a most important feature of the model is that only one property of the fluid is directly altered by the variation in temperature. This property is the density, which declines as the temperature rises. In other words, the fluid expands when it is heated, which is the usual relation for real gases and liquids.

There is a well-defined theoretical method for investigating the effects of the temperature gradient in this hypothetical experiment. First one is to imagine that a parcel of fluid has been displaced upward or downward from its original position; then the forces acting on the displaced parcel are to be analyzed. It is these forces that will determine all subsequent motions in the fluid. The parcel can have any size and shape, but the displacement must be small. (Formally, the Rayleigh theory holds only for displacements that are infinitesimal, or of vanishingly small magnitude.) The initial displacement need not be a response to any imposed force; since the molecules of a fluid are constantly in motion, their positions fluctuate randomly, and any small displacement can be expected to occur spontaneously if one waits long enough.

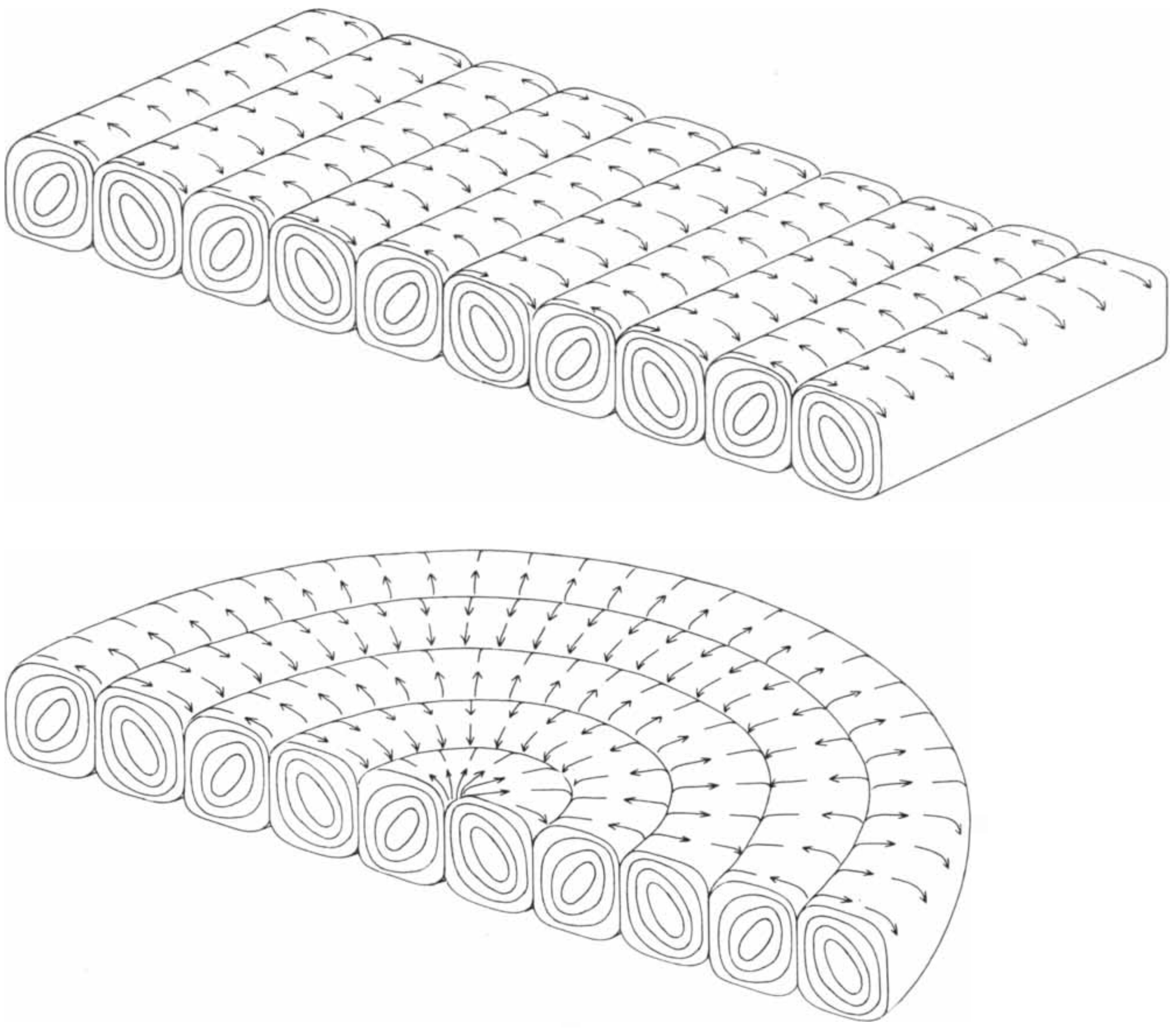
Consider a small parcel of fluid near the bottom of the layer. Because of the elevated temperature at the bottom the parcel has a density that is less than the average density of the entire layer. As long as the parcel remains in place, however, it is surrounded by fluid of the same density, and so it has neutral buoyancy. All forces acting on it are in balance, and it neither rises nor sinks.

Suppose now that through some random perturbation the parcel of fluid is given a slight upward motion. What effect does the displacement have on the balance of forces? The parcel now is surrounded by cooler and denser fluid. As a result it has positive buoyancy, so that it tends to rise. The net upward force is proportional to the density difference and to the volume of the parcel. Thus an initial upward displacement of the warm fluid is amplified by the density gradient, and the amplification gives rise to forces that cause further upward movement. A similar analysis could be made for a slight downward displacement of a parcel of cool, dense fluid near the top of the layer. On moving downward the parcel would enter an environment of lower average density, and so the parcel would become heavier than its surroundings. It would therefore tend to sink, amplifying the initial perturbation. Natural convection is the result of these combined upward and downward flows, and it tends to overturn the entire layer of fluid.

The implication of this analysis is that convection should be observed in the fluid whenever there is a temperature gradient, no matter how small it is. Even in an infinitesimal gradient any random upward motion of the warm fluid or downward motion of the cool fluid would seem to be enough to establish a flow. Actually such extreme sensitivity is not observed; instead the temperature gradient must reach a certain threshold value before the convective flow begins. It was Lord Rayleigh's most significant contribution to explain why this is so.

Lord Rayleigh pointed out that a theory of convection must take into account at least two other influences on the motion of a fluid particle. One of these is viscous drag, the fluid equivalent of friction. The drag force is always directed opposite to the motion of the fluid, and its magnitude is determined in part by an intrinsic property of the fluid, the shear viscosity, which measures the resistance to relative motion of any two adjacent regions. If the velocity is not too great, the magnitude of the drag force is proportional to the viscosity multiplied by the radius of the parcel and the velocity. Clearly if the viscous drag is equal to the buoyant force, there can be no motion.

The second disruptive influence con-



ROLL-SHAPED CELLS are a stable configuration in convection driven by buoyancy forces rather than by surface tension. The fundamental unit of the pattern consists of two rolls that rotate in opposite directions; the width of this unit is twice the depth of the fluid layer. The planform of the pattern depends strongly on the boundaries of the layer. In a rectangular container the rolls are parallel to the shorter sides; in a circular container they form concentric rings. A stable roll pattern is usually observed only when the fluid has no free surface.

stitutes an adjustment for the fact that convection is not the only mode of heat transport in a fluid. Radiation and conduction, or heat diffusion, can also carry heat. At the comparatively low temperatures of most convection experiments radiation makes such a small contribution that it can be ignored. Heat diffusion, however, is not always negligible; it tends to dissipate the temperature gradient that drives the convective flow.

The effect of heat diffusion can be explained by considering again a parcel of warm fluid displaced upward from its equilibrium position into a cooler environment. According to the fundamental definition of heat, the molecules in the warm parcel must have a higher average speed than those in the surrounding cooler fluid. Molecules can freely cross the boundary that defines the parcel, and the effect of many such exchanges in both directions will be to equalize the average speeds of the two populations. In other words, heat will flow out of the displaced warm parcel, so that the parcel cools off and the surrounding fluid warms up, until eventually they reach equilibrium at the same temperature. For a parcel of cool fluid displaced downward the heat flow is in the opposite direction: from the warm environment into the cool parcel. In either case as the local temperature difference is reduced so is the buoyant force derived from it.

The time required for a fluid parcel to reach thermal equilibrium with its surroundings depends in part on a property of the fluid, the thermal diffusivity. The time scale of this process is inversely proportional to the diffusivity constant and directly proportional to the surface area of the parcel. If this thermal-diffusion time is comparable to the time required for the parcel to move some characteristic distance, such as its own diameter, the buoyant force will be abolished. In other words, if the fluid moves no faster than it loses heat through diffusion, a convective flow cannot be sustained. The heat input to the system through the bottom plate is then carried through the fluid layer by a purely conductive or diffusive mechanism, without any bulk motion.

Lord Rayleigh's analysis shows that the mere existence of a temperature gradient is not enough to ensure the onset of convective flow. It is necessary for the buoyancy resulting from this gradient to exceed the dissipative effects of viscous drag and heat diffusion. The gravitational potential energy liberated by the sinking of denser fluid and the rising of lighter fluid must be greater than the energy dissipated by drag and diffusion. The relation of these effects can be expressed as a dimensionless ratio: the buoyant force divided by the product of the viscous drag and the rate of heat diffusion. The ratio is dimensionless in that all the units of measurement associated with the three quantities cancel exactly, leaving a pure number whose value is the same no matter what system of units is adopted. The ratio is now called the Rayleigh number. Convection begins when the Rayleigh number exceeds a critical value.

The meaning of the Rayleigh number can be stated more precisely through an examination of the stability of various possible modes of motion in the fluid. It is convenient to define stability in terms of a potential curve or potential surface, which gives the energy of a system as a function of some variable. The system is usually found in whatever state has the minimum energy, which corresponds to the lowest point on the potential surface.

It is easy enough to imagine a tangible model of a potential surface, namely a hemispherical bowl with a marble in it. In equilibrium the marble remains at rest at the bottom of the bowl, where its gravitational potential energy is lowest. If some random perturbation then displaces the marble slightly, it will roll back toward its equilibrium position; perhaps it will overshoot the bottom of the surface and oscillate about it, but the dissipative effects of friction will eventually damp out the oscillations and at some future time the marble will again be found at rest at the point of minimum energy. Because a marble at the bottom of the bowl returns to its original position after a perturbation it is said to be in stable equilibrium.

Another model potential surface is formed by turning the hemispherical bowl upside down and carefully balancing the marble at its apex. This is also a state of equilibrium in that all forces acting on the marble are in balance, and in the absence of any disturbance the marble might remain motionless indefinitely. In practice, however, some external influence (say a breeze or a passing truck) will always disturb this precarious balance if the experimenter waits long enough. After such a perturbation the marble does not return to the point of equilibrium but instead departs farther and farther from it. No matter how small the initial disturbance is, the marble will eventually be found at a large distance from the center. The perturbation is amplified, and so the state of equilibrium is said to be unstable.

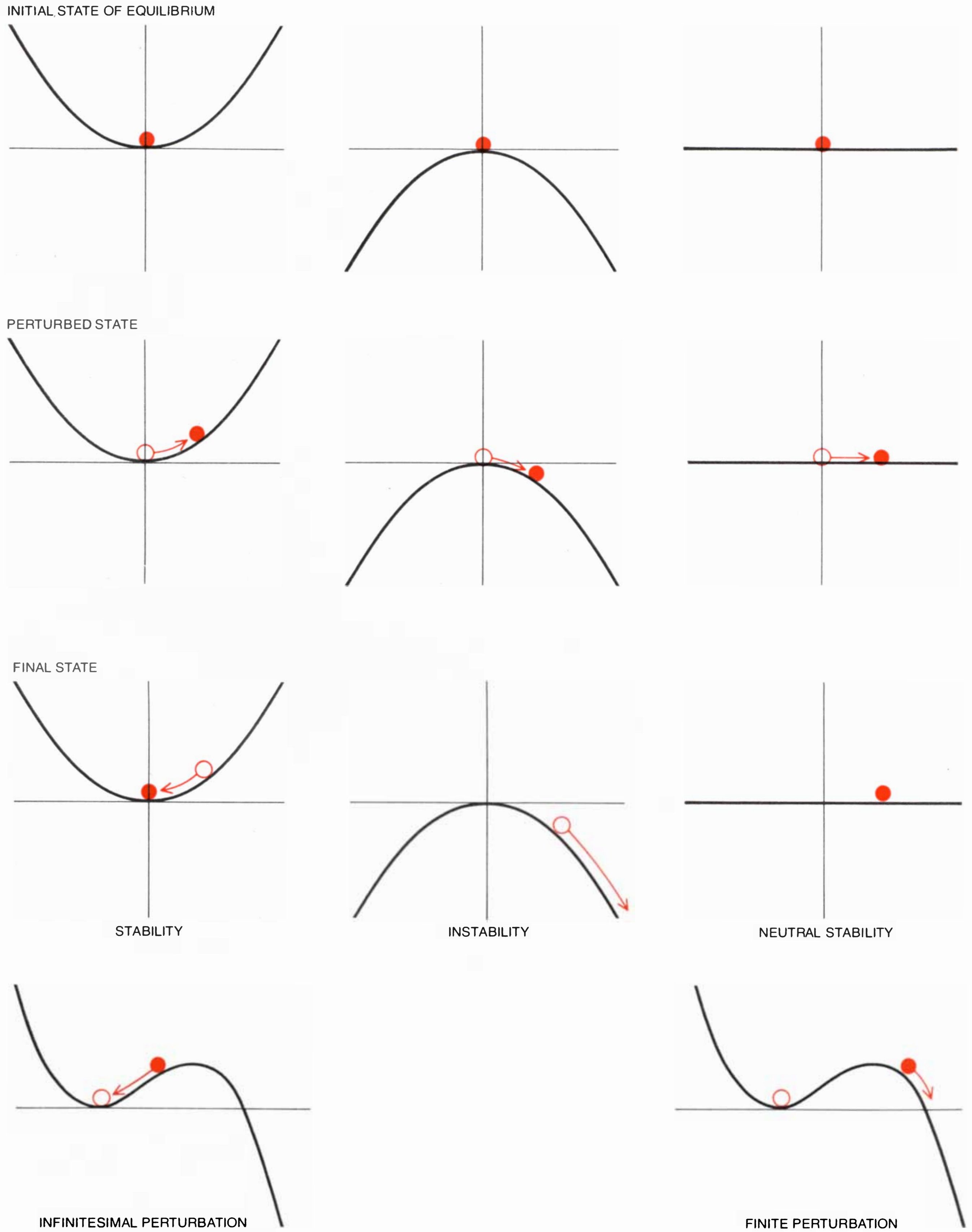
There is a third possibility: the marble might also be placed on a flat surface. In this case when the marble is moved, it neither returns to its original position nor continues moving farther away. It simply remains at equilibrium in its new position. On a flat potential surface every point is said to represent a state of neutral or marginal stability.

A little further thought devoted to this model potential surface suggests that the absolute stability of a system can be demonstrated only by testing its response to all possible perturbations. For example, a marble inside the bowl will return to the center after an infinitesimal disturbance or even after a small but finite one. It will not return to its starting point, however, if the perturbation is large enough to propel it entirely out of the bowl. Because an infinite number of possible disturbances would have to be checked, it is difficult to prove that a state of equilibrium is stable; instability, on the other hand, can be demonstrated by exhibiting a single perturbation that grows spontaneously.

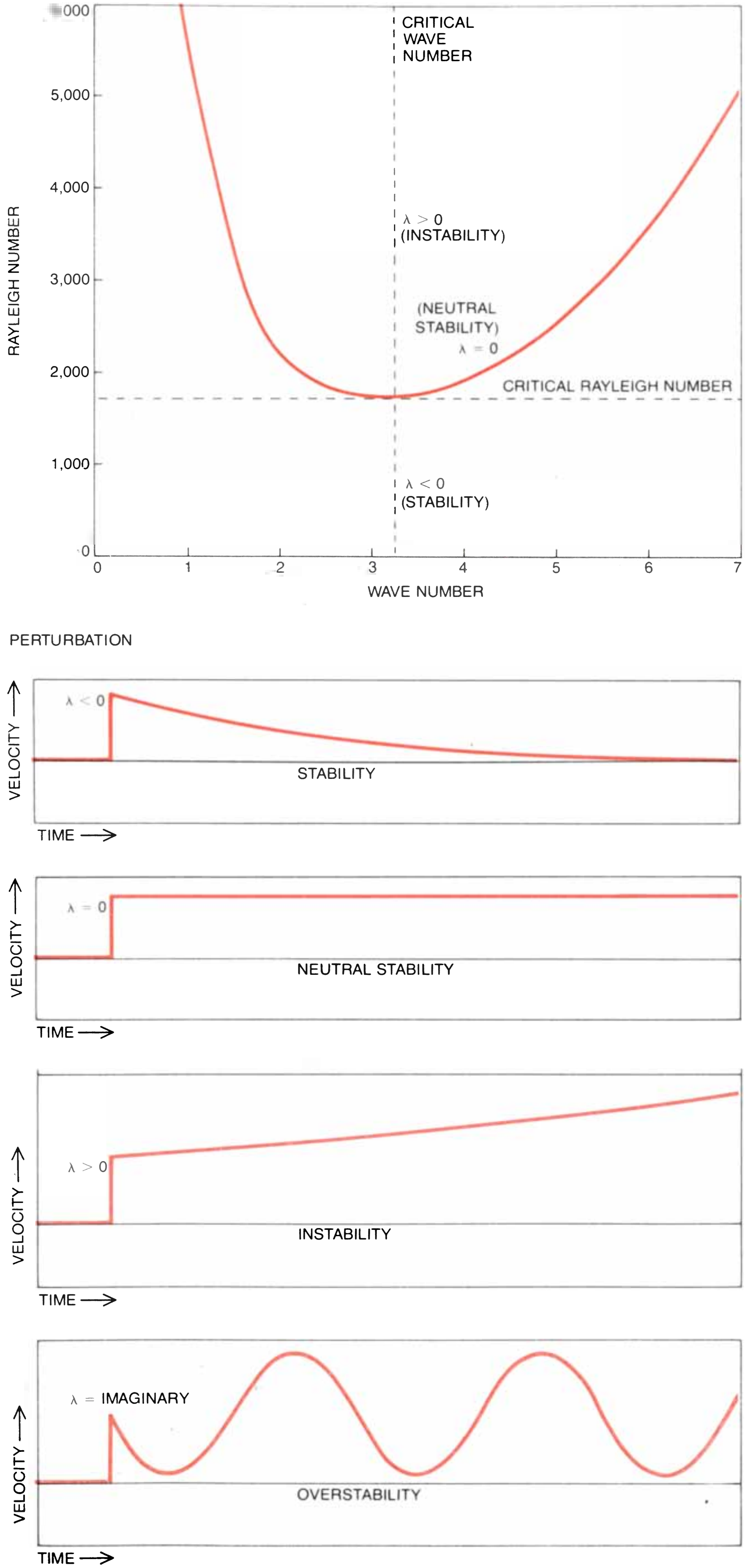
The application of these ideas to the convection problem is straightforward. It is easily shown that a motionless fluid layer heated uniformly from below is in an equilibrium state, even if viscous damping and heat diffusion are neglected. Although the lighter portion of the fluid is overlain by denser material, so that the gravitational potential energy could be reduced by interchanging these positions, in the absence of perturbations all the forces acting on any given parcel of fluid are in balance. The question a theory of convection must answer is whether the equilibrium is stable or unstable or exhibits neutral stability. In other words, the theory must define the shape of the potential surface.

It is the value of the Rayleigh number that determines the curvature of the potential surface. If the Rayleigh number is zero, say because the temperature gradient and the buoyant force are zero, the motionless state is obviously stable and the potential surface is concave upward, like the inside of a bowl or a trough. Setting the fluid in motion requires an increase in its energy. If the Rayleigh number is very large, on the other hand, so that buoyancy overwhelms all dissipative effects, the fluid can reduce its total energy by establishing a convective flow. Any perturbation of the stationary equilibrium will then be amplified. Thus the potential surface slopes downward like an inverted bowl.

Continuity demands that there be some value of the Rayleigh number intermediate between these extremes where buoyancy and the dissipative forces are of equal magnitude. This is the critical Rayleigh number, and it designates the flat, transitional potential surface of neutral stability. As the Rayleigh number increases from zero (because of an increasing temperature gradient, for example) the potential surface starts out concave and gradually flattens; at the critical Rayleigh number it is perfectly flat, and as the number continues increasing the surface becomes convex. Not until the critical value is exceeded does the motionless equilibrium become unstable. For the model experiment considered here, calculations indicate a critical Rayleigh number of 1,708. In a typical laboratory experiment, where the fluid is a layer of silicone oil a few millimeters thick, the crit-



STABILITY OF A PHYSICAL SYSTEM can be judged from its response to an arbitrary perturbation, such as a small displacement of a marble at rest on a surface. If the surface is concave, the marble eventually returns to its equilibrium position at the bottom of the bowl, and so that position is said to be one of stable equilibrium. On a convex surface the marble can be balanced at the apex, but the equilibrium there is unstable: the slightest disturbance is amplified as the marble reduces its potential energy by rolling downhill. On a flat surface the marble neither returns to its starting position nor moves farther away from it; the surface exhibits neutral stability. A system can be stable with respect to some perturbations but not others, as in the case of a surface with both concave and convex regions. In a simple fluid a necessary condition for the onset of convection is an unstable distribution of some property such as density or surface tension.



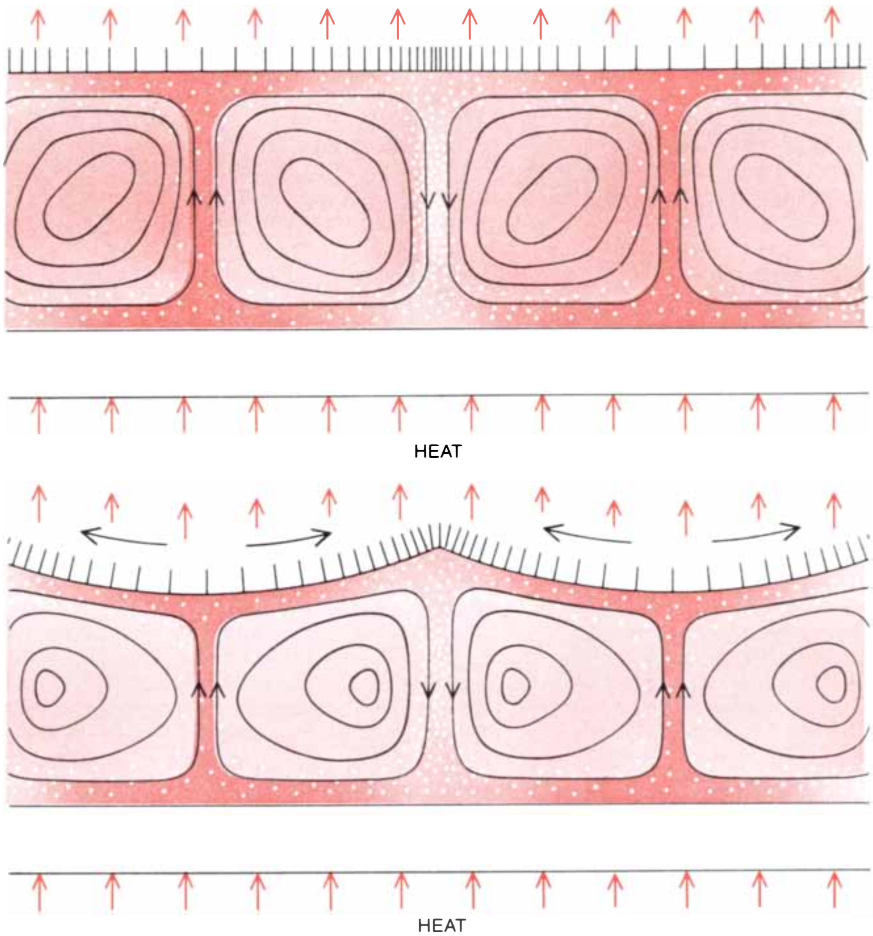
ical Rayleigh number is reached when the temperature gradient is a few degrees Celsius.

The hypothetical experiment this account of Lord Rayleigh's theory is based on includes many simplifying assumptions, some of which are known to be contrary to fact. The theory is nonetheless remarkably successful in predicting the conditions necessary for the onset of convection in real fluids. For example, experiments by Peter L. Silveston of the University of British Columbia and by Ernest L. Koschmieder of the University of Texas at Austin have yielded values for the critical Rayleigh number of 1,700 ± 50, in good agreement with the theoretical value.

The balance of buoyant and dissipative forces provides a criterion for the onset of convective instability, but what is observed in the model experiment once the flow is under way? The Rayleigh theory supplies only limited guidance in answering this question, and even the more comprehensive theories to be discussed below cannot account for all the observed features of fully developed convective circulation. If the evolution of the flow cannot be deduced mathematically, however, at least a qualitative description is possible.

In a fluid layer heated uniformly from below, the temperature gradient should be independent of horizontal position and so should the resulting buoyant force. When the critical Rayleigh number is exceeded and the motionless equilibrium becomes unstable, the warm fluid therefore tends to rise everywhere and the cool fluid tends to sink everywhere. Both things obviously cannot happen at once: at any point the fluid can be ascending or descending, but it cannot move in both directions at the same place and time. This impasse is avoided by the spontaneous division of the layer into a pattern of convection "cells," in each of which the fluid circulates in a closed orbit.

STABILITY OF A FLUID with respect to buoyancy-induced convection is determined by the value of the Rayleigh number. If the number is subcritical, any perturbation is damped out; if the number exceeds the critical value, a perturbation grows continuously. The system is most sensitive to perturbations with a particular wave number, or scale of length, corresponding to twice the depth of the fluid layer. The difference between the actual Rayleigh number and the critical number determines the value of a parameter, λ; in the theory devised by Lord Rayleigh the velocity of the convective flow depends exponentially on λ. If λ is negative, the velocity decays to zero; if λ is equal to zero, the velocity remains steady; if λ is positive, the velocity grows continuously. Because λ is a complex number and can have an imaginary value the flow can oscillate, a phenomenon called overstability.



VARIATIONS IN SURFACE TENSION alter the pattern of convective transport in a fluid with a free surface. The magnitude of the surface tension varies with temperature, being greatest where the fluid is coolest. Hence any difference in temperature across the surface gives rise to a gradient in surface tension. The motionless state becomes unstable if the gradient is large enough to overcome the dissipative effects of viscosity and heat diffusion. The magnitude of the tension is represented here by the density of the hatching. Fluid is pulled along the surface toward cooler regions of greater surface tension and is replaced by warmer fluid from below.

Theoretical arguments give a tentative indication of the favored scale of the individual elements in the convection pattern. These arguments are derived from the varying sensitivity of the marginally stable state to perturbations with different scales of length. One must be careful here to avoid confusing the amplitude of a perturbation, which in the model experiment corresponds to the vertical displacement of a fluid parcel, with the scale of the perturbation, which measures the size of the parcel. If the Rayleigh theory is to yield meaningful results, the amplitude must always be infinitesimal, whereas the scale can be as large as the dimensions of the apparatus will allow.

It is customary to express the scale of a perturbation in terms of a wave number, which is the reciprocal of a length. This practice reflects the fact that the geometry of a disturbance is generally complicated, so that it does not have a single, clearly defined size; the disturbance can be decomposed, however, into a spectrum of fundamental modes, or spatial frequencies, just as a complex sound can be analyzed into a combination of pure tones. A wave number represents the contributions of a particular scale of length to the random fluctuations. Larger wave numbers correspond to smaller scales.

The stability of the motionless steady state is vulnerable to being upset by disturbances at some wave numbers more than it is at others. One can imagine an experiment in which the critical Rayleigh number is measured in a fluid whose fluctuations can be controlled in such a way that they are always characterized by a single wave number. Such an experiment would reveal that instability sets in soonest when the wave number describes perturbations with a horizontal dimension of about twice the fluid depth. For wave numbers that are either larger or smaller more extreme conditions (a larger Rayleigh number) are required to induce convection. The calculated value of 1,708 for the critical Rayleigh number is the value found when the fluctuations are of the optimum size.

The sensitivity of the fluid to disturbances at a particular scale implies that those disturbances will be amplified sooner than any others when the layer becomes unstable. Therefore the pattern observed when convection begins might be expected to have features at about this scale. It is not obvious that those features will persist once the convective flow is fully developed, but it turns out they do, provided the Rayleigh number does not exceed the critical value by too large a margin.

The wave number specifies the overall scale of the pattern but not its detailed form; convection cells of many different shapes could be constructed so that they had the same wave number. The pattern actually observed depends strongly on the geometry of the experimental apparatus. The pattern cannot now be deduced from first principles, but there are well-established empirical rules that make reliable qualitative predictions.

In experiments such as the hypothetical one described here, where the upper and the lower surfaces are constrained by rigid boundaries, the basic element of the pattern is a "roll," which has a long, tubular form. Warm fluid rises along one edge of the roll, traverses the upper surface and thereby loses its heat, then plunges to the bottom of the layer along the opposite edge. The circulation then carries the fluid across the lower boundary, where its temperature is raised again. Adjacent rolls have an opposite sense of rotation.

When a roll is viewed in cross section, its form is approximately square: the width is equal to the height, and the height, of course, is determined by the depth of the fluid layer. Thus the proportions of the roll are constant, but the size depends on the fluid depth. Because the repeating unit in this pattern consists of two rolls rotating in opposite directions the scale of the pattern is properly defined as the width of two rolls. Hence the scale is equal to twice the fluid depth, in agreement with the predictions of the wave-number analysis.

The planform of the cellular pattern (its appearance from above) is determined largely by the shape of the apparatus. Because details of this kind do not directly enter into the theory the planform cannot readily be derived from it; experimental observation, however, gives an ample basis for prediction. If the container is rectangular, the rolls tend to be aligned parallel to the shorter sides. The width of each roll and hence their number is of course determined by the depth of the layer. In a circular vessel the rolls form concentric rings.

As was pointed out above, Lord Rayleigh's analysis of convection was inspired mainly by experimental observations Bénard made in about 1900. It is now known that the Rayleigh theory is not the appropriate one for the convective mechanism investigated by Bénard. The experimental conditions Bénard employed differed in a subtle but crucial way from those described here, and the importance of the change is immediately apparent in the pattern of the convective flow. In Bénard convection the rolls imposed by the geometry of the apparatus may appear transiently when the flow is just beginning, but they soon give way to a more complicated pattern: a polygonal tessellation of the fluid surface. Initially the polygons are somewhat irregular, having from four to seven sides, although the mean number is six. When the pattern is fully developed, it becomes an almost perfect array of regular hexagons, arranged as in a honeycomb. The center of each hexagonal cell is a region of upwelling warm fluid, which spreads out over the upper surface and sinks at the perimeter, where adjacent cells are joined.

In the Bénard experiments, as in the hypothetical experiment described above, the fluid forms a thin layer that is heated from below. The crucial change is that the layer is not confined between two rigid boundaries but instead is open to the air at its upper surface. Since the surface is free, surface tension can affect the flow; indeed, surface tension is now recognized as the dominant influence in Bénard convection, being more important than the buoyant force. It is therefore no surprise that the Rayleigh theory, in which buoyancy is explicitly assumed to be the only force operating, is inadequate to explain convection of this kind. The predictions of the Rayleigh theory are in error even for such fundamental quantities as the temperature gradient needed to initiate convective flow. A successful alternative theory was not introduced until 1958, by J. R. A. Pearson of the Imperial College of Science and Technology in London.

Surface tension is the cohesive force whose net effect is to minimize the surface area of a fluid. For example, a free drop of liquid tends to assume a spherical form under the influence of surface tension, since that is the configuration of minimum area. The tension can be imagined as a network of stretched elastic bands extending in all directions over the free surface. If at any point the forces exerted by the various bands are not in balance, the surface layer will flow toward the region of greater tension until equilibrium is established. The surface flow is communicated to the bulk of the fluid as a result of its shear viscosity.

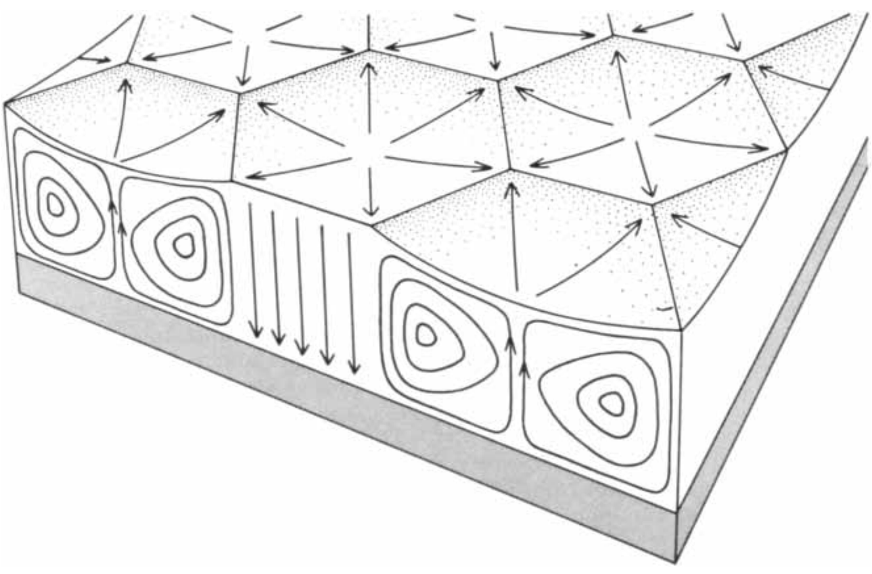
Surface tension can act as the propulsive force in a convective flow because the tension varies with temperature: like density, the surface tension is reduced as the temperature increases. Therefore any temperature gradient established across the surface of the liquid will be accompanied by a gradient in surface tension. The cooler regions will exhibit stronger surface tension and in the warmer regions the tension will be reduced. If the surface-tension gradient leads to an imbalance of forces, a flow will result.

The onset of convective instability in the Bénard system can be analyzed in much the same way that the onset of a buoyancy-induced flow is analyzed. Suppose a parcel of warm liquid is displaced upward through some random fluctuation. Whether or not this motion is sustained by buoyant forces, it will have an effect at the surface of the layer, slightly raising the temperature and thereby reducing the surface tension in the area directly over the fluctuation. Nevertheless, forces at the surface remain in equilibrium because the surrounding surface pulls equally on this region in all directions. In order to initiate the flow a second perturbation is required, causing a horizontal displacement of some small element of the surface in the area of reduced tension. The tensile forces acting on the displaced element are then out of balance, and if the surface-tension gradient is large enough, the displacement will be amplified. The surface element will be pulled into the cooler region of higher tension, and it will drag part of the bulk fluid with it. More liquid will therefore be drawn up from the warmer, underlying strata, reinforcing the surface gradients in temperature and tension. Meanwhile fluid that has cooled during its transit across the surface will begin to sink and the cellular pattern will be established.

As in buoyancy-driven convection, the mere existence of a temperature gradient does not guarantee that a convective flow will be sustained. The gradient must be large enough to overcome the dissipative effects of viscous drag and of heat diffusion. The balance of these effects in the Bénard system is expressed through another dimensionless ratio, in this case named after the 19th-century Italian investigator C. G. M. Marangoni. The formula for the Marangoni number is the same as that for the Rayleigh number except that the buoyant force is replaced by the surface-tension force; in other words, the Marangoni number is the ratio of the surface-tension gradient to the product of viscous drag and heat diffusion. Bénard convection sets in when the Marangoni number exceeds a critical value.

An intriguing feature of convection driven by surface-tension gradients is that it alters the contour of the surface. Regions of enhanced surface tension tend to pucker, so that they reduce their total exposed area. The consequences of this effect are contrary to intuition. In the center of a Bénard cell, where the fluid is rising, the surface is depressed; at the edges of the cell, where the fluid is falling, the surface is raised. Gravitational forces oppose the formation of the surface cusps, since the gravitational potential energy is minimized by a flat surface. Thus the interactions of gravitation and surface tension are subtle and complex. A theory that incorporates both buoyant and surface-tension forces was formulated in 1964 by D. A. Nield of the University of Auckland.

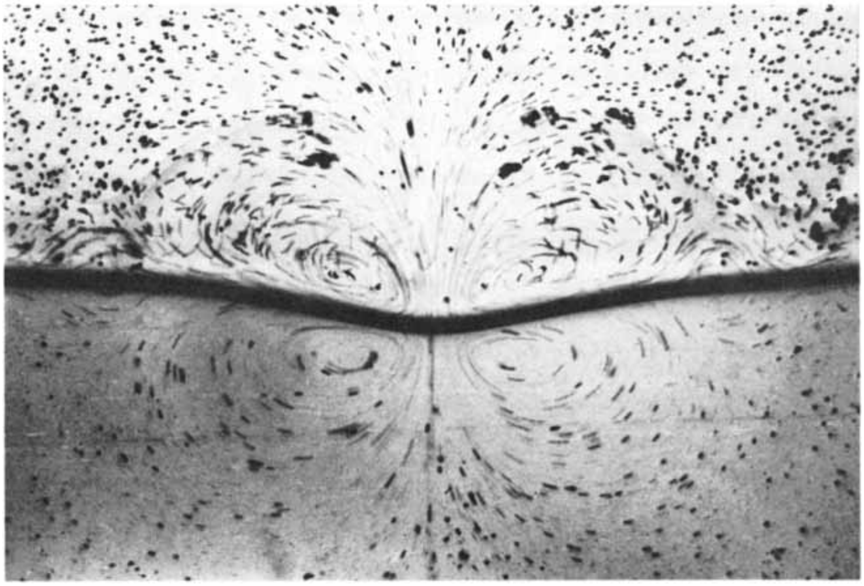
The predominance of surface tension in Bénard convection has now been established beyond doubt. One unequivo-



TESSELLATION OF THE SURFACE with hexagonal cells is a characteristic feature of convection driven by a gradient in surface tension. Where the tension is greatest the surface becomes puckered, so that its area is reduced. Over the ascending plume in the center of each cell the surface is depressed; the fluid must flow uphill before descending at the edge of the cell.

DATSUN ENGINUITY!
NEW DATSUN NAPS-Z ENGINE DELIVERS MORE POWER FROM LESS GAS!
"The target was higher fuel economy, lower exhaust emissions, better driveability and increased horsepower, all at once from the same engine!"—Car & Driver.
It sounds impossible. And to Datsun engineers, the task must have seemed like a wild gas chase. But they did it; 250 patents prove the NAPS-Z (Nissan Anti-Pollution System) engine achieves those "incredible performance goals."
How? It uses a new cast-aluminum SOHC layout with crossflow hemispherical combustion chambers. This race-proven hemi-head design allows the engine to breathe more deeply through larger valve openings, and cuts the distance spark-plug flame must travel to insure complete burning of the fuel-air mixture.
"Engine is cleanest burning yet!"—Popular Science.
Recirculated exhaust gases help keep this special "quick-burn" system from forming more than the slightest trace of NOx emissions. "The payoff," says Car & Driver, "is 9% more power, roughly 20% higher fuel economy, fewer noxious nasties out the tailpipe, and an absolute minimum of the spitting and surging typical of emissions-controlled engines."
"The system offers the added benefits of smoother acceleration and an increase in total operating efficiency."—Motor Trend.
"The added power has been put to good use," according to Road & Track. The fuel-injected version of the NAPS-Z powers the 1980 Datsun 200-SX to "60 mph in 11.6 seconds and through the quarter-mile in 18.8 seconds"—seconds faster in both respects than last year's 200-SX.
Tests by the USAC scored acceleration times (in seconds) of: 0 to 50 mph in 8.4; 0 to 55 mph in 9.9; and the quarter-mile in 18.2.
But you must experience this new engine for yourself. Visit your Datsun dealer today and test-drive a Datsun 200-SX Sport Coupe or the carbureted NAPS-Z in a Datsun 510 Hatchback, Sedan or Wagon. And be sure to ask about air conditioning at its best: Datsun's "A-System." It's another example of Datsun "enginuity."
Two spark plugs meet stringent Calif. emissions standards. Other states use one.
DATSUN
WE ARE DRIVEN
NISSAN
1980 Datsun 200-SX
28 EST. MPG 40 EST. HWY.
31 EST. MPG 43 EST. HWY.
1980 Datsun 510 Hatchback
EPA estimates for comparison. Standard 5-speed manual.
Actual mileage may differ depending on speed, trip length and weather. Actual highway mpg will probably be less. California mileage lower.

cal demonstration is that the characteristic hexagonal convection cells appear even when the fluid layer is heated from above rather than from below. Under those circumstances the density gradient opposes convective flow, so that the forces resulting from the surface-tension gradient must overcome it. Convective flow attributed to surface tension was also observed in experiments conducted during two of the Apollo spaceflight missions, where gravity and buoyancy were negligible.



ASCENDING PLUME of warm fluid creates a gradient in surface tension along the interface between two liquids. The depression in the surface marks the point of minimum tension, and fluid is drawn along the interface to the right and left away from this point. The circulation of the bulk fluid is driven by the surface traction. A similar mechanism operates in a fluid whose surface is open to the air. The photograph was made by H. Linde of the Central Institute for Physical Chemistry in Berlin. Particles of plastic suspended in the liquids make the flow visible.

The Rayleigh theory and theories modeled on it give approximations of the conditions required for the onset of convection. What happens once the flow has been established? In describing such fully developed convective motions theories of this kind are much less satisfactory.

In the Rayleigh theory the velocity of the flow is given by an exponential function: the velocity is proportional to Euler's number, or e (which has a numerical value of about 2.7), raised to some power. The exponent is equal to the time, t (measured in seconds, say, from some moment of reference), multiplied by a coefficient, λ, determined by the Rayleigh number. Hence the velocity is proportional to $e^{\lambda t}$. The predictions of the theory can be catalogued by observing the evolution of this expression for different values of λ.

If the Rayleigh number is less than the critical Rayleigh number, λ is negative. As time passes, therefore, the velocity is given by progressively larger negative powers of e; under these conditions the value of the exponential approaches zero as t increases. In other words, the velocity decays to zero, and any random motions of the fluid are damped out. When the Rayleigh number is exactly equal to the critical value, λ is zero, and so the exponent λt is zero for all time. Any number raised to the zero power is equal to 1, which implies that the disturbance is neither damped nor accelerated but retains its initial value forever.

Both of these predictions are in accord with intuition and with the analysis of the fluid layer's stability. A negative value of λ corresponds to the stable motionless state; the value $\lambda = 0$ specifies marginal stability. The interpretation of the theory becomes problematical, however, when the Rayleigh number is greater than the critical value and λ is positive. This is just the condition under which a convective flow can be established.

When λ is greater than zero, the power to which e is raised increases continuously with time, and the expression displays the well-known characteristics of exponential growth. If λ is equal to $+1$ and the initial velocity is one centimeter per second, then after one second the speed will have increased to 2.7 centimeters per second and after two seconds it will have reached 7.4 centimeters per second. The velocity grows without limit, which quickly leads to absurd predictions; for example, the predicted convection current reaches the speed of light in less than half a minute.

The dependence of the flow velocity on the value of λ has been presented here in a simplified form. In general λ is a complex number, one having both a real part and an imaginary part; the latter includes as a factor the square root of -1. We have considered so far only the variation in the real part of λ. If the imaginary part is not zero, oscillatory flows can arise, a condition called overstability. Such oscillating currents have been observed in real fluids, and they make up an interesting subclass of convective phenomena. In the model described by the Rayleigh theory, however, the imaginary part of λ vanishes and the problem of exponential growth must be confronted directly.

The increase in the speed of a convective flow obviously cannot continue very far along the exponential curve. For this reason the predictions of the Rayleigh theory can be considered realistic only when the Rayleigh number is near the critical value (so that λ is small) or only for a short interval after the onset of convection (so that t is small). The physical source of these limitations lies in the simplifying assumptions that were adopted in developing the theory. In particular the temperature gradient was assumed to be constant and therefore unaffected by the convective circulation. This assumption is clearly contrary to fact: when hot fluid rises into the cooler upper part of the layer, the temperature difference between the upper and lower boundaries is reduced. The buoyancy force is reduced in proportion, so that the flow is self-limiting. This mechanism of self-limitation, however, does not appear in the mathematics of the Rayleigh theory. There the temperature gradient remains constant no matter how rapid the flow is, and the buoyancy force supplies the impetus for an acceleration that continues indefinitely.

Given this failure, it may seem surprising that the Rayleigh theory yields acceptable results under any circumstances. It can do so only because it includes another assumption, namely that the parcel of fluid is never given more than an infinitesimal displacement. If this condition is met, the assumption of an unvarying temperature gradient is quite reasonable. Even a finite but small motion of the parcel can cause only a small disturbance of the temperature distribution, so that the predictions of the theory are still approximately valid. In applying the Rayleigh theory to a fully developed flow, however, the assumptions are violated and the predicted evolution of the motion ends in an implausible infinity.

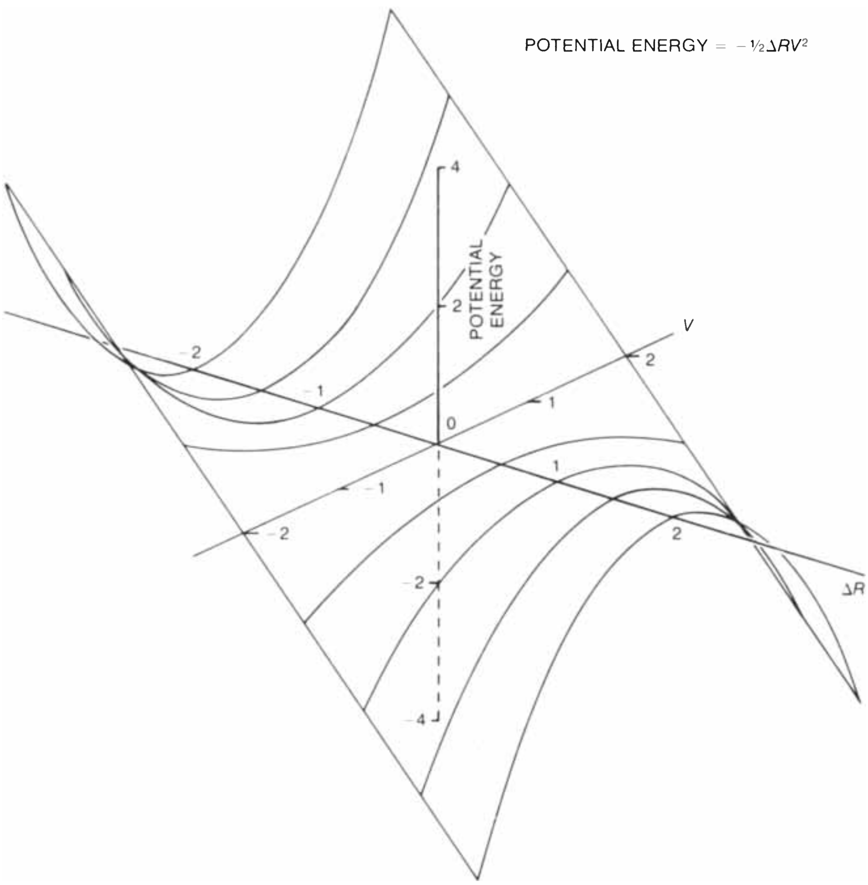
A workable theory of well-established convection must somehow recognize the feedback loop through which the flow itself alters the force that drives the flow. There is no known practical method for solving the problem exactly, but there are approximations that do better than the Rayleigh theory. The one we shall discuss is based on ideas introduced in 1937 by the Russian theorist L. D. Landau, who developed his theory in order to describe certain kinds of phase transitions, such as the onset of magnetization in ferromagnetic iron. With V. L. Ginzburg, another Russian physicist, he later adapted it to the description of superconductivity in metals. These phenomena have certain features in common with convection; most notably, they require that fluctuations in structure at many scales of length be described simultaneously. When the Landau theory is adapted to the convection problem, it incorporates the Rayleigh theory as a first approximation.

What the theory must provide is an equation of motion for the fluid, an equation that gives the velocity and acceleration of a parcel of fluid for any specified combination of externally imposed conditions. The equation of motion might be constructed directly, but the meaning of the Landau theory is clearer if instead one defines a potential surface, from which the equation of motion can be derived. (The equation of motion is simply the equation that gives the slope of the potential surface.)

The potential surface can be conceived of as an undulating landscape, where height above or below some reference plane represents the relative energy of the fluid system. The tendency of the energy to assume a minimum value implies that the point representing the state of the system will "roll downhill" whenever possible. An axis on the reference plane defines the line of zero velocity; displacements to the right and left of this line denote increasing positive (upward) and negative (downward) velocity of some parcel of fluid. Position along the zero-velocity axis is given by the difference between the actual value of the Rayleigh number and the critical value; this difference can be designated ΔR. Thus the value of ΔR and the velocity, V, together specify a point on the reference plane; the height of the potential surface at the point is the energy of the system in that state.

It is necessary to discuss in some detail the equation that specifies the topography of the potential surface. The equation can be written as the sum of an infinite series of terms, each term in the sequence including a higher power of the fluid velocity. The first term is a quadratic one: $-\frac{1}{2}\Delta R V^2$. In the next term the velocity enters as V^3, in the next as V^4, and so on. Each of these higher powers of the velocity is to be preceded by a coefficient measuring its contribution to the form of the surface.

As might be expected, it is impractical to calculate the sum of the entire infinite series; even if formal methods for doing so were available, the coefficients are not known for all the terms. In general, however, the coefficients are expected to become smaller as the power to which V is raised grows larger. Hence there is some hope of obtaining reasonable accuracy even from a truncated series, one in which all terms beyond a certain power of V are neglected. If the velocity is not too large, the contributions of those higher terms should be small; in particular, if V is less than 1 in some natural system of units, the higher powers of V will converge on zero and the approximation will be a good one.



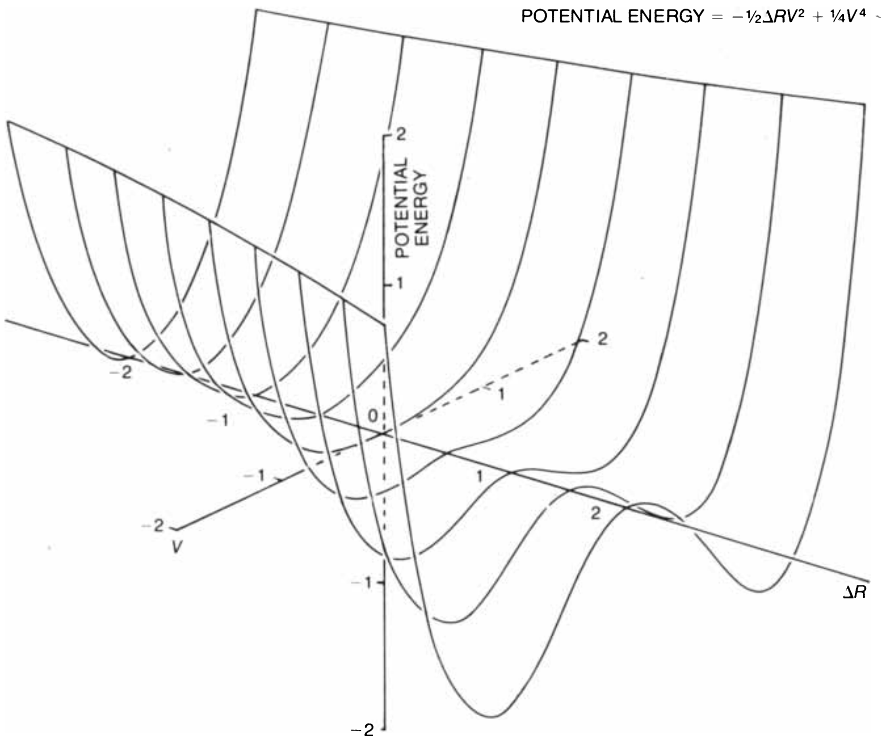
POTENTIAL SURFACE associated with the Rayleigh theory defines the relative energy of the fluid for any combination of Rayleigh number and velocity. The surface is described by a quadratic equation in which the coefficient, ΔR, is the difference between the actual value of the Rayleigh number and the critical value. As long as ΔR is negative the zero-velocity axis designates the state of minimum energy and a convective flow can be sustained only at the cost of increasing the energy. When ΔR is positive, the slope of the surface is reversed and the zero-velocity axis defines a state of unstable equilibrium; the fluid can reduce its energy by establishing a convective circulation. The major deficiency of the Rayleigh theory is that once the flow has begun the velocity continues increasing without bound, an unrealistic prediction.

The Rayleigh theory can be viewed as such a truncation of the infinite series, one that retains only the first term, namely $-\frac{1}{2}\Delta R V^2$. The surface described by this quadratic expression has two lobes, each with a parabolic cross section; one lobe is concave upward and the other is concave downward. It is obvious from the equation that whenever V is zero (that is, everywhere along the zero-velocity axis), the relative energy is zero. If the Rayleigh number is subcritical, so that ΔR is negative, the surface sweeps upward on each side of the axis and the energy increases whenever the velocity is greater than zero. In other words, the motionless state is a minimum in the potential surface, a state of stable equilibrium. When the Rayleigh number exceeds the critical value, so that ΔR is positive, the situation is just the opposite: the surface falls away from the zero-velocity axis, which is now a line of unstable equilibrium and maximum energy.

These properties of the surface are just the ones predicted by the elementary analysis of stability presented above, and they incorporate both the strengths and the weaknesses of the Rayleigh theory. In the immediate neighborhood of the origin, where both ΔR and V are small, the behavior of the system can be correctly inferred from the changing curvature of the surface. When ΔR is slightly negative, the fluid returns to rest after any small perturbation; if ΔR is slightly positive, the perturbation is amplified and a convective flow begins. When the Rayleigh number has exactly the critical value, so that ΔR is zero, the surface is flat and a random fluctuation in velocity is neither damped out nor amplified. At larger values of ΔR and V, however, a problem that by now should be familiar is encountered once again. The potential surface slopes toward negative infinity, with the result that the velocity increases without bound.

In the Landau theory this flaw can be remedied by retaining additional terms from the infinite series. Indeed, major improvements are possible with just a few more terms. For the hypothetical experiment from which the Rayleigh theory was derived considerations of symmetry constrain the choice of terms. In that experiment the nature of the circulation would be unchanged if the apparatus were cooled from above instead of heated from below, or if all velocities reversed direction. This invariance implies that the potential surface is symmetrical with respect to the zero-velocity axis; such a symmetrical surface is described by an equation that includes only even powers of the velocity (such as V^2, V^4 and so on). Whenever the exponent is an even number, V and $-V$ raised to the same power yield the same result, whereas the results have opposite signs when the exponent is an odd number. Hence all terms in which V is raised to an odd power (such as V^3, V^5 and so on) must have coefficients of zero.

Interesting results can be obtained by including one more term in the Landau potential beyond the quadratic term, namely the quartic term $\frac{1}{4}V^4$. The topography of the potential surface is then determined by evaluating the expression $-\frac{1}{2}\Delta R V^2 + \frac{1}{4}V^4$. When ΔR is negative, the surface closely resembles the simpler quadratic surface, although the energy increases more steeply when the velocity departs from zero. When ΔR is positive, the form of the surface is significantly altered. The energy decreases on each side of the zero-velocity axis, but the decrease does not continue indefinitely. Instead the energy reaches a minimum value and then rises again as the velocity increases further. Both the depth of the energy minimum and the velocity at which the minimum is reached increase as ΔR increases.



LANDAU THEORY yields a more realistic potential surface by including additional terms in the equation that defines the surface. In the simplest case, shown here, a quartic, or fourth-power, term is added to the quadratic term of the Rayleigh theory. The motionless steady state again becomes unstable when ΔR exceeds zero, but the velocity does not continue growing indefinitely. Instead there is a minimum in the potential surface at a finite velocity. The theory predicts that a new equilibrium will be reached at the velocity that minimizes the total energy. A theory of this form was devised in 1937 by L. D. Landau to describe certain phase transitions in magnetic materials and has recently been adapted to the convection problem.

For a limited range of Rayleigh numbers and velocities this comparatively simple version of the Landau theory can give realistic predictions. As before, when the Rayleigh number is less than critical, any random fluctuation in velocity is damped out and the stationary state is the state of minimum energy and stable equilibrium. At supercritical values of the Rayleigh number a perturbation grows larger, but the growth does not continue without limit. When the fluid reaches some finite velocity, determined by the value of ΔR and corresponding to the minimum in the potential surface, a new stable equilibrium is attained. Any departures from this velocity are then damped out.

The Landau theory that incorporates quadratic and quartic terms avoids some of the more spectacular failures of the Rayleigh theory, but it is still an approximation and it still remains valid only if the velocity is not too large. When V is large, high powers of V make a significant contribution even if they are preceded by a small coefficient; for this reason a theory that neglects all the higher-power terms cannot accurately represent the shape of the potential surface far from the zero-velocity axis. Moreover, in many convective systems one direction of flow is favored over the other, so that the symmetry of the potential surface is lost and odd powers of V must be included in the equation.

Both the Rayleigh theory and the Landau theory are derived from hypothetical experiments in which as many properties of the fluid as possible are considered to be constant. Real fluids are rarely this simple, and the interactions of the various properties can become quite tangled. For example, in the model it was assumed that only density varies as a function of temperature. In reality, viscosity and thermal diffusivity also change with temperature in most fluids. Since these quantities enter into the definition of the Rayleigh number, variation in them can have an important influence on the onset of a convective flow and on its subsequent evolution. It was also assumed that the fluid is incompressible; because many real fluids are compressible pressure becomes a significant variable, one that in turn affects the density and many other properties. A relation of notable intricacy connects temperature and viscosity. In general viscosity declines with increasing temperature, but at the same time the energy dissipated through viscous drag appears in the form of heat, and so it raises the temperature.

A theory that explicitly took into account all the known interactions of fluid properties would be too unwieldy to be practical. In describing a real fluid system rather than a mathematical model what must be sought is an optimum compromise between the complexity of the fluid and the complexity of the theory. The nature of the necessary compromise can be illustrated with a few examples of convection in the world outside the laboratory.

In the earth's atmosphere convection is observed at several scales of length. The temperature gradient between the Tropics and the poles drives a global circulation, which can be decomposed


McDONNELL DOUGLAS DC-10
DC-10

into at least three large convective cells in each hemisphere. Distortions of these patterns caused by the rotation of the earth give rise to the trade winds of the Tropics and the prevailing westerlies of the temperate zones. Local heating of the atmosphere near the earth's surface gives rise to smaller-scale convective flows, including those of most storms. Cumulus clouds, which form when warm air rises and cools and thereby becomes supersaturated with moisture, often mark the convective overturning of the atmosphere.

A theoretical analysis of atmospheric convection must take into account the large compressibility of air, which gives rise to a density gradient even when the temperature is constant with height. An accurate description of the atmospheric circulation would also have to include the compressive heating of air when it sinks into a region of higher pressure. Viscosity and other properties of air also vary with pressure and temperature, and the presence of water vapor, which gives up heat when it condenses, adds still another level of complexity. Clouds formed as a result of convective circulation are themselves often unstable to further convective motion: the cloud is cooled at the top by the loss of heat into space and warmed at the bottom by radiation absorbed from the ground. If the magnitude of these effects is large enough, a convection cell within the cloud can become established.

In spite of these complications convective motion in the atmosphere often manifests the same basic patterns observed in simpler laboratory experiments. The linear cloud formations called cloud streets or mare's tails are produced by cells of the roll type; satellite photographs occasionally reveal arrays of polygonal cells that extend over thousands of square kilometers. It is not the case, however, that the results of laboratory experiments can simply be extrapolated to atmospheric scale. Convection cells in the laboratory are always about as wide as they are high, whereas atmospheric cells are much wider, by a factor of as much as 50. In addition the sense of the circulation in small-scale experiments is always the same (with gases the flow is downward in the middle of each cell) whereas in the atmosphere both directions of flow are observed.

Convection in the oceans also spans a wide range of scales, from a meter or two to the size of the ocean basins themselves. The simplest of these flows has a straightforward explanation. Because some wavelengths of solar radiation penetrate a few tens of meters into the ocean the water is heated to a considerable depth. Cooling, on the other hand, results almost entirely from evaporation and from the loss of heat through conduction and radiation to the atmosphere, processes that are essentially confined to the surface. Hence heat is introduced into the oceans at a level below that at which it is removed, and a layer of water several meters deep can become unstable to convection.

The compressibility of seawater is small, and it can influence convective motion only at the greatest depths, but another variable has an important influence: salinity. The density of seawater varies not only with temperature but also with the concentration of dissolved salts; the density increases along with the salinity. As a result two independent factors can work together to establish a density gradient. The interaction of these factors can give rise to new kinds of convective motion not seen when only a single gradient is present.

If the temperature is greater at the bottom of a layer but the salinity is greater at the top, both gradients act in concert to encourage convection. When the temperature and salinity gradients act in opposite directions, subtler effects come into play. If warm salty water overlies cold fresh water, the temperature gradient favors stability but the salinity gradient upsets it. Even if the two opposed gradients combine to yield uniform density, convection can sometimes become established, for reasons having to do with dissipative effects that act differently on the two gradients. The temperature gradient is dissipated mainly by heat diffusion, whereas the salinity gradient is dissipated chiefly by the molecular diffusion of the salt and water molecules. Heat diffusion is much faster, often by a factor of 100. Initially the temperature and salinity of the two layers might be adjusted to give them precisely equal density. If a parcel of the warm salty fluid is then displaced downward into the cold fresh layer, it will lose its heat long before molecular diffusion can significantly reduce its salinity. As a result it will become denser and the motion will be amplified.

The opposite arrangement, with cold fresh liquid overlying warm salty fluid, can lead to the oscillatory phenomenon known as overstability. A parcel of warm saline water, on rising slightly, cools off but retains its salt concentration. As a result it becomes denser than it was initially and sinks back into the bottom layer. Indeed, it may overshoot its original position and continue oscillating about it. The oscillations may either grow or be damped out, depending on the magnitudes of the two gradients.

One of the most complex of convective systems is the one that apparently operates in the earth's mantle, creating a chain of rifts in the sea floor and propelling the continents across the surface of the earth. The heat that drives the circulation is liberated not at a boundary but rather throughout the volume of the material, mainly as a result of the decay of radioactive elements. A temperature gradient is formed under these circumstances because heat is lost from the system only at the surface, so that the temperature increases with depth. There is little doubt that the gradient is large enough to induce convection, but the properties of the system are so complicated and the mantle is so inaccessible to measurement that the form and dimensions of the convection pattern are highly uncertain. Viscosity increases sharply with depth, and at some level in the convective zone the material evidently undergoes a transition from one crystalline phase to another.

At a much smaller scale an interesting convective process of considerable complexity can be observed in a drying film of paint or lacquer. Here the driving force is not buoyancy but surface tension, as in Bénard's experiments. The mechanism ultimately responsible for the flow is the evaporation of solvent from the free surface of the film. If some perturbation increases the rate of evaporation in a region, that region will be cooled, which increases its surface tension. Moreover, the intrinsic surface tension of the pigments or other large molecules in the film is usually greater than the tension of the solvent, so that a deficiency in solvent raises the surface tension independently of the temperature. The liquid is drawn across the surface to regions of elevated surface tension, where it sinks to the base of the film and resumes the cycle. As the concentration of solvent is reduced, however, the viscosity increases, and ultimately the Marangoni number falls below the critical value. Convection then stops.

Convection cells in paint films often have a hexagonal form, or at least a polygonal form that approaches the ideal of regular hexagons. The flow can cause "flooding" of pigments, which is observed in the dry film as an irregularity in coloring. In some cases the three-dimensional pattern of the convection cells remains frozen in the dry film. It should be noted that the phenomenon is not always undesirable: paint with a "hammer" finish acquires its texture by this means.

The generality of the concept of convection is suggested by these diverse examples: the spontaneous overturning of the earth's atmosphere and oceans, and the circulation of paint in a film a few tenths of a millimeter thick. The theories that describe these fluid motions require many simplifying assumptions if they are to be of any practical use, and even then they are far from simple. It is therefore all the more remarkable that these few related theories, governed by a handful of dimensionless numbers, can account for phenomena that differ so greatly in scale.

The Physiology of the Koala

This gentle marsupial eats mainly eucalyptus leaves (which are toxic to other animals) seldom drinks and uses no shelter. Recent studies indicate how the animal manages to survive in its distinctive niche

by Robert Degabriele

A common reaction of people who see a koala (or more likely a picture of one, since the only place outside of Australia where koalas live is the San Diego Zoo) is to perceive the animal as a living teddy bear and to wonder if they could have one as a pet. The koala is of course not a bear and they cannot have one. It is a marsupial, a primitive mammal whose young are born in a virtually embryonic state and spend the early part of their life in a pouch that covers the mother's teats. And although it was once possible to keep koalas as pets, the practice is now forbidden by rigorous protective laws enacted in Australia because the animal was hunted for its pelt.

The scientific name of the koala is *Phascolarctos cinereus,* from the Greek for pouched bear and ash gray. The animal is formally characterized as an arboreal folivore, meaning that it lives in trees and eats leaves. The koala's tree is the eucalyptus, specifically 35 or more of the 600 or so species of the genus *Eucalyptus* that grow in Australia. The diet of an adult koala is almost exclusively eucalyptus leaves, and since the oils of eucalyptus leaves are toxic to most other mammals, one wonders what adaptations enable the koala to thrive on them. This subject has been the focus of my own work with koalas, and I shall be returning to it.

A fully grown koala weighs on the average about nine kilograms (20 pounds) and its body is at most about 62 centimeters (two feet) long. Its fur is thick and woolly, its limbs are long and its toes are strongly clawed. On each front foot the two innermost digits can be opposed to the others like two thumbs, as can the innermost digits on each rear foot. It is this articulation that makes the koala a good climber. On the ground, to which it normally descends only to move from one tree to another, the koala goes on all fours. Most of the time it is in a tree, sleeping and foraging among the leaves intermittently both by day and at night.

Koalas usually breed every second year. The mating season extends from early spring to midsummer, and the gestation period is about 35 days. A newborn koala (usually a single birth) is about 19 millimeters (three-quarters of an inch) long and weighs about 5.5 grams (less than a quarter of an ounce). This tiny creature climbs unaided to the pouch, which is unusual in that it opens to the rear, and fastens onto a teat. About six months later, having attained a length of some 20 centimeters and grown a good coat of fur, the young koala emerges from the pouch and thereafter, for another six months, is carried on its mother's back. At six months of age it is weaned on the emulsified droppings of the mother and then is ready for its adult diet of eucalyptus leaves.

For mammals a diet of leaves is a poorer source of energy than many other kinds of food would be. Moreover, mammals are less efficient than other animals in digesting leaves. For this reason many mammalian arboreal folivores are thought to be living close to the limits of their energy budget. Therefore even though the koala is biologically successful, the foundation of its success is precarious.

The question of the role of the volatile oils of eucalyptus leaves in the feeding behavior of the koala has taken up much time and effort. The common observation that the preference of the animal varies from species to species of eucalyptus and even from tree to tree within a species has led to the widely held assumption that a koala actively chooses leaves on the basis of their oil content and has a particular reason for doing so. The reasons put forward for the choice include the antimicrobial action of the oils in the voluminous cecum of the koala's intestinal tract and the "need" for the heat-generating effect of the oils in maintaining body temperature.

Ian Southwell of the Museum of Applied Arts and Sciences in Sydney recently examined the relation between the koala and the composition of volatile oils in eucalyptus leaves. He found no correlation between variations in the level of the volatile oils in the leaves and the feeding preferences of koalas. What his finding demonstrates is that the koala's considerable success in occupying the eucalyptus habitat is due to a general ability to detoxify oils that are harmful to other animals. In other words, the koala is one marsupial that has been able to overcome the chemical defense mechanism of eucalyptus trees.

In this way the koala has gained access to an abundant resource. How does the animal utilize it? How adequately does a eucalyptus provide food, water and shelter?

The first step in answering these questions is to consider the anatomy and physiology of the koala's digestive system. In common with other herbivorous mammals the koala is unable to digest cellulose and so must rely on microorganisms that do it. The location of the microorganisms in the digestive tract provides a basis for classifying herbivorous mammals into pregastric and postgastric digesters. In the pregastric category are such eutherian (placental) mammals as cattle and such marsupials as the kangaroo and the wallaby; the postgastric group includes such eutherian mammals as the horse and the rabbit and such marsupials as the brushtail possum and the koala.

Among the postgastric digesters the commonest site for the microorganisms is the cecum, an expansion of the hindgut at the junction of the small and the large intestine. The cecum is the most remarkable feature of the koala's digestive system. Recent observations have established that it takes up about 20 percent of the length of the postgastric intestine. Thus the koala has a fermentation chamber where the passage of leaves can be delayed so that the microbes can digest the cellulose.

The possession of such a large cecum

BROWSING KOALA is seen in its characteristic habitat, a eucalyptus tree. Koalas subsist on the leaves of about 35 of the approximately 600 species of the genus *Eucalyptus* that grow in Australia. The leaves provide all the nutrients the koala needs and usually enough moisture so that the animal does not have to drink. The koala feeds and sleeps intermittently in the trees throughout the day and night.

means that the koala's dietary requirement for carbohydrate is likely to be met at all times through the microbial digestion of cellulose. The animal's position with respect to nitrogen, that is, protein, is less assured. In order to determine the koala's ability to remain in positive nitrogen balance my colleagues and I at the University of New South Wales in Australia maintained a group of koalas in cages, feeding them exclusively on fresh leaves of the gray gum (*Eucalyptus punctata*), which is known to be a food tree for koalas. We measured the nitrogen taken in and excreted in both summer and winter.

As one might expect, we found that in such a regimen the koala is quite capable of remaining in positive nitrogen balance throughout the year. The way the balance is achieved was not so predictable. Although the intake of digestible nitrogen seemed to be about the same at all times, the dietary intake was significantly higher in winter. In other words, the koalas had to consume more leaves in winter to remain in nitrogen balance.

The explanation probably lies in seasonal differences in the quality of the leaves. Eucalyptuses such as the gray gum grow rapidly in spring and early summer but not much after they flower. It is therefore likely that winter leaves are older and more fibrous and contain less digestible nitrogen. Here is one bit of evidence that mammalian arboreal folivores may be living close to the limits of their energy budget.

The precarious nature of the koala's energy budget is further emphasized by the phenomenon known as "wasting disease." In times of drought koalas have become comatose and have died with a full stomach. I suspect that a nitrogen deficiency is the major factor, since during a drought eucalyptuses grow few leaves or none. The remaining leaves are comparatively old and contain progressively less digestible nitrogen. A koala's response to such a change in the quality of the available leaves is likely to be to eat more leaves. If the quality of the leaves deteriorates enough, the quantity of leaves that is required and even the capacity of the koala's digestive system can become limiting factors. Under such circumstances it may become physically impossible for a koala to satisfy its nutritional requirements.

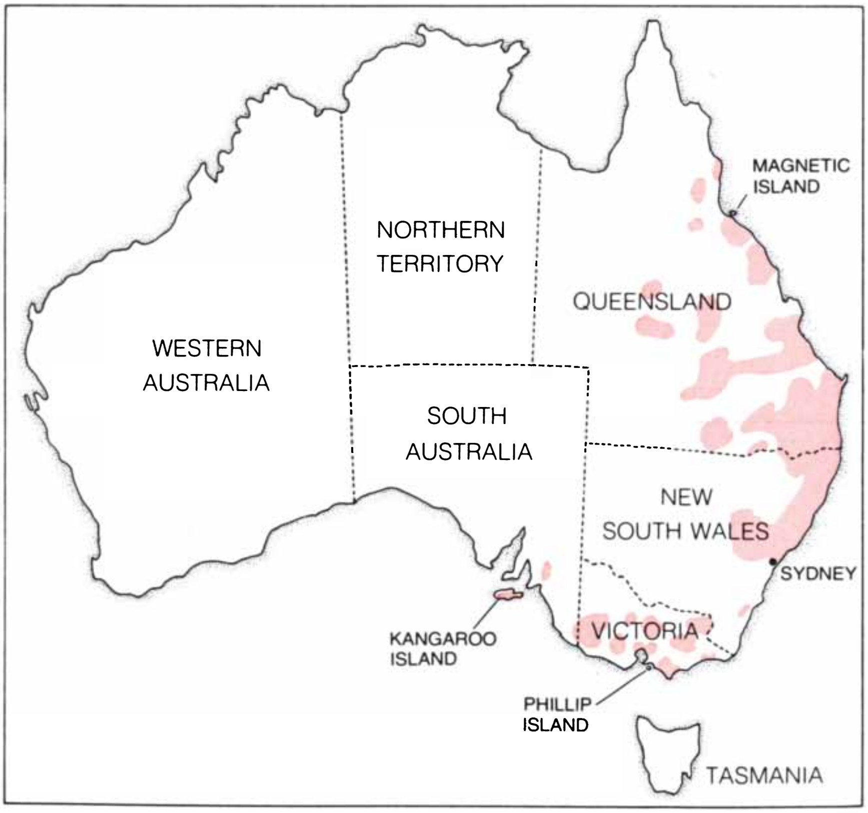
The koala also has a reputation for seldom or never drinking water. This is a common theme running through the various legends of the Australian aborigines about the koala. Some of the aboriginal names for the animal, such as *koobor*, mean "Does not drink water." Curiously, the name koala, an aboriginal name from the Hawkesbury River district near Sydney, has no such connotation.

The available evidence suggests that in normal conditions a koala gets the water it needs from dew and from eucalyptus leaves. In general water is lost in the urine, in the feces (either as free water or as water that would have been produced by the metabolic breakdown of food) and through evaporation. The urinary water loss is regulated by the kidney. The free-water content of the feces is regulated by the large intestine, and the metabolic-water content depends on digestive efficiency. The evaporative loss is tied closely to the animal's mechanisms for regulating its body temperature.

However the koala's water balance is achieved, the turnover rate of water is useful as an indication of both the availability and the utilization of water. One can determine that rate (and also the total content of water in the body) by observing the dilution and the rate of disappearance of water labeled with tritium, the radioactive isotope of hydrogen. We studied the water metabolism of free-ranging koalas in three widely separated populations: one at Magnetic Island, near the northern limits of the koala's range, one at Sydney, near the center of the range, and one at Phillip Island, the southern limit of the range. No significant differences were found in total body water and turnover rate.

The finding indicates that the microhabitat of the koala is reasonably uniform in terms of water requirements and water supply no matter where the animals live. The free-water content of gray-gum leaves ranges from 40 percent (old and fibrous leaves) to 65 percent (new growth). The leaves of other species of eucalyptus contain at least as much water as those of the gray gum, since leaves with a water content below 40 percent dry out and die. One can conclude that in normal conditions the leaves of eucalyptus trees provide the koala with both food and water in adequate amounts.

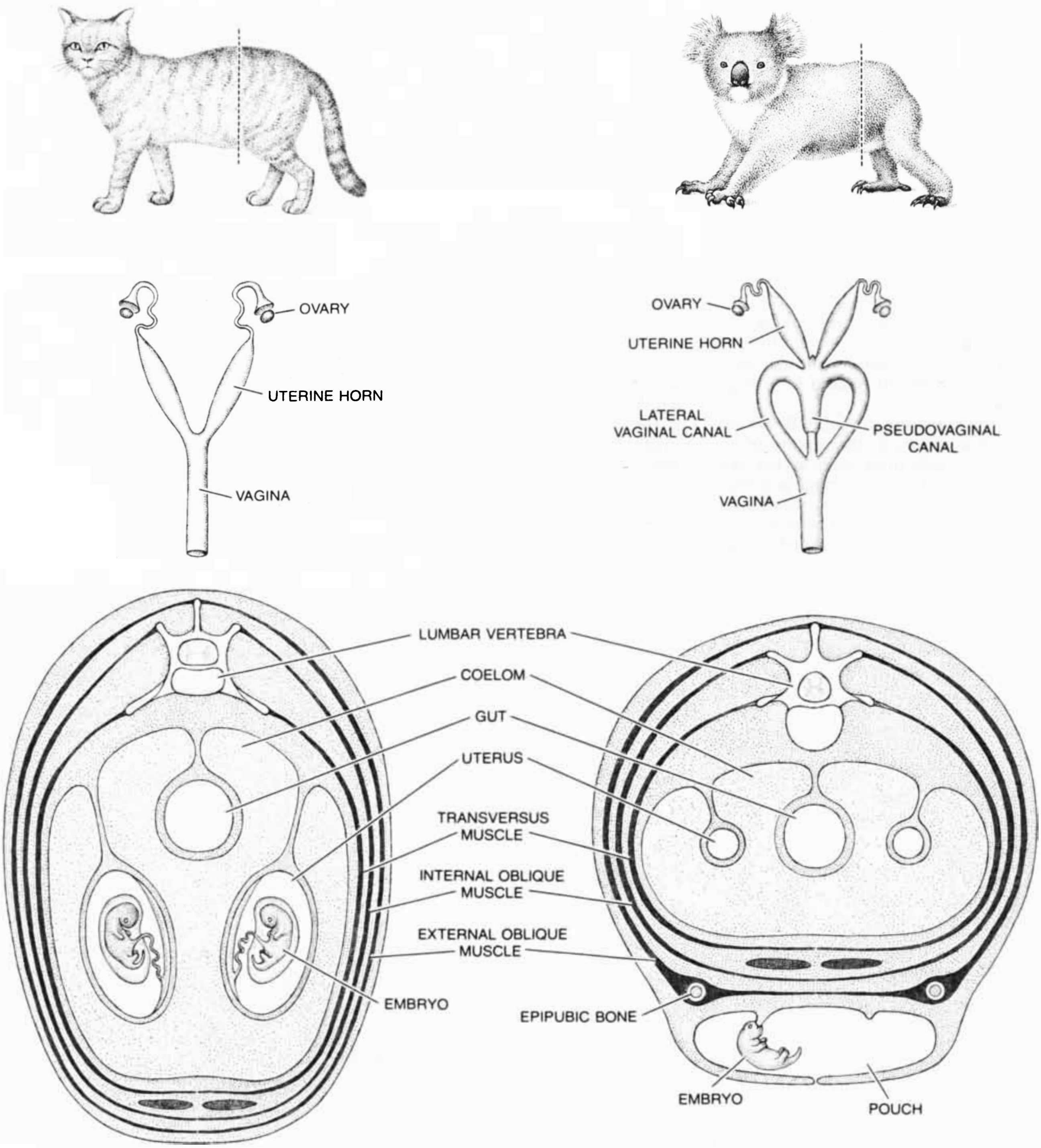
The koala's content of body water is relatively high (77.4 percent of the animal's weight). That is close to the water content of the fat-free component of the carcass of any mammal, a relation suggesting that the total absence of fat deposits in all the koala carcasses I have examined is characteristic of the koala. The lack of fat may be a consequence of the koala's precarious nutritional bal-



DISTRIBUTION OF THE KOALA is indicated by the colored areas on this map. The animal was once so vigorously hunted for its pelt that it was threatened with extinction. With the enactment of protective laws Australia also made efforts to reestablish the species, particularly in Victoria, where colonies were set up on islands and in mainland forests. The northernmost colony appears to be one that developed from a group of koalas introduced to Magnetic Island.

ance. It is also what enables the animal to carry such a high proportion of body water. (Much of the water is in the cecum, which can hold large amounts of moist food.)

The usefulness of carrying a considerable amount of water becomes clear when the rate of water turnover by the koala is examined closely. When that rate is expressed in terms of the four-fifths power of body weight, the effect of varying body size is negated and one can make comparisons within and between species. We have compared the koala and the short-nosed bandicoot (*Isoodon macrourus*), a rat-size terrestrial marsupial that feeds on invertebrates. The water-turnover rate of short-nosed bandicoots living on an island with no available drinking water is similar to that of the koala (179 grams per .8 kilogram of



REPRODUCTIVE SYSTEMS of the cat (*left*), a typical eutherian (placental) mammal, and of the koala (*right*), a marsupial, are compared. The drawings at the bottom show transverse sections through the body just forward of the hind legs. The epipubic bone, which is distinctive in marsupials, helps to support the pouch. The young of a placental mammal are much more fully developed at birth than the young of a marsupial. A koala at birth is less than an inch long and is still virtually an embryo; it crawls into the pouch and attaches itself to a teat. It stays in the pouch for about six months, by which time it has grown to a length of about 20 centimeters (eight inches) and has developed a good coat of fur. After emerging from the pouch the infant is carried on its mother's back for another six months.

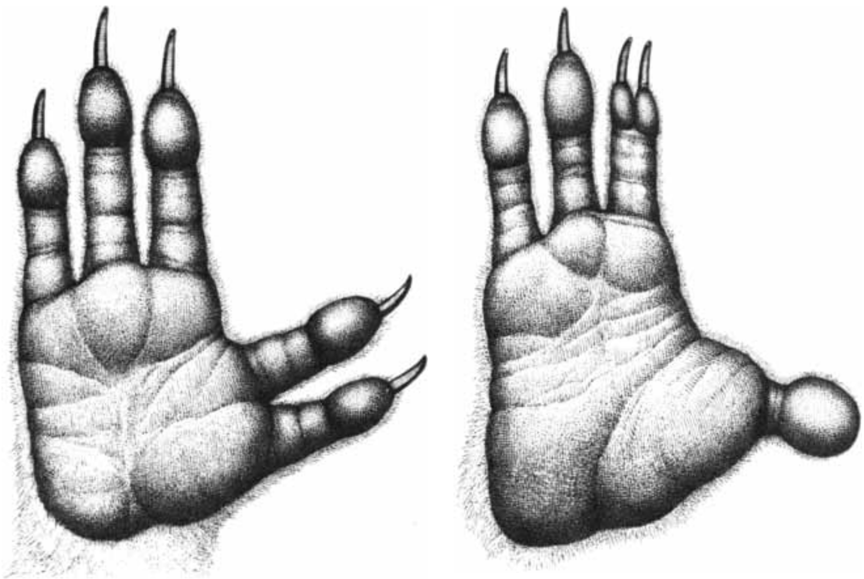
body weight per day). In bandicoots living on the water-rich mainland the rate is much higher (243.8 grams). This comparison reinforces the idea that the koala's food incorporates a supply of water and that in normal circumstances the animal does not drink water. The idea is further reinforced by the fact that the koala has a simple type of kidney that is not capable of a high degree of water conservation. Animals that have evolved to survive with a low intake of water typically have a kidney that can produce highly concentrated urine, so that more of the available water stays in the body.

We investigated the way the koala achieves water balance by making measurements on individually caged koalas. Each koala was given a daily supply of freshly cut branches of gray gum. Drinking water was always available in the summer but was sometimes withheld in the winter. Inputs and outputs of water were determined as changes in weight.

The rate of water turnover of the caged koalas was no more than half the rate of their free-living counterparts. The most likely reason for the difference is the reduction in activity imposed by a cage. As would be expected in an arboreal mammal, the leaves were the major source of water.

Although the koalas consumed more food and water in the winter, drinking water contributed only about a fourth of the intake in both summer and winter. It is likely, then, that the increased intake of water in the winter is the result of a need to consume a greater volume of nutritionally inferior leaves. In other words, nutritional requirements would dictate the level of water intake. Nevertheless, the water intake associated with the food remains reasonably constant at from 40 to 50 grams per kilogram of body weight per day. Therefore, provided that the leaves being eaten are nutritionally adequate, the koala can get enough water from its diet.

The koala loses water mainly through evaporation from its respiratory surfaces. The loss of water in the urine is the least significant component. Evaporative loss and the production of urine show remarkably little variation between seasons and when drinking water is not available. The slight decrease in urinary output in the absence of drinking water suggests that the koala has a water-conservation mechanism of the kind found in the pig and the beaver, whose kidney actively reabsorbs urea when drinking water is limited. In the koala, however, the reduction in the volume of urine is so slight that I suspect the urea is being reabsorbed for nutritional reasons rather than for water-conservation ones.



HANDS AND FEET of the koala are adapted for movement in trees by having long and strong claws and digits that are opposable to other digits as the human thumb is. The hand (*left*) has two such digits, the foot one digit. The joined second and third toes of the foot are often employed by the animal in combing its fur. The palms and soles have granular pads for gripping.

The loss of water in the feces is probably the most effective means the koala has of regulating its water balance. The feces are always quite dry. When drinking water is not available, the water content of the feces drops from 52 to 43 percent, which is the same as the water content of the feces of a dehydrated camel. It is evident that the koala can manage quite well with a regularly available but not overabundant water supply.

The high rate at which water is lost through evaporation from the respiratory surfaces is an indication of the importance of the relation between water balance and thermal balance. An examination of the way in which the koala regulates its body temperature clarifies this relation.

The koala is unusual among the arboreal marsupials in that it does not seek any kind of shelter, a pattern of behavior (or rather a lack of one) found in only one other arboreal marsupial group, the tree kangaroos. (Such behavior is quite common among tree-dwelling placental mammals.) One is therefore stimulated to find out what kind of protection against extremes of environmental conditions is provided by the fur of the koala.

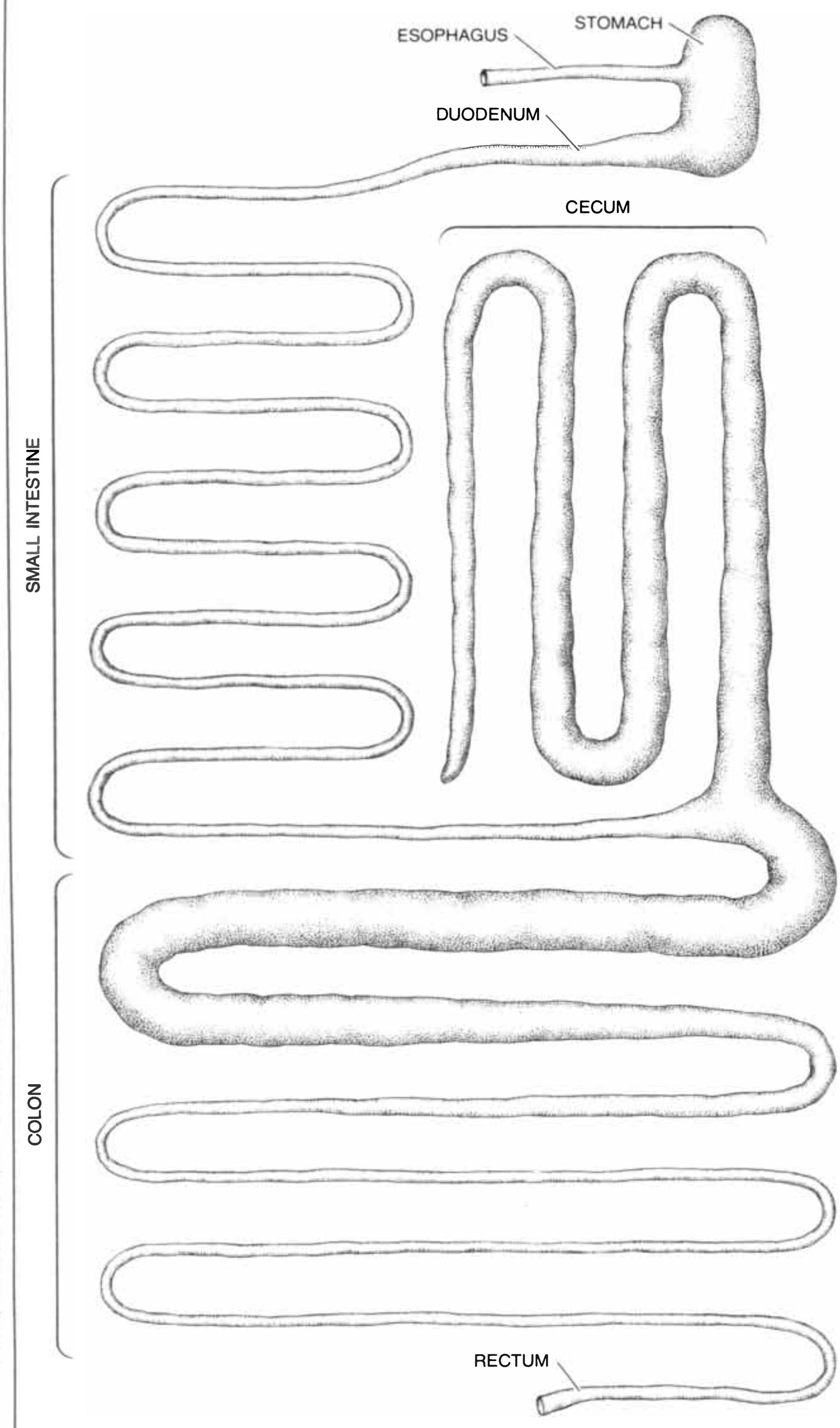
The dorsal fur, which is the thickest at 54.4 hairs per square millimeter, covers 77 percent of the animal's body surface. The ventral hair has only half the density of the back fur and covers 13 percent of the body surface. The differences in density are not paralleled by differences in hair length, which for both the longer guard hairs and the shorter underfur is virtually the same all over the body. One does, however, find seasonal differences in hair length: the length of the long hairs differs from that of the short ones more during the summer than during the winter.

The thick dorsal fur is darker than the sparse ventral hair; it therefore tends to absorb solar heat as well as to provide insulation. The sparse ventral hairs can be erected, so that the amount of insulation the ventral hair provides can be adjusted. The combination of these kinds of covering gives the koala a type of environmental control reminiscent of that of the guanaco, a llama-like South American animal with a densely matted dorsal surface (40 percent of the total area) and sharply defined areas of almost bare skin on the ventral surface. By changing its posture the guanaco can achieve as much as a fivefold variation in thermal insulation in still air and five times more than that in the wind.

Casual observations of koalas in trees on windy days indicate that as the wind speed rises a koala tends increasingly to roll up into a compact ball in order to present only its curved middorsal surface to the wind. Koalas did this even when the temperature was high. As the wind speed increased further their ears were folded forward so that almost nothing projected into the airstream.

The picture of a portable shelter worn by every koala was enhanced by measurements we made of the insulation characteristics of koala pelts. (Our sample was small and, being fortuitous, consisted of more summer pelts than winter ones, but there was not much difference between the two kinds.) The koala's dorsal fur proved to have the highest insulation value among the 12 marsupials so far investigated. Indeed, it is in the lower range of the values obtained for arctic animals.

The effect of wind on the insulating power of the dense, matted dorsal fur is small. Moreover, compared with the fur of a number of other animals the koala's fur shows the smallest decrease in insulating value as the wind speed increases, at least up to about 10 miles per hour.



ALIMENTARY TRACT of the koala is notable for the size of the cecum, which accounts for some 20 percent of the length of the intestine. Since the koala cannot digest the cellulose in the leaves that are its staple diet and relies on microorganisms to do it, the cecum serves as a fermentation chamber in which the passage of leaves through the digestive tract can be delayed so that the microbes can do their work. The koala can detoxify the oils of eucalyptus leaves.

The average decrease was 14 percent and the minimum was a remarkable 3 percent. These data suggest both that the fur would maintain a significant level of insulation at much higher wind speeds and that for an animal living in the exposed treetops of the open forest the fur provides excellent thermal protection.

The temperature-regulating effect of fur is complemented by metabolic activity. The koala's basal metabolic rate is 74 percent of the rate predicted for marsupials in general. (Among placental mammals such folivores as the sloth and the potto show a similar divergence from the predicted rate.) The koala responds to high environmental temperatures by panting. At low environmental temperatures the low metabolic rate is accompanied by a high level of total body insulation, with the fur contributing 50 percent (a rather large proportion). The metabolism-insulation pattern of the koala appears to be characteristic of many tropical arboreal mammals; perhaps the arboreal marsupials in general differ from the terrestrial ones in this way.

In examining the relation between water balance and thermal balance I began with the fur and led into thermoregulatory mechanisms. Now I can complete the circle by considering the relation between evaporative water loss (a thermoregulatory characteristic) and metabolic water production (an aspect of water balance). The amount of water produced by metabolism for every gram of oxygen consumed is calculated on the basis of the nutrient composition of the leaves and of the feces. The ratio of metabolic water production to evaporative water loss is then derived from the relation between oxygen consumption and each of those processes. Such calculations have shown that the koala's requirements for evaporative cooling are adequately provided for up to an ambient temperature of about 30 degrees Celsius (86 degrees Fahrenheit), which is rarely exceeded for long in the animal's environment.

In summary, the koala is an animal whose ecological niche can be described (in the most general sense) as the arboreal environment provided by trees that belong mainly to the genus *Eucalyptus*. They provide the animal with a supply of food and water and with a place to live. With its specialized digestive system the koala can overcome the toxic effects of eucalyptus oils and can extract enough nutrients and water from the eucalyptus leaves. The koala's thermoregulatory pattern is attuned to that level of water supply, and so (other considerations aside) the animal is potentially able to thrive in most of the forests of Australia.

Supercoiled DNA

In many forms of DNA the double helix itself forms a helix. This supercoiling, which has important biological consequences, is best described and analyzed by means of a simple mathematical model

by William R. Bauer, F. H. C. Crick and James H. White

DNA, the primary genetic material of most organisms, is usually visualized as a double helix in which two chains of complementary nucleotides (the subunits whose sequence constitutes the genetic message) wind around a straight common axis. It is now clear, however, that often the axis of the double helix is not linear but curved. Indeed, the double helix can wind in space to form a new helix of a higher order, in which case it is said to be supercoiled. It appears that a large proportion of the known DNA's exhibit some form of supercoiling in at least one stage of their life cycle. An appreciation of this structural feature and its consequences is therefore essential to a complete understanding of the biology of DNA.

Affecting DNA's in a wide range of sizes and shapes (including some that are not organized into a double helix), supercoiling takes a variety of forms. For example, in the chromatin (DNA complexed with protein) of higher organisms the DNA is wound around a core of protein to form a left-handed solenoidal superhelix. Here we shall be concerned mainly with a different type of supercoiling in which no protein core is needed. It is the supercoiling of closed circular DNA: double-helix molecules in which each polynucleotide chain forms an unbroken loop.

The existence of these twisted rings of DNA was first postulated to explain a surprising discovery about the DNA of the polyoma virus: a small virus that causes tumors in the mouse. In 1963 it was reported that when this DNA is suspended in a solvent and spun in a centrifuge, it resolves into three components distinguishable by the velocity at which they move through the solvent. Further investigation revealed that these components, labeled I, II and III in order of their decreasing velocity of sedimentation, do not differ in molecular weight. Hence there had to be some variation in the molecular compactness of the three species accounting for the difference in their velocities. In other words, the molecules had to have different shapes.

Viewing the components by means of electron microscopy indicated that this was indeed the case. Although the molecules of component III were clearly linear, those of component I and component II appeared to lack ends. The revelation that components I and II were circular, however, did not solve the problem. There was still no clue to what distinguished the two components. Then in 1965 Jerome Vinograd of the California Institute of Technology put forward an ingenious proposal that solved the mystery and introduced the concept of supercoiling in DNA.

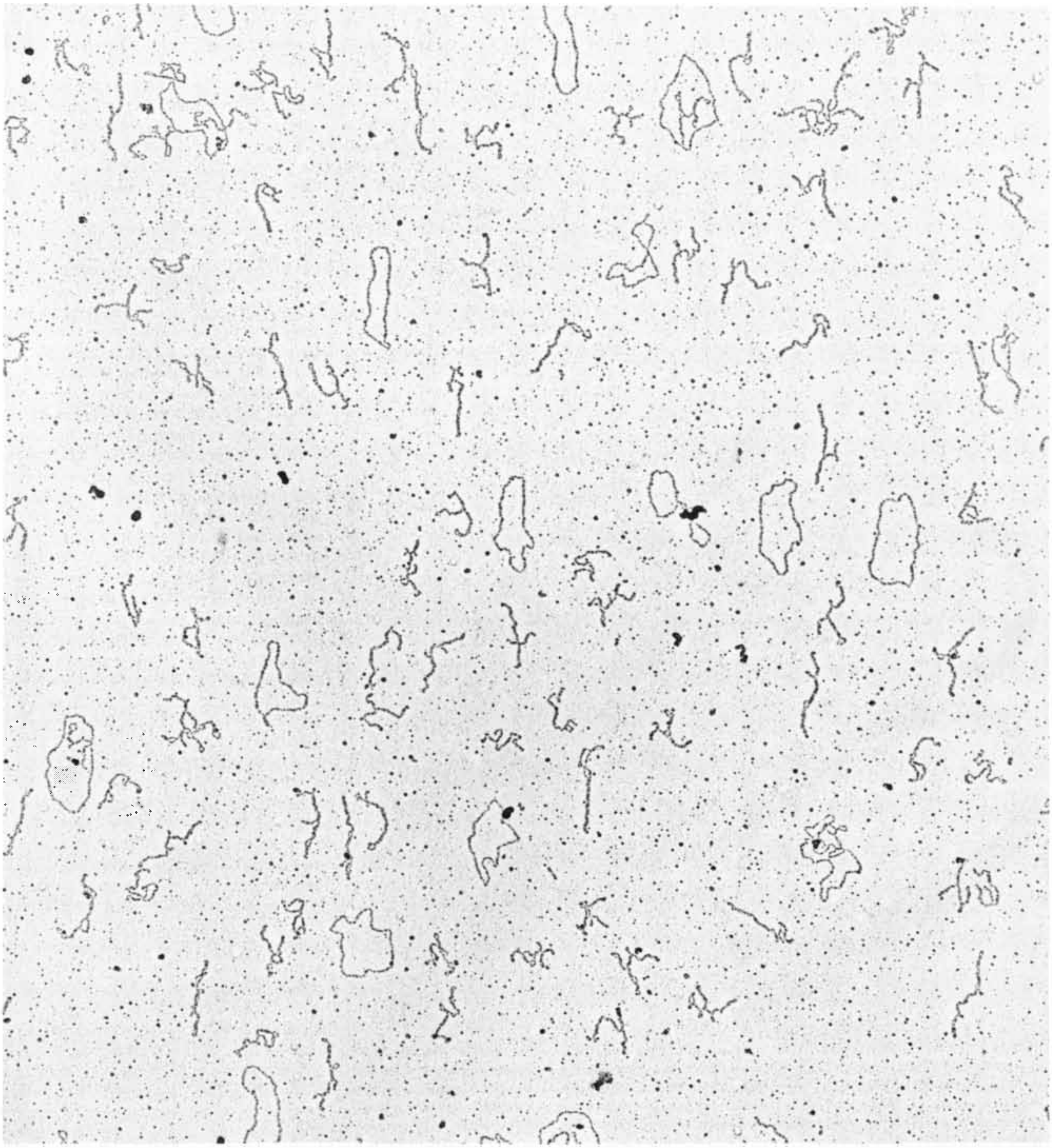
Vinograd suggested that a closed circular molecule of DNA (a circular molecule in which both polynucleotide chains are completely intact) is usually underwound compared with a linear piece of DNA of the same length. In other words, in the closed circular molecules of component I there are fewer helical turns than there are in the molecules of either component II or component III. Component I behaves as though before the ends of the double helix were joined a number of 360-degree twists had been made in the direction opposite to that of the helix's normal winding.

The difference between components I and II of polyoma DNA, then, is that the circular molecules of component II are "nicked": in each molecule there is at least one nick, or break, in one of the polynucleotide chains. In contrast, circular molecules of component I are closed and underwound. Because the forces that stabilize the double helix are strong the closed circular molecules resist such underwinding, and for reasons we shall discuss they compensate by forming supercoils. Hence the closed circular molecules of component I of polyoma DNA are more compact than the nicked circular molecules of component II and therefore have a higher sedimentation velocity.

Supercoiling is a widespread phenomenon, characterizing many of the medically and biologically most interesting DNA's. An entire class of DNA tumor viruses, including the polyomas and the human papilloma (wart) viruses, contain such DNA, and the DNA of mitochondria (the energy-transducing organelles) of human and other animal cells is supercoiled. Animal cells have also been found to contain a vast array of very small supercoiled DNA molecules whose function has not been determined. It is particularly striking that the majority of the known smaller genomes (sets of genes) fall in this category, including genetic factors for fertility and drug resistance. Moreover, the phenomenon of integration, in which a small piece of DNA is physically inserted into a larger DNA molecule, appears to require that the integrated element be supercoiled. Thus the vectors, or vehicles, for DNA in genetic engineering are of this type. Perhaps most remarkable is the fact that RNA tumor viruses, including the cancer-causing Epstein-Barr virus, often employ the enzyme reverse transcriptase to make DNA copies that become supercoiled in the course of infection. It should also be noted that supercoiling has been found in molecules with as few as 350 nucleotides in each polynucleotide chain and with as many as 1,750,000.

Virtually every physical, chemical and biological property of DNA—its transcription, hydrodynamic behavior, energetics, enzymology and so on—are affected by closed circularity and the deformations associated with supercoiling. Understanding the mechanism of supercoiling and the consequences of this structural feature for DNA, however, presents problems of considerable mathematical complexity. Fortunately there are two branches of mathematics that offer substantial help in this effort: topology, which studies the properties of structures that remain unchanged when the structures are deformed, and differential geometry, which applies the methods of the differential calculus to the study of curves and surfaces. In what follows we shall first describe a mathematical model of closed circular DNA and then discuss the implications of the model for real DNA. We shall also take up some of the methods that have been developed for measuring supercoiling and its effects.

Consider first the physical structure of the double helix: it is formed by two sugar-phosphate backbones to which the nucleotide bases of the DNA are attached. The bases on the opposing backbones are paired, with about 10 base pairs for each turn of the helix. (The exact number depends on the particular configuration of the DNA molecule.) To study supercoiling mathematically it is most convenient to construct a model in which the structure is represented as a narrow twisted ribbon of infinitesimal thickness. The most obvious way to design such a model is to specify that the edges of the ribbon follow the sugar-phosphate backbones of the DNA. Carrying out the construction in the simple case of a linear double-helix molecule,

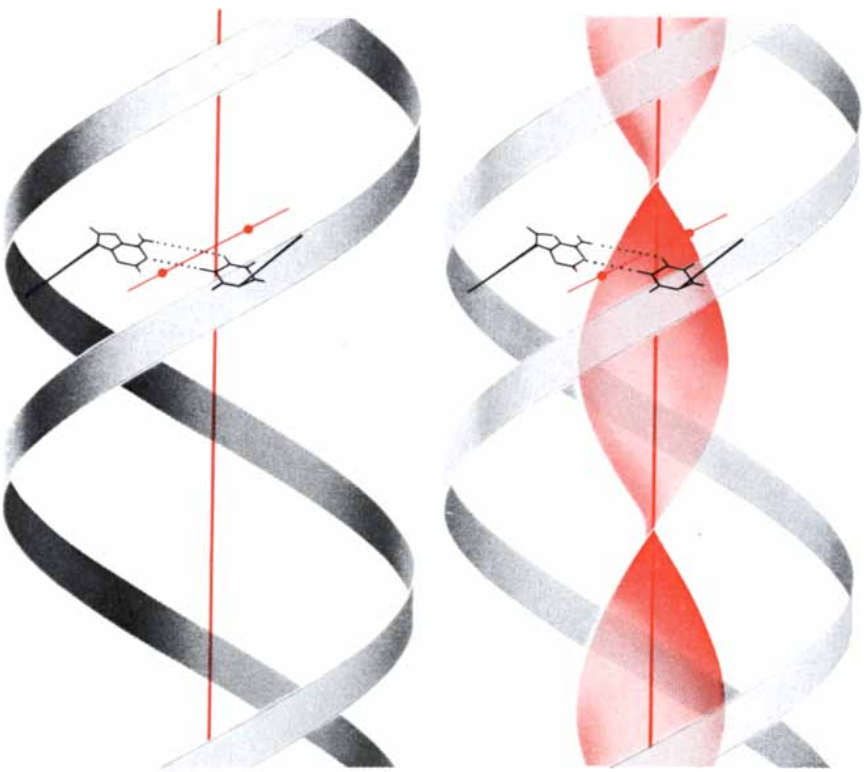


MANY SUPERCOILED DNA MOLECULES appear in this electron micrograph. Each molecule is a loop of the DNA double helix. Where the loops are open one of the strands of the double helix is "nicked," or broken, so that the loop is relaxed, or not supercoiled. Where the loops are not open but kinked they are un-nicked. In such molecules the DNA usually has fewer helical turns than a linear molecule of the same length. In un-nicked loops the stress introduced by this underwinding is compensated for by supercoiling. These molecules are plasmids extracted from the bacterium *Escherichia coli* and separated from other *E. coli* DNA molecules by centrifugation. They are enlarged some 25,000 diameters. The micrograph was made by Gary Cohen of the State University of New York at Stony Brook.

however, demonstrates that this choice is unsatisfactory: the axis of the model (the line halfway between the edges of the ribbon) winds around the straight axis of the double helix. For our purposes it is preferable to choose a model whose axis coincides with that of the double helix. In addition we specify that the ribbon must always lie perpendicular to the pseudodyads, or twofold axes of rotation, that are distributed along the double helix [*see illustration below*]. (There is one dyad axis associated with each nucleotide pair and one associated with the space between successive pairs; these axes are all perpendicular to the axis of the double helix, and their location is independent of the sequence of bases in the molecule.)

This ribbon model follows the axis of the DNA double helix and twists as the two chains of the molecule twist around that axis. In addition, because the sequences of atoms in the two polynucleotide chains run in opposite directions the edges of the ribbon will be assigned opposite orientations. (It does not matter what directions are chosen for the two edges as long as they are opposite.) This model can be analyzed mathematically in a number of different ways, but we shall be most concerned here with the relation between the oppositely directed edges of the ribbon.

When the ends of a ribbon are joined, each edge describes a closed curve in three-dimensional space. Furthermore, when the ribbon represents a closed circular molecule of DNA, a number of 360-degree twists are introduced before the ends of the ribbon are joined, and so the two curves described by its edges are linked. In other words, it is impossible to separate the curves without "cutting" one of them. If each loop in a linked pair represents a covalently bonded molecule, as is the case with the two polynucleotide chains of the double helix, the two are said to be joined by a topological bond. This is a peculiar type of bond in that although no part of one molecule is covalently joined to any part of the other, it is nonetheless necessary to break a covalent bond in order to separate the two. (The concept of topological bonding was first introduced into chemistry by Edel Wasserman and his colleagues at Bell Laboratories.)



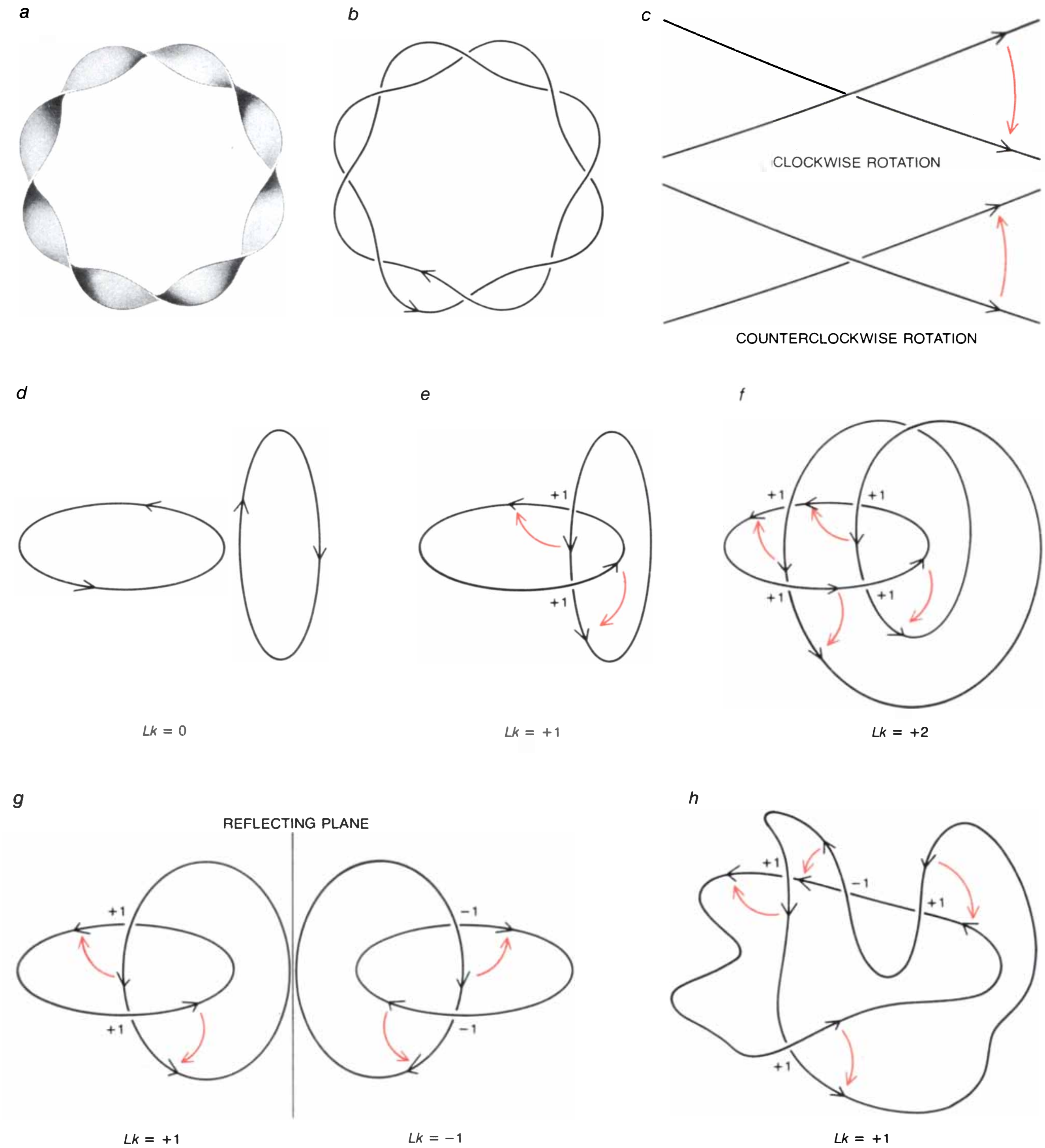
MODEL OF SUPERCOILED DNA makes use of two structural features of the double helix (*left*): the axis (*thick colored line*) around which the two sugar-phosphate backbones of the molecule wind and the series of "pseudodyad" axes defined by the nucleotide bases attached to the backbones. The bases, whose particular sequence constitutes the genetic message, are paired, and there is one dyad axis associated with each base pair and one with the space between successive base pairs. (Here only one base pair is shown.) The dyad axis (*thin colored line*) defined by each base pair is the line perpendicular to the helical axis about which either one of the bases can be rotated into the position of its oppositely oriented mate. The model constructed with respect to these features (*right*) is a narrow, infinitesimally thin ribbon whose axis, or centerline, follows the axis of the double helix and whose surface always lies perpendicular to the dyad axes defined by the base pairs of the double helix. This ribbon bends as the axis of the DNA molecule bends, and it also twists as its two polynucleotide chains twist around each other.

In mathematical terms the linking of two closed curves is a topological property: no matter how the curves are deformed (pulled, twisted and so on), as long as neither one is broken they will remain linked in exactly the same way. It will be useful, then, to assign a numerical value that describes the way the loop formed by one edge of the ribbon representing a molecule of DNA is linked to the loop formed by the other edge.

It seems reasonable to define such a linking number *Lk* so that its value will be 0 for a pair of unlinked closed curves, 1 for a curve that loops through another just once, 2 for a curve that loops through another twice and so on. Since the two edges of the ribbon have been given orientations, however, it turns out that a more useful definition of a linking number can be made in which either a plus or a minus sign is assigned to the number depending on the orientation of the curves. There are a number of ways to compute such a linking number for a particular configuration of curves, and one of the most convenient is to examine all the points in a projection, or two-dimensional representation, of the configuration where a piece of one of the curves crosses over a piece of the other [*see illustration on opposite page*]. Each of these crossing points can be assigned an index number of $+1$ or -1, according to the direction in which the top piece must be rotated so that it coincides with the bottom piece. Adding up the index numbers of all the crossing points and dividing by 2 (the number of linked curves) gives the linking number *Lk*.

Thus the linking number *Lk* is defined as a signed integer that describes a property of two closed curves in space. To separate a pair of curves without actually cutting them the value of *Lk* must be 0 (although the converse is not always true). If the curves in question are the edges of a closed ribbon with N turns in it, their linking number is $+N$ or $-N$, depending on the direction of turning. Moreover, as long as the ends of the ribbon remain joined the number will remain unchanged when the ribbon is deformed. Notice that although the edges of the ribbon model of DNA were not chosen to coincide with the sugar-phosphate backbones of the double helix, the linking number of the ribbon will be exactly that of the backbones.

Therefore for a molecule of relaxed closed circular DNA 5,000 base pairs long, with 10 base pairs for each turn of the helix, the linking number will be $+500$. By convention the number is positive because the DNA double helix is right-handed. (A closed circular DNA molecule is said to be relaxed if its axis lies entirely in a plane. As we shall relate, there are many ways in which



CLOSED CIRCULAR DNA is modeled by a twisted ribbon whose ends have been joined (*a*). Since the sequences of atoms in the two chains of the double helix run in opposite directions, it is also convenient to assign the edges of the ribbon opposite directions. When the edges are viewed as directed closed curves in three-dimensional space (*b*), they are found to be mathematically linked; in other words, there is no way to separate them without breaking one or the other. This relation can be described mathematically by a linking number *Lk* whose magnitude expresses the number of times one curve is linked through the loop of the other and whose sign depends on the way the curves are labeled. One way to compute the value of this quantity is to examine a projection, or two-dimensional representation, of the two curves and to each point where one curve crosses over the other (but not where a curve crosses over itself) assign an index number according to the following rule: If a clockwise rotation is required to move the top piece so that it coincides with the bottom piece, then +1 is assigned to the crossing point, and if a counterclockwise rotation is required, then −1 is assigned (*c*). (This convention is the reverse of the one normally employed in mathematics, but it ensures that the linking number of right-handed closed circular DNA will be positive.) *Lk* is then calculated by adding up the index numbers and dividing by 2 (the number of curves). The linking number obtained in this way is a signed integer equal to 0 if the two curves are unlinked (*d*), to +1 or −1 if one curve links through the other just once (*e*), to +2 or −2 if one curve links through the other twice (*f*), and so on. The sign of the number will change if the orientation of either one of the curves is changed or if the pair of curves is viewed in a mirror (*g*). The value of the linking number remains the same no matter how the two curves are deformed (*h*), and so since the linking number of a twisted ribbon is equal to the linking number of the sugar-phosphate backbones of the DNA molecule it models, the linking number expresses an important constraint on the possible supercoiled structure of the DNA.

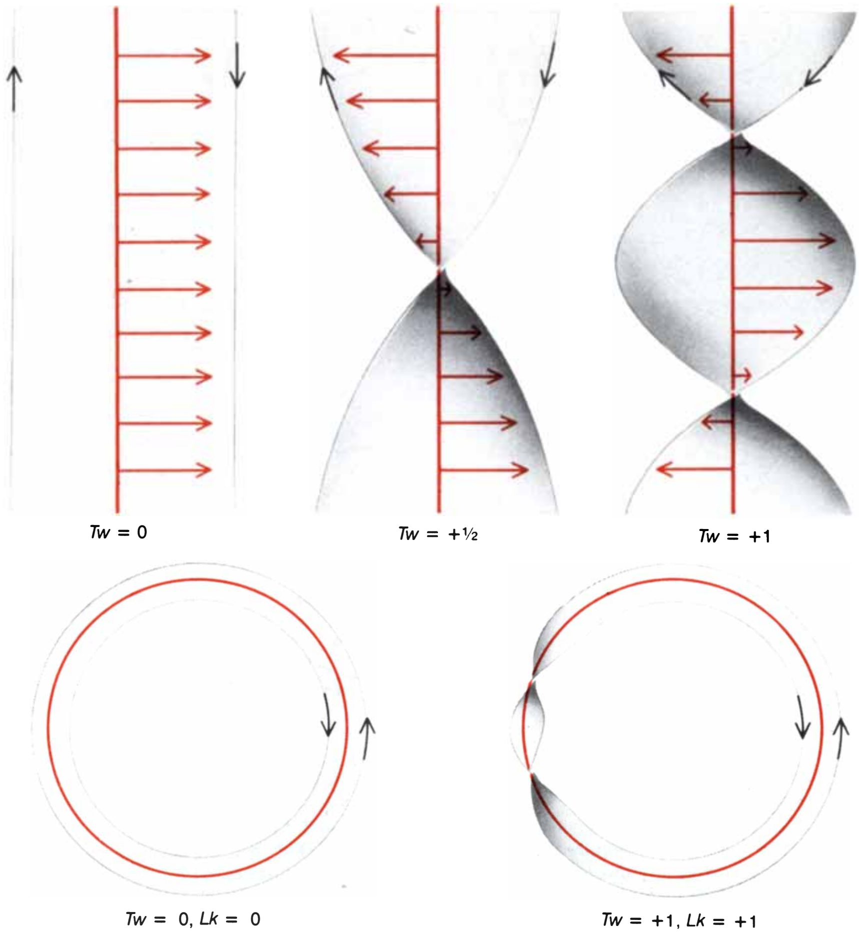
naturally supercoiled molecules can be made to relax. A nicked circular molecule is relaxed in its natural state, but since its polynucleotide chains are not both intact, strictly speaking it has no linking number, the quantity that is our principal concern here.)

Another way to analyze the ribbon model of DNA is by looking not at the relation between its edges but at the way the ribbon twists. For a ribbon whose axis follows a straight line the idea of a numerical value expressing twist is intuitively obvious. Here we shall adopt the convention that a right-handed twist of 360 degrees has a value of +1 and a left-handed twist has a value of −1. The definition of twist is less obvious, however, for a ribbon whose axis is not straight. Perhaps the best way to understand this concept is to imagine a small arrow placed perpendicular to the axis of the ribbon, pointing to one of its edges [*see illustration below*]. As the arrow is moved along the twisting ribbon it rotates about the axis, and the twist of the ribbon can be defined as the integral of the arrow's angular rate of rotation with respect to the arc length of the axis curve.

In the special case where the axis of the ribbon is confined to a plane this value can be measured simply as the number of rotations the arrow completes about the axis as it is moved along the ribbon. For example, when the ribbon models a closed circular piece of DNA 5,000 base pairs long that is relaxed (that is, its axis lies in a plane), the arrow makes one complete rotation for every turn of the double helix, and so the total twist *Tw* equals +500, with the plus sign arising once again because the double helix is right-handed.

For a relaxed circular piece of DNA 5,000 base pairs long, then, both the linking number and the twist are equal to +500. From this example one might well assume that linking number is just another way of expressing twist, but that is not the case. Indeed, it is particularly important to understand the distinction between these two quantities. To begin with, linking is a topological property, whereas twist is geometrical: if a ribbon is deformed, its twist may be altered. Moreover, to compute the linking number (which is always an integer) the ribbon must be considered as a whole. On the other hand, twist (which might not be an integer) can be considered locally, and the twist values of individual sections can be summed to obtain the total for the ribbon.

It is important to note that this ribbon model does not come close to representing the detailed mechanical properties of DNA, not least because the mathematical ribbon is assumed to be of infinitesimal thickness. A real ribbon (with finite thickness) subjected to the helical manipulations we are concerned with would experience numerous local deformations that are not taken into account here. The ribbon model, however, was designed to emphasize the linking number of DNA, which because it is quantized is not affected by such small deformations.



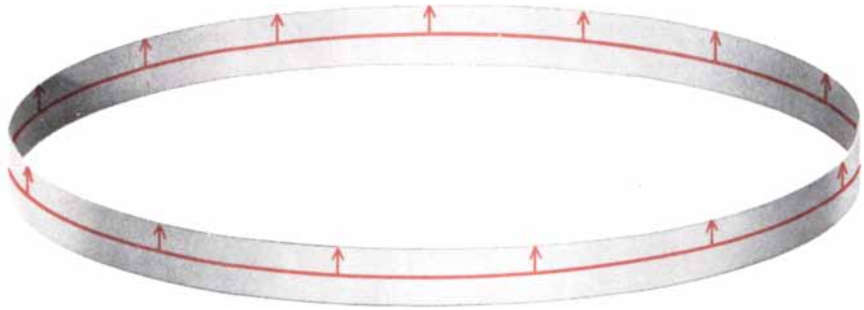
TWISTING OF A RIBBON can be assigned a numerical value by placing a small arrow on the ribbon perpendicular to its axis and pointing to one of its edges. As the arrow moves along the ribbon it rotates about the axis, and the magnitude of the total twist *Tw* can be defined as the integral of the angular rate of this rotation with respect to the arc length of the curve described by the axis. Whether this quantity is positive or negative depends on whether the rotation of the arrow about the axis is respectively right-handed or left-handed. When the axis of the ribbon lies entirely in a plane as is shown here, then the total twist is easily computed, being equal to the number of rotations the arrow makes about the axis. The twist can be computed separately for different parts of the ribbon and can then be summed to obtain the total value.

The realization that linking and twisting are distinct properties raises another question: Is there a geometrical significance to the difference between these properties, that is, to the difference between the linking number of a ribbon and its total twist? Quite fortuitously at about the same time that biochemists were first studying circular DNA mathematicians were independently taking a look at linking and twisting in ribbons. In 1968 one of us (White) proved that the linking number of a ribbon and its total twist differ by a quantity that depends exclusively on the curve of the axis of the ribbon. (This quantity is well known to mathematicians as the Gauss integral of the axis curve.) In other words, assume that the axes of two closed ribbons follow the same curve in three-dimensional space; then even if the ribbons themselves turn and twist in entirely dissimilar ways, their values of linking number and total twist will differ by exactly the same amount.

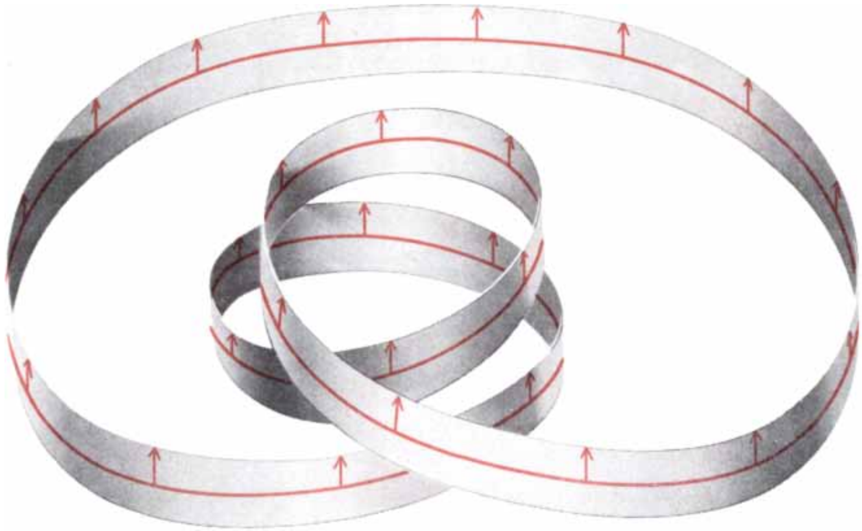
At about the same time Vinograd, not knowing of this result, asked F. Brock Fuller, a mathematician at Cal Tech, to tackle the mathematics of supercoiling. Fuller, while working on the relation between linking and twist, suggested the picturesque name writhing number for the quantity by which the two differ. Thus for a closed ribbon in three-dimensional space the writhing number *Wr* equals the difference between the linking number *Lk* and the total twist *Tw*, or $Wr = Lk - Tw$. The writhing number of a ribbon is a powerful quantity whose value generally changes if the axis of the ribbon is deformed in space. Hence writhing, like twisting, is not a topological property of the ribbon but a geometrical one.

The writhing number can be obtained by computing the Gauss integral, but it is generally far easier to calculate it by evaluating the linking number and the total twist of the ribbon in question and then taking their difference. It is only in certain special cases that it is convenient to compute Wr directly. For example, if the axis of a ribbon lies entirely in a plane or entirely on the surface of a sphere, then it can be shown that Wr is zero. Substituting this value into the equation $Wr = Lk - Tw$ gives $Lk = Tw$, which explains why in the example of the relaxed closed circular molecule of DNA both Tw and Lk were found to be $+500$. Now consider what happens if the axis of this DNA molecule is made to writhe in such a way that its writhing number is no longer zero. (There is no intuitive way to estimate the writhing number of a curve. A ribbon that writhes in the ordinary sense of the word may turn out to have an overall writhing number of zero.) When the writhing number of the molecule is made to change, the linking number remains the same (it can be altered only if one of the backbones of the double helix is broken) and so the twist must change. It is this relation that underlies the phenomenon of supercoiling.

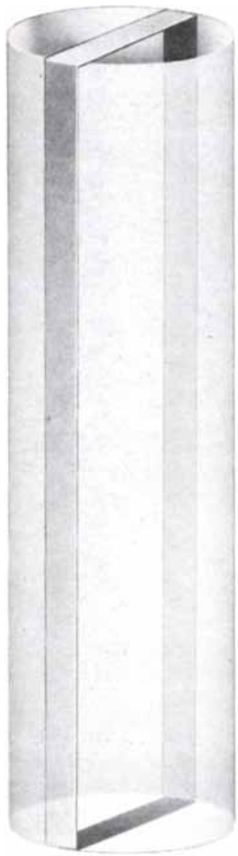
A ribbon's linking number, total twist and writhing number do not depend on the ribbon's location or orientation in space. They are also independent of scale, but if one axis of space is inverted (as is the case when the ribbon is reflected in a mirror) or three axes are inverted (as is the case when the ribbon is inverted through a point), then the sign of all three quantities is changed. On the other hand, if any two axes are invert-



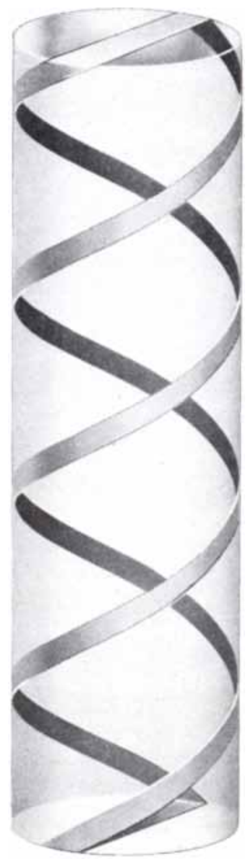
$Lk = 0$, $Tw = 0$, $Wr = 0$



$Lk = +2$, $Tw \sim 0$, $Wr \sim +2$

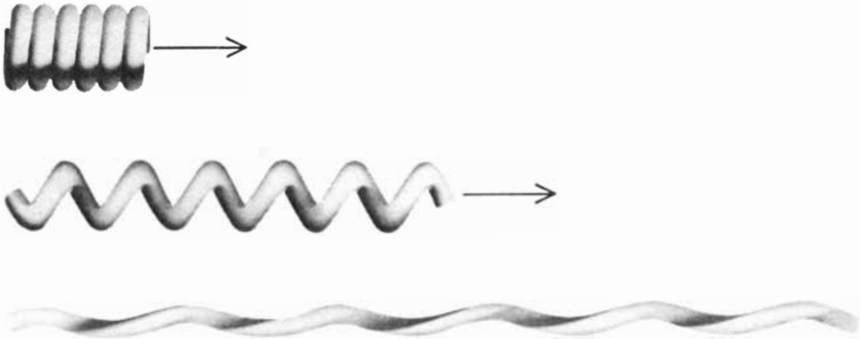


$Lk = 0$, $Tw = 0$, $Wr = 0$

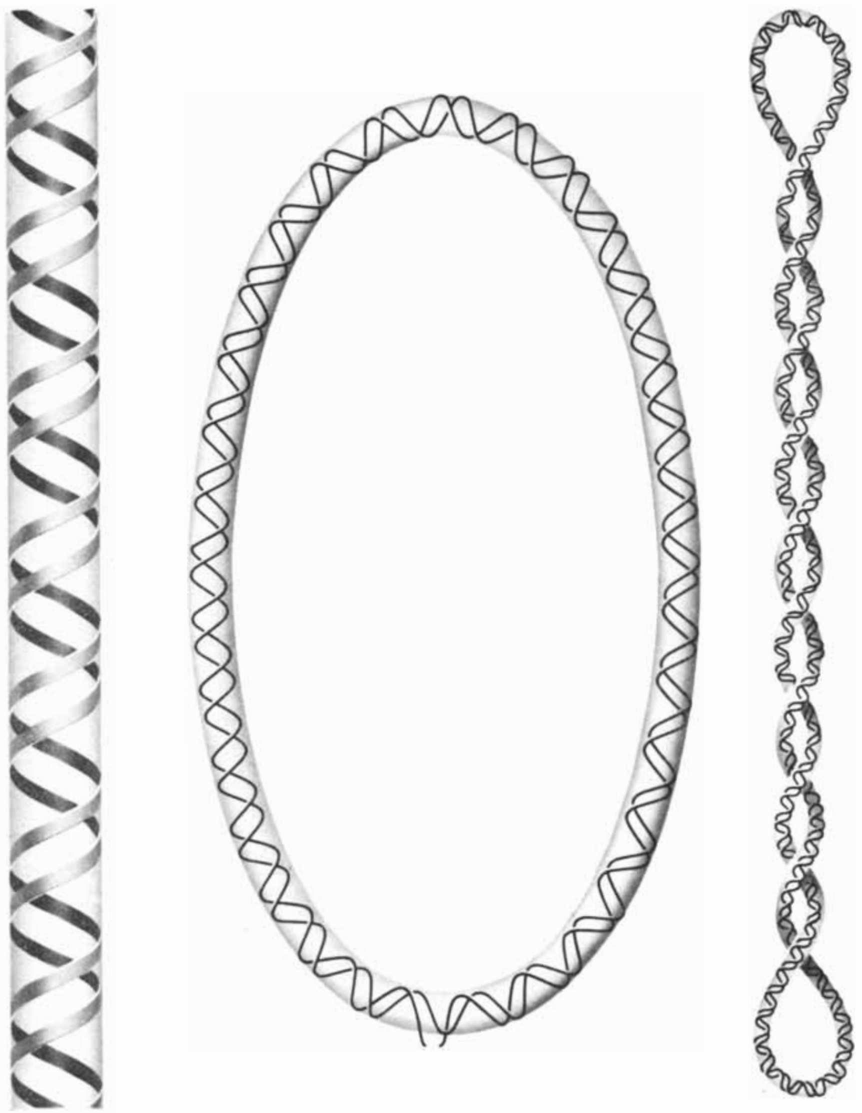


$Lk = 0$, $Tw > 0$, $Wr < 0$

LINKING NUMBER and total twist of a ribbon are equal when the axis of the ribbon lies entirely in a plane as is shown at the top. That these two quantities do not necessarily have the same value is shown by the configuration in the middle, in which the linking number *Lk* is a positive integer and yet the total twist *Tw* is close to zero. The linking number and total twist of a ribbon differ by a quantity called the writhing number *Wr*, that is, $Wr = Lk - Tw$. The writhing number depends exclusively on the curve of the axis of the ribbon. In the ribbon wound around the cylinder shown at the bottom left *Lk*, *Tw* and *Wr* are all zero. Rotating the top of the cylinder counterclockwise about its central axis through two full turns generates the right-handed interwound helical ribbon shown at the bottom right, in which *Lk* is clearly unchanged but (neglecting end effects) *Tw* equals $+4\sin\alpha$ and therefore *Wr* equals $-4\sin\alpha$. (In the first case the total twist *Tw* was made to be zero, and so the final configuration is a simplified model of supercoiled DNA, omitting the intrinsic twist that arises from the turns of the double helix.)



RELATION BETWEEN WRITHING AND TWIST is demonstrated with a coiled telephone wire. When the wire is in a relaxed state (*top*), its twist is small and its writhing is substantial: the axis of the wire traces a helix in three-dimensional space. When the wire is stretched out (*middle*) so that its axis is nearly straight (*bottom*), its writhing is small and its twist is large.



BASIC CONFIGURATIONS OF THE DOUBLE HELIX include a linear form (*left*), a circular form that is nicked and therefore relaxed (*middle*) and a circular form that is closed (unnicked) and therefore supercoiled (*right*). (The three molecules shown here are drawn to different scales.) The supercoiling of the closed circular molecule into an interwound superhelix can be understood in terms of the relation between linking, writhing and twist. Because molecule is underwound it has a deficit in linking number compared with a relaxed molecule of the same size. It compensates by writhing and by twisting and bending, satisfying equation $Wr = Lk - Tw$.

ed, as happens if one looks into an ordinary microscope (that is, not a dissecting one), their signs are unchanged. There are two other cases of practical interest. Making a contact print of a projection of the ribbon changes the signs of *Lk, Tw* and *Wr*. On the other hand, if the negative of a photograph of this projection is placed correctly in an enlarger, the resulting print will show a ribbon with the signs of all three quantities unchanged. Conversely, if the negative is turned over in the enlarger, the signs of *Lk, Tw* and *Wr* in the resulting print will be reversed.

In short, any operation that turns a right-handed screw into a left-handed one without introducing other distortions will change the sign of the linking number, the total twist and the writhing number. One of us (White) and Thomas F. Banchoff of Brown University have shown that there is one other special mathematical operation that changes the sign of these quantities but leaves their magnitude unaltered: inverting the ribbon through a sphere. This result explains why the writhing number of a ribbon whose axis lies on the surface of a sphere is zero. Under this operation the closed curve described by the axis is transformed into itself.

Although the equation $Wr = Lk - Tw$ demonstrates that linking and twist are mathematically distinct, the physical difference between these quantities may not yet be evident. It may be helpful, then, to consider what happens when a mathematical ribbon is wound around a cylinder in such a way that its surface is always flat against the cylinder [*see illustration on preceding page*]. (Note that it is only because a mathematical ribbon has infinitesimal thickness that it can be made to lie flat against the surface of a cylinder in this way.) We shall call the pitch angle of the helix described by this ribbon α. In other words, α is the angle at which each turn of the helix inclines away from the horizontal, so that when α is small, the helix is shallow, and when α is large, the helix is steep.

Now assume the ribbon is wrapped around the cylinder N times before its ends are joined in the most straightforward way. Then if end effects are ignored, it can be demonstrated that the linking number of the ribbon Lk equals N, whereas the total twist Tw equals $N\sin\alpha$. Therefore when the helix is stretched out so that the pitch angle α increases, the number of turns and thus the linking number remain the same, but the twist goes from a small value to a large one, clearly demonstrating the difference between linking and twist.

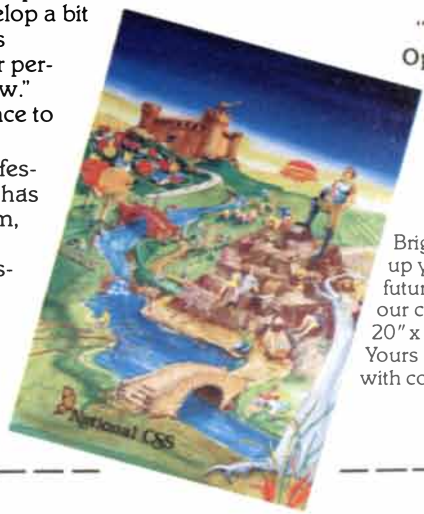
Moreover, since a ribbon's writhing number is defined as the difference between its linking number and its total twist, the value of Wr for this ribbon is $N - N\sin\alpha$, or $N(1 - \sin\alpha)$. As this formula indicates, when α is small and the

Launch Yourself...

. . . into a world where bold ideas are welcomed, and where innovation, creativity, and performance are valued and highly respected characteristics. As a DP professional — a programmer, tech. rep, systems engineer, or programmer analyst — you've probably come up with a lot of good ideas . . . ideas you may have wanted to develop a bit further just to see where they would take you. But usually, something always seemed to get in the way. Maybe it was a lack of resources or opportunity. Or perhaps someone told you "There just isn't enough time to try that idea right now." But those good ideas are still there, waiting . . . waiting . . . waiting for a chance to be tested.

Dare To Be Different . . . At National CSS we know what creative DP professionals are looking for, because since the beginning, information processing has been our only business. We know that by offering people a chance to perform, and by providing them with the resources and a receptive environment, that they will come up with the ideas that shape the future of information processing. That philosophy has made National CSS **the place** for DP professionals . . . a place where you can dare to be different and come out a hero.

Sure, you may not be looking for a new job right now, but . . . maybe not today, maybe not tomorrow, but someday you're going to have another one of those good ideas, and when that happens, we hope you'll remember the name National CSS . . . it could launch your career in a bright, new direction.



"Land of Opportunity" poster
FREE

Brighten up your future with our colorful 20″ x 28″ poster. Yours free with coupon.

I'd like to know more about National CSS. Please send me your free brochure and poster.

Name: ______________ *Address:* ______________

City: ______________ *State:* ______________ *Zip:* ______________

National CSS is an equal opportunity employer.

National CSS

Mail to: National CSS, Inc.
Dept. ADC 03074, P.O. Box 209,
West Haven, CT 06516

twist is small, the writhing is substantial, but when α is large and the twist is large, the writhing is minimal. The relation can be easily observed in a coiled telephone wire: when such a wire is unstressed, it assumes a highly writhed form with little twist; when its ends are pulled out, a highly twisted form that writhes only slightly is obtained.

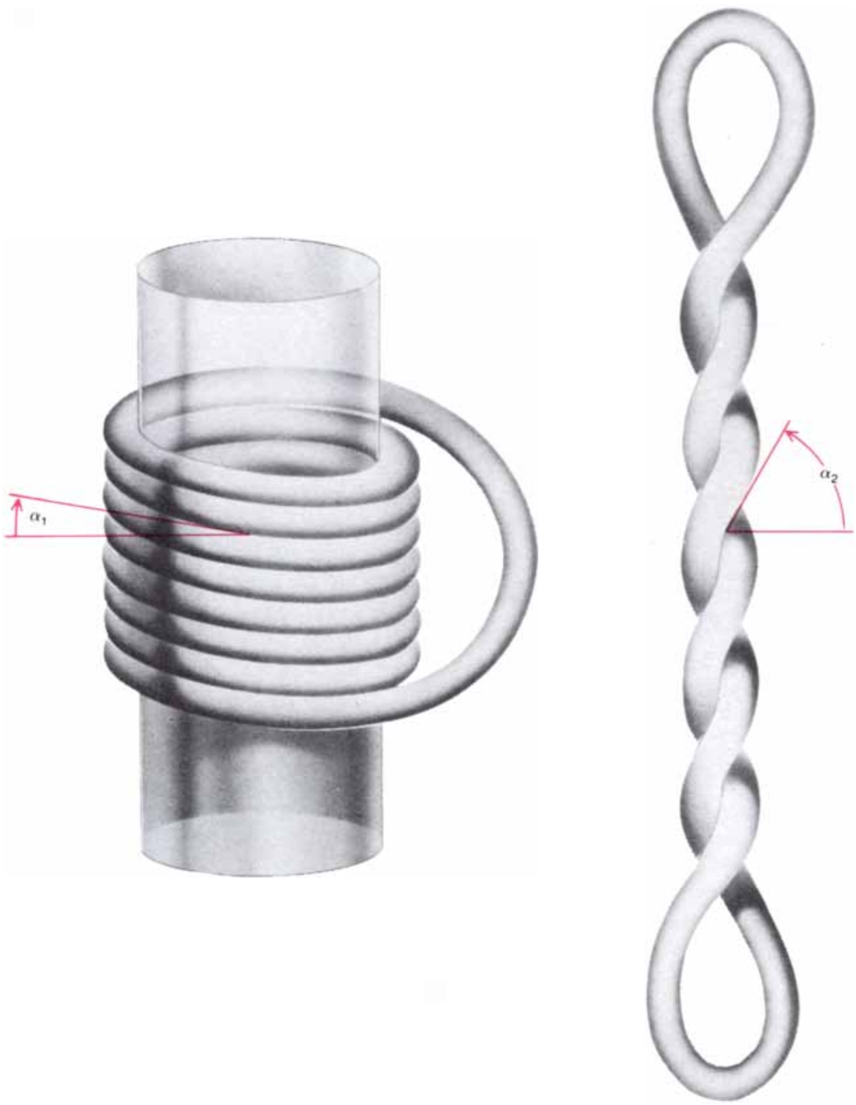
Consider how these findings apply to a real DNA, say to the polyoma-virus DNA. Remember that this DNA can be resolved through sedimentation into three components: I and II, which are circular, and III, which is linear. It can be determined experimentally that the average linking number for a population of relaxed circular molecules of this DNA is about +500. (Even in a highly purified preparation not all molecules are relaxed or supercoiled to the same extent and so the average linking number of a population is not necessarily an integer.) On the other hand, for a population of closed circular molecules (the supercoiled molecules that make up component I) the average linking number is about +475. Just as Vinograd predicted, the closed circular molecules of the polyoma-virus DNA are underwound, having a deficit in their winding number of about 25. This finding suggests a way to define supercoiling. It is equal to ΔLk, the difference between the linking number of a molecule in the natural closed circular state and the linking number of the same molecule in the relaxed closed circular state (where the energy of deformation is at a minimum and the writhing number is zero). For example, for the DNA's of both polyoma virus and the monkey virus SV40, ΔLk is approximately −25. (It is assumed that the linking number is measured with both the relaxed and the supercoiled DNA in solution under standard conditions of temperature, salinity and so on.)

It can now be understood why a deficit in the linking number of a molecule of DNA causes the molecule to supercoil. A linear molecule of DNA in solution normally assumes a form known as the *B* configuration, in which the nucleotide bases are approximately perpendicular to the helical axis with 3.4 angstrom units between them and in which there are about 10 base pairs for each turn of the double helix. This is a configuration of minimum energy, and if the molecule is bent or twisted, its energy is increased. If a long molecule is simply circular, however, the diameter of the circle is large compared with the thickness of the double helix. Hence the curvature of the molecule is small and its energy is increased only slightly. As a result nicked circular molecules such as component II of polyoma-virus DNA hardly depart from the *B* configuration.

The situation is quite different for a closed circular molecule with a deficit in linking number. To satisfy the condition that the value of *Lk* be less than that of a relaxed molecule (say 475 rather than 500) the double helix would have to be untwisted, a transformation that would substantially increase the deformation energy of the molecule. By supercoiling, however, the closed circular molecule minimizes the amount by which it departs from the *B* configuration.

More precisely, as the analysis of the ribbon model revealed, one way that underwound DNA can reduce its deformation energy is by writhing. Since writhing and twist are interconvertible, it is apparent that by changing the extent of writhing it is possible to minimize the twist of a molecule, thereby minimizing the twisting component of its deformation energy. On the other hand, writhing always introduces some curvature, and so it increases the bending contribution to the energy of the molecule. Therefore the supercoiled configuration that the underwound DNA molecule assumes is one that minimizes twist while introducing the smallest possible amount of bending.

Thus when the axis of DNA is made to writhe, the double helix responds by twisting and/or bending. The forms that this supercoiling can take range from a left-handed solenoidal superhelix to a right-handed interwound superhelix, although in nature the interwound



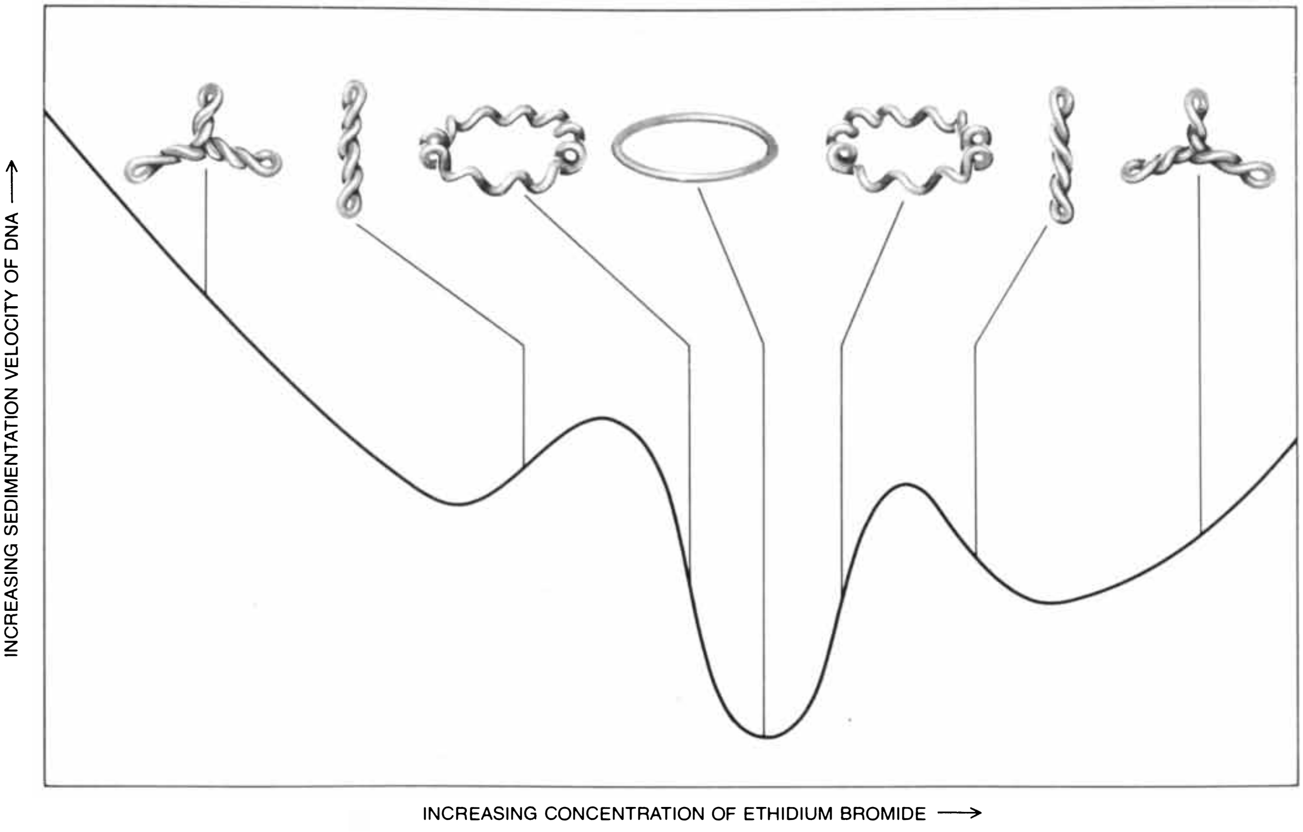
RUBBER TUBE that has been coiled around a cylinder with its ends joined in such a way that all twist is relieved (*left*) jumps into an interwound helix coiled in the opposite direction when the cylinder is removed (*right*). If there are N right-handed turns in the first configuration, and if the pitch angle α_1 at which each helical turn inclines away from the horizontal is small, then in the second configuration there will be approximately $N/2$ turns going up and $N/2$ going down and the new pitch angle α_2 will be large. (This will be only roughly true.) An examination of the changes in the values of the linking number, the total twist and writhing number that accompany this transformation explains why most naturally occurring DNA, having a deficit in linking number, supercoils into interwound helix that like double helix is right-handed.

configuration is usually preferred. The trade-off between writhing and twist that results in the favoring of the interwound form may be made clearer by a mechanical example. Consider a piece of thick, flexible rubber tubing wound into a left-handed coil around the surface of a cylinder. Assume that as the tubing is coiled it is allowed to twist freely about its own axis, so that its twisting energy is zero. Its ends are then brought together (in such a way that any strain due to twisting is relieved) and permanently joined. If at this point all restraint on the tubing is removed, however, it will jump into an interwound superhelix that is right-handed [*see illustration on page 126*]. Why?

Without undertaking a detailed analysis of the mechanics of this situation it is possible to gain a general understanding of how a left-handed superhelical coil that has little or no twist but is highly writhed is transformed into a right-handed interwound superhelix by considering the linking number, the total twist and the writhing number of the two configurations. (To visualize the meaning of linking number in this context imagine a pair of parallel lines drawn on opposite sides of the unwound relaxed tubing.) Remember that in the first configuration the total twist was deliberately made to be zero, and so (ignoring end effects) the linking number of the tubing must be equal to the writhing number. In the example of the helical ribbon wound around a cylinder we observed that the writhing number for such a configuration is given by the formula $Wr = -N_1(1 - \sin\alpha_1)$, where N_1 is the number of coils in the tubing and α_1 is their pitch angle. (The negative sign arises because the coils are left-handed.) Hence for the first configuration the linking number equals $-N_1(1 - \sin\alpha_1)$. (Because we are neglecting end effects this number will not usually be an integer.)

Turning now to the second configuration of the rubber tubing, the right-handed interwound superhelix, we shall call the total number of right-handed turns (up plus down) of the helix N_2 and its pitch angle α_2. Then (once again neglecting end effects) it can be shown that for such an interwound superhelix the writhing number Wr equals $-N_2\sin\alpha_2$. Note that the left-handed superhelical coil was transformed into the interwound superhelix without breaking the rubber tube or separating its ends. Therefore the linking number of the second configuration must be the same as the linking number of the first, that is, Lk must be equal to $-N_1(1 - \sin\alpha_1)$. Substituting these values of Wr and Lk into the equation $Wr = Lk - Tw$ yields $-N_2\sin\alpha_2 = -N_1(1 - \sin\alpha_1) - Tw$, or $Tw = -N_1(1 - \sin\alpha_1) + N_2\sin\alpha_2$.

The pitch angle of the first configuration of the rubber tubing α_1 was small, but the pitch angle α_2 of the second configuration was fairly large. Therefore matters can be further simplified by substituting the approximations $\sin\alpha_1 = 0$ and $\sin\alpha_2 = 1$ into the equation, yielding $Tw = -N_1 + N_2$. This simple formula explains why the rubber tube when it was released jumped into an interwound configuration of reversed handedness. To these approximations the twist of the tubing Tw is at a minimum when the turns in the second configuration are equal in number and opposite in handedness to those in the first configuration. Moreover, the curvature of the second



SUPERCOILS ARE RELAXED when the dye ethidium bromide, which has a planar molecular structure, is added to supercoiled DNA in solution, as is demonstrated by this graph showing how the sedimentation velocity of DNA changes as the dye is added. The sedimentation velocity is the rate at which the molecules of DNA move through the solvent when the solution is subjected to the strong gravitational field generated by an ultracentrifuge: more highly supercoiled molecules are more compact, and so they have a greater sedimentation velocity. Increasing the concentration of ethidium bromide serves to gradually decrease the writhing of the DNA molecules, not by reducing the deficit in their linking number (that can be done only by creating nicks in their polynucleotide chains) but by reducing their twist. Molecules of the dye are intercalated between the base pairs of the DNA, creating a local untwisting of the double helix. The graph shows that with a sufficiently high concentration of dye the DNA molecules become fully relaxed, and that if more dye is added, the DNA begins to supercoil in the opposite direction. Changes in the structure of DNA shown here are confirmed by electron microscopy.

configuration is clearly less than that of the first. Therefore when the twist is small, the interwound superhelix will have an energy of deformation much lower than that of the superhelical coil, and so the rubber tube will of course assume the interwound form. (To apply this argument to DNA we should have to paint the double helix onto the unwound relaxed rubber tubing, but that would not affect the energy calculation.)

To make a detailed analysis of the supercoiling of real DNA requires a fairly precise knowledge of all its elastic constants, and it is necessary to take into account not only end effects but also such matters as charge repulsion and thermal motion. The preceding arguments show quite clearly, however, why a closed circular DNA molecule, with a deficit in its linking number (or as is sometimes the case an excess) will writhe into a supercoil. In nature most closed circular DNA is negatively supercoiled, that is, the supercoiling results from a deficit in linking number. The analysis of the rubber-tube model explains why such molecules can be expected to assume the configuration of an interwound superhelix whose handedness, rather surprisingly, is the same as that of the double helix, namely right-handed.

The backbones of a double helix have enough flexibility so that if in a supercoiled molecule one of them is nicked, the other can rotate about it to put the molecule in a relaxed configuration. The un-nicked molecule cannot lose its supercoiling in this way without one intact backbone's passing through the other, which is a physical impossibility. Given a naturally supercoiled molecule of DNA with a deficit in its linking number, however, is it possible for its writhing to be reduced without one of its backbones' being nicked? If the backbones must remain intact, then the linking number cannot change, and since the writhing number is negative in this case, the only way for the magnitude of the molecule's writhing to be reduced is for its total twist to be reduced.

Adding ethidium bromide, a dye with a planar molecular structure, to a solution of DNA serves this purpose. Leonard S. Lerman of the State University of New York at Albany was the first to show that molecules of this type are intercalated between the base pairs of the double helix. In supercoiled DNA such insertion causes a local untwisting of the double helix. Indeed, a sufficiently high concentration of ethidium bromide will completely relax the molecule.

More precisely, this critical concentration will reduce the total twist of the molecule. For example, in the presence of the dye a linear piece of polyoma-virus DNA would have a total twist of 475 rather than 500. Therefore a closed circular molecule of that DNA with a linking number of 475 would in the presence of this critical amount of dye have a total twist equal to its linking number, so that its writhing number would be zero. The axis of the molecule would cease writhing. Furthermore, if the concentration of ethidium bromide is increased past the critical point, the total twist will be still further reduced, causing the closed circular molecule to writhe into an oppositely oriented supercoil.

These effects can be observed directly by studying the sedimentation velocity of DNA molecules, as was first shown by Lionel V. Crawford of the Imperial Cancer Research Institute in Britain and Michael J. Waring of the University of Cambridge and independently by Vinograd and one of us (Bauer). Adding ethidium bromide to closed circular polyoma-virus DNA in solution causes the molecules of DNA to writhe less, so that they become less compact and sediment slower. As more ethidium bromide is added their sedimentation velocity reaches a minimum; then as higher concentrations of the dye cause the molecules to supercoil in the opposite direction and become more compact the sedimentation velocity begins to rise again.

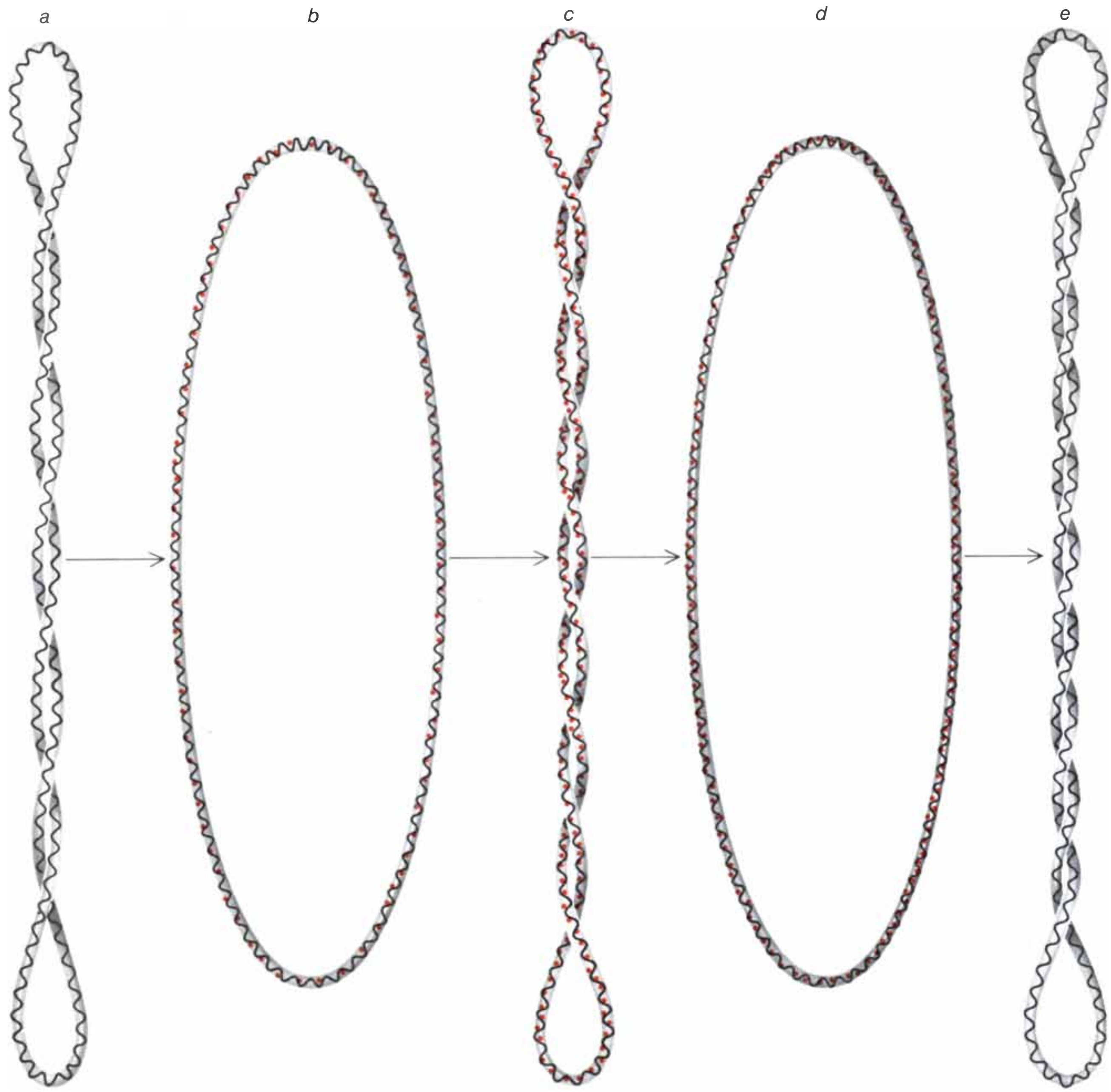
To create DNA molecules with different degrees of writhing it would be convenient to be able to make a nick in one of the backbones of a DNA double helix, relax the molecule by a few turns and then close the nick. Astonishingly, enzymes have been identified that do just that. The first of them, an enzyme known as ω-protein, was discovered in *Escherichia coli* by James C. Wang of Harvard University in 1971. Similar enzymes have now been discovered in a variety of sources, including other bacteria, the vaccinia virus and the nucleus and mitochondria of animal cells. These nicking-closing enzymes, which are also called topoisomerases, generally require no energy source to function. They always act to reduce the supercoiling of a DNA molecule and thereby lower its energy.

Nicking-closing enzymes have proved to be invaluable tools for studying the physical chemistry of DNA. To begin with, adding a topoisomerase to a solution of supercoiled DNA gradually reduces the supercoiling until all the DNA molecules are in or near the relaxed state, but much more ingenious applications have been devised for these enzymes. For example, suppose a high concentration of ethidium bromide is added to closed circular DNA, causing it to supercoil in the direction opposite to its usual one. If a topoisomerase is then added as well, the molecules in the solution will become relaxed, but because of the large amount of intercalated ethidium bromide they will be considerably more underwound than they are in their natural state. Now, if first the

topoisomerase and then the dye is removed, the molecules will compensate for the additional reduction of the linking number by supercoiling even more strongly in the usual direction. In this way super-supercoiled molecules can be created, molecules that have many more superhelical twists than are usually present. This type of energy storage and transfer may be central to the role of supercoiling in the cell. We can even conceive of a chemical engine that by going through a similar cycle of operations turns chemical energy into work.

It is also interesting to note that Martin Gellart and his colleagues at the National Institutes of Health have found an enzyme in the bacterium *E. coli* that given an energy source such as adenosine triphosphate (ATP) will reduce the linking number of relaxed circular DNA, thereby increasing its writhing number and making it supercoil. The activity of an enzyme of this type, called a gyrase, is essentially opposite to that of the nicking-closing enzymes. So far gyrases have been found in a variety of microorganisms but not in any higher organisms.

Nicholas R. Cozzarelli and his colleagues at the University of Chicago have recently shown that a gyrase seems to act by bringing two segments of the DNA molecule close together. The en-



SUPER-SUPERCOILED MOLECULES OF DNA, with more superhelical turns than are normally present, can be created by employing ethidium bromide (*colored dots*) in conjunction with a nicking-closing enzyme. An enzyme of this type acts to reduce the supercoiling of a DNA molecule by making a temporary nick in one of its sugar-phosphate backbones, relaxing the molecule and then reclosing the nick. Therefore if enough ethidium bromide is added to supercoiled DNA in solution (*a*) for the DNA to become first completely relaxed (*b*) and then supercoiled in the opposite direction (*c*), the addition of the nicking-closing enzyme will bring the DNA back to the relaxed state (*d*). A substantial amount of dye will be bound to the DNA, however, so that the DNA will be more underwound than is usual. Hence if the enzyme and then the dye are removed, the DNA will supercoil more strongly than usual in its original direction (*e*).

zyme then cuts both of the backbones of one of the segments and passes the other intact segment through the resulting gap before it rejoins the originally cut backbones. It is easy to show that such a process would alter the linking number of the DNA in increments of two rather than one, and that is exactly what has been observed experimentally.

How is supercoiling measured? One experimental method is based on a discovery made by Walter Keller of the University of Heidelberg in 1974. Keller observed that if supercoiled molecules with even slightly different linking numbers are moved through an agarose gel by electrophoresis, they do not all travel at the same rate. The more compact molecules travel faster, and so a pattern of discrete bands is created in the gel. Moreover, as was later shown by Vinograd, if one of these bands of molecules is cut out of the gel and rerun, the result is again a single band in exactly the same location. In fact, as was shown later by Wang and one of us (Bauer), even if the band is heated to near melting (so that the bases in the double helixes become unpaired) and then is slowly cooled (so that the double helixes are rejoined), it will still form a single band in the same location.

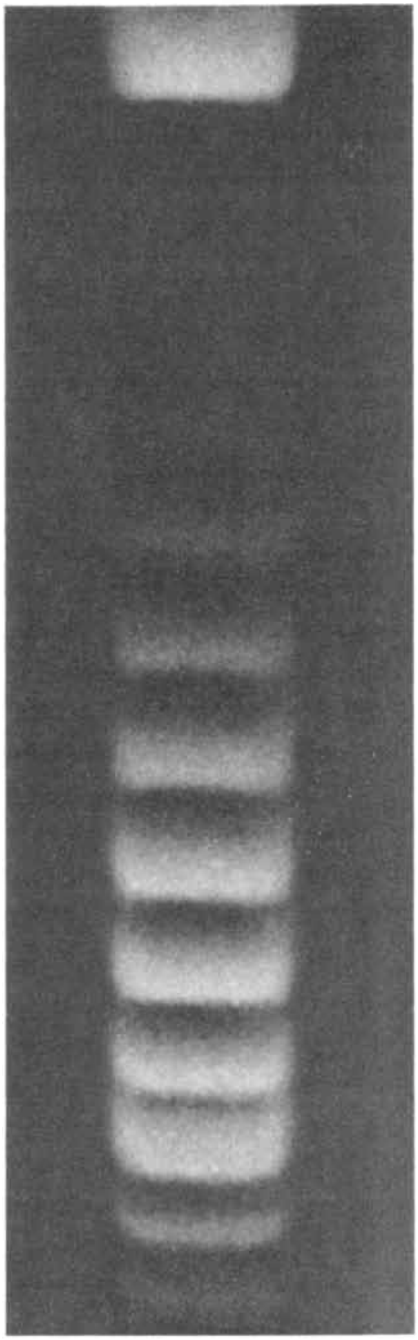
Hence the molecules in adjacent bands differ by a quantity that survives heating and cooling and, since the bands are discrete, is quantized. This quantity can only be the linking number. The only reasonable interpretation is that the linking numbers of the molecules in adjacent bands on the gel differ by a value of 1. Therefore when a population of DNA that includes molecules in all the states from relaxed to supercoiled is run on the gel, ΔLk, the difference in linking number that determines supercoiling, can be evaluated directly by counting the bands. This experimentally determined value for ΔLk can be put to good use. For example, it can be applied to calibrate the untwisting effect of ethidium bromide on the double helix. It turns out that every intercalated molecule of ethidium bromide produces a local untwisting of about 26 degrees, and so ethidium bromide can be utilized as a subsidiary standard for measuring the value of ΔLk for any supercoiled molecule. (It should also be noted that the striking effect of the enzyme gyrase on the linking number of DNA can be observed clearly in gel experiments.)

One of the most important results to emerge so far from the study of supercoiled DNA is the direct proof that the two backbones of DNA do indeed wind around each other at intervals of about 10 base pairs. Here the most elegant experiments are those that have been done by Wang. He compared similar circular DNA molecules constructed by the methods of genetic engineering, each having about 5,000 base pairs. The exact length of these molecules was not known, but the difference in length between any two of them was known precisely. Some of them differed by only a single base pair, others by as many as 400 pairs.

The differences in the extent of supercoiling observed when these different-sized molecules were relaxed by means of a nicking-closing enzyme and then run side by side on a gel enabled Wang to deduce that the number of base pairs for each turn of the double helix must be 10.4 plus or minus .1. His experiment was a pretty demonstration of the precision and sophistication of modern techniques of molecular biology. Moreover, several models for DNA have recently been proposed in which the two polynucleotide chains do not coil around each other to form a double helix but instead lie side by side over most of their length, having only a few helical turns. Wang's results, together with those from related experiments, show that these new models must be incorrect. This topological argument is very powerful in that it eliminates all models of the side-by-side type, regardless of their molecular detail.

How does supercoiling arise? In the SV40 virus and the polyoma virus the DNA is supercoiled because it is usually not naked in the nucleus of the cell. During replication the double helix is wound on nucleosomes (beads of protein consisting of eight histone molecules with one or two associated molecules). When the DNA molecule slips off this supporting structure, it supercoils (in much the same way that the coiled rubber tubing discussed above does). The fact that the linking number is always reduced in naturally occurring DNA implies that when DNA winds around the nucleosomes in a solenoidal coil, the coil is left-handed. Uncertainties in the available data and arguments about their interpretation make it difficult to determine the exact number of supercoils that are generated by each nucleosome, but it appears likely that the number will turn out to be between 1 and 2. Nucleosomes are found in association with DNA in all higher organisms; in lower organisms the origins of supercoiling are not completely clear.

Much work remains to be done before all the implications of supercoiling in DNA can be understood. In particular the enzymes that are related to supercoiling (topoisomerases and gyrases) and the role of supercoiling in the replication of DNA are currently the subject of intense research efforts. It is clear, however, that the present knowledge of supercoiling, both experimental and theoretical, provides a sound basis for the investigation of this surprising manifestation of the double helix.



EXPERIMENTAL TECHNIQUE for measuring supercoiling involves moving a population of closed circular DNA molecules through an agarose gel by electrophoresis, in this case from top to bottom. Because the more highly supercoiled molecules are more compact they move faster through the gel. As a result a set of discrete bands of molecules is created such that the molecules in each band all have the same linking number and the numbers associated with adjacent bands differ by a value of 1. (The DNA shown here has been stained with a fluorescent dye, and so the bands containing more molecules are brighter.) If the population includes molecules with various degrees of supercoiling, as is shown here, counting the resulting bands provides a direct measure of the deficit in linking number that determines degree of supercoiling. Uppermost band in gel contains nicked DNA.

The Screw Propeller

Ship screws and aircraft propellers look quite different, but both are designed according to the circulation theory of lift. The theory is the traditional one developed to account for the lift of an airfoil

by E. Eugene Larrabee

In 1845, early in the history of steam navigation, the effectiveness of the screw propeller was established when the steam sloop *Rattler,* equipped with such a propeller, towed the steam sloop *Alecto,* equipped with paddle wheels, stern first at a speed of nearly three knots in a tug-of-war staged by the British Admiralty. The closely matched ships were driven by slow-turning steam engines of about 200 horsepower. The test was significant in that it demonstrated the validity of screw-propeller propulsion in the face of the intuitively more understandable concept of the paddle wheel.

The screw propeller has remained the most generally useful means for driving watercraft of all types and sizes. The chief exceptions are a few special-purpose craft such as shallow-draft or hydrofoil boats. And when it comes to moving a piston-engine airplane through the air, there is no alternative to the propeller. The propeller has also been retained on turboprop aircraft powered by gas turbines. Most commercial airliners, however, are now driven by turbofan engines, in which the propeller goes by another name. In a turbofan engine the high-velocity gas output from the fuel-combustion chamber flows through a power turbine that drives a "fan," or axial compressor, at the front of the engine. The fan is really a many-bladed screw propeller designed to operate within the engine case. The fan compressor provides the air for combustion and also creates a low-velocity, high-volume flow of air that bypasses the combustion chamber and mixes with the hot exhaust of the power turbine to propel the airplane. In the turboprop engine the power turbine turns an external visible airscrew that provides most of the thrust. As we shall see, the turboprop, by virtue of its efficiency, may be about to rise again.

Anyone who has looked closely at a motorboat propeller and a typical airplane propeller is aware that the blades of a marine screw fill a much larger fraction of the total disk area, the area swept out by the propeller blades, than the blades of an airscrew do. The aeronautical engineer would say that the marine screw has greater "solidity." An extreme contrast can be seen if one compares the marine screw, or "wheel," of an ultralarge crude carrier (oil tanker) with the airscrew of the *Gossamer Albatross,* the ultralight aircraft that Bryan Allen pedaled across the English Channel on June 12 of last year. The propellers of the largest tankers have five or six blades, a diameter of up to nine meters, turn at about 95 revolutions per minute and can absorb 45,000 h.p. to drive a ship of more than 500,000 deadweight metric tons at 16 knots (8.23 meters per second). The twin screws of the 31-knot *Queen Elizabeth 2,* each of which absorbs 55,000 h.p., are appreciably smaller, turn much faster and have an even greater solidity than the screws of the largest tankers.

The *Gossamer Albatross* propeller has two slim blades and a diameter 70 percent as large as the screws of the *Queen Elizabeth 2;* when the 4.1-meter propeller is turned at 95 r.p.m., it absorbs .25 h.p. to drive an airplane with a flying weight of 96 kilograms at 5.4 meters per second. Marine screws and airscrews are both the result of similar computer analyses that account for the fluid density, the interaction of the flow fields, the hydrodynamic or aerodynamic load on the blades and finally the practical limitations imposed by the vehicle and its mode of operation. Why are the geometries of the two kinds of propeller so different?

It is not because air is a gas and water is a liquid. Strangely enough, the *Gossamer Albatross* propeller comes much closer than the ship screw to the ideal form of a propeller of highest efficiency adapted to operation in a nearly incompressible fluid of low viscosity, that is, a fluid such as water. Its shape conforms to theoretical considerations set forth by Albert Betz and Ludwig Prandtl of the Kaiser Wilhelm Institute for Fluid Dynamics Research at Göttingen in 1919 and refined by a visiting British scholar, Sydney Goldstein, in 1929. Their ideas came out of the revolution in theoretical hydrodynamics at the beginning of the century, when Prandtl invented boundary-layer theory, which explained the drag, or resistance, of streamlined bodies, and when W. M. Kutta of Germany and Nikolai Joukowsky of Russia independently invented the circulation theory of lift, which explained the lift of wings and propeller blades.

Until then theoretical hydrodynamics, which had been invented to study the resistance of ship hulls in motion, had been largely an academic discipline concerned with the mathematical flow patterns that could be formed by various combinations of flow fields. Although the streamlines of some of the combined mathematical flow fields could be made to resemble the flow around a ship hull, Jean Le Rond d'Alembert had shown in 1742 that the theoretical resistance of such a flow on the mathematical hull form was exactly zero. Moreover, if the hull form was mathematically yawed (turned at an angle to the flow field), the hydrodynamic force perpendicular to the direction of hull motion was also zero. Since real hulls do not exhibit zero response to real flows, these findings became known as d'Alembert's paradox.

In the absence of a useful theory 19th-century naval architecture became an experimental science and marine propellers were developed intuitively. The propeller was likened to a machine screw that advances as it turns in a threaded hole. Unlike metal, however, water yields under the thrust of the propeller. Hence the "effective pitch," or the distance traveled in one revolution, is less than the "geometric pitch" of the propeller by an amount called the slip. In 1865 the Scottish engineer William J. M. Rankine developed an "actuator disk" theory of the propeller, later improved by William Froude, in which the

PROPELLER OF "QUEEN ELIZABETH 2" is one of two screws that together absorb 110,000 horsepower in driving the liner of 67,000 gross tons at a top speed of 31 knots. The propeller is 5.8 meters in diameter and generates maximum thrust when it is turning at 174 revolutions per minute. The screw's diameter is sharply constrained by the requirement that it not project below the keel or beyond the vessel's beam. In order to achieve the required thrust within the limited diameter the total blade area is 89.5 percent as large as the area of a solid disk of the same diameter. In marine terminology the ratio of blade area to disk area is called the expanded area ratio; in aeronautics the same ratio is called solidity. The *Queen Elizabeth 2* propeller was designed and made by SMM Propeller Ltd. of Liverpool.

PROPELLER OF "GOSSAMER ALBATROSS," the ultralight aircraft Bryan Allen pedaled across the English Channel on June 12, 1979, bears only a remote resemblance to the screw of *Queen Elizabeth 2* shown at the top of the page. Both propellers, however, act as "pushers" and are among the most efficient of their type ever designed. The airscrew of the *Gossamer Albatross* is 4.1 meters in diameter, rotates at 95 r.p.m. and absorbs .8 h.p. at takeoff and .25 h.p. in cruising flight. The propeller was computer-designed with the author's algorithms to have minimum energy in its vortex wake and minimum friction loss in slightly climbing flight. The *Gossamer Albatross* was designed by AeroVironment Incorporated of Pasadena, Calif. It is shown here at the Johnson Space Flight Center in Houston, Tex.

Physician, did you miss any of these significant developments in medical science?

• **The isolation of coxsackievirus B4 from the damaged pancreatic cells of a 10-year-old boy who died of diabetic ketoacidosis adds to the growing evidence that implicates viruses in some forms of insulin-dependent diabetes mellitus (IDDM).**

• ***Campylobacter fetus*** **subsp.** ***jejuni*** **causes at least as much acute gastroenteritis as does** ***Shigella*** **or** ***Salmonella;*** **the culturing and antimicrobial therapy for** ***C. fetus*** **subsp.** ***jejuni*** **differs from that for the other two bacteria.**

• **Miconazole shows great promise in patients with disseminated and meningeal coccidioidomycosis who do not respond to or cannot tolerate amphotericin B.**

• **Pittsburgh pneumonia agent (PPA) and atypical** ***Legionella*****-like organisms (ALLO) are particularly dangerous in immunosuppressed patients.**

• **Lithium batteries for pacemakers have made mercury-zinc cells virtually obsolete; they can power a unit for 8 to 20 years. Microcircuitry has revolutionized the therapeutic and diagnostic potential of these devices.**

Do some of these advances sound unfamiliar? With 2,000 or more journals to scan and innumerable meetings to attend, significant information slips by all too easily. Textbooks are obsolete before they are published.

SCIENTIFIC AMERICAN *Medicine* is the definitive remedy for such chronic information loss.

Comprehensive. The most advanced treatment of medicine available, SCIENTIFIC AMERICAN *Medicine* is written by leading scholar-clinicians from Harvard and Stanford for the seasoned clinician.

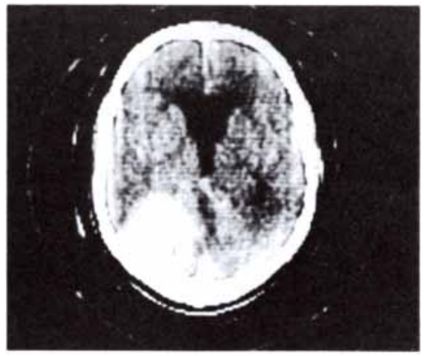
Up-to-Date. SCIENTIFIC AMERICAN *Medicine* keeps you up-to-date. Each month, you receive five to eight new chapters (and a new index), reporting significant new advances.

Complimentary CME. At no extra cost, we offer you an "office study" CME program of eight patient management problems that earn more top credits* than you can get from most week-long courses.

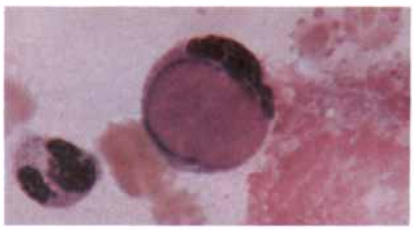
Trial Offer. We would like to put this superb information instrument at your disposal for two months, at no cost. On receipt of the coupon we will send you the two-volume text and two monthly updates. You may also take a CME test for credit. At the end of 60 days, if you decide to continue the subscription, you will owe us $185 for the full 12 months (renewal is $160); otherwise, just return the volumes to us. Please mail the coupon today and let us take the hassle out of keeping up.

*As an organization accredited for continuing medical education, the Stanford University School of Medicine designates this continuing medical education activity as meeting the criteria for 32 credit hours in Category 1 for Educational Materials for the Physician's Recognition Award of the American Medical Association, provided it has been completed according to instructions.

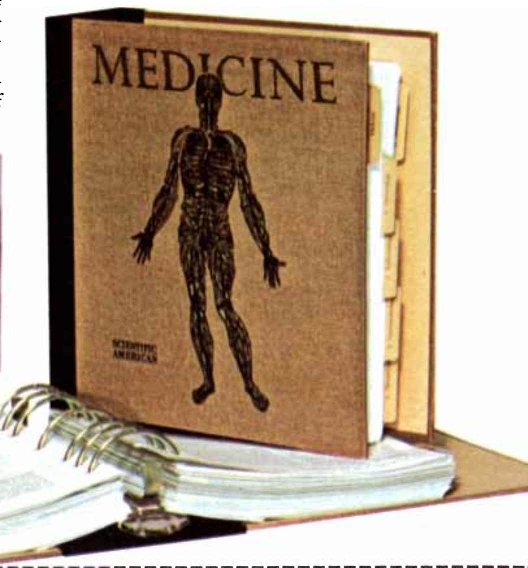
This program is approved by the American Academy of Family Physicians for 32 elective hours of CME credit.



CT scan reveals large meningioma and mild hydrocephalus.



A lupus erythematosus cell.



SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, New York 10017

Please enroll me as a subscriber to SCIENTIFIC AMERICAN *Medicine.* On receipt of this coupon you will send me the advanced two-volume text described in your announcement and update it regularly by sending me new monthly subsections. I understand that the annual subscription of $185 for SCIENTIFIC AMERICAN *Medicine* is tax deductible. If I am not entirely satisfied, I may cancel at any time during the first 60 days, returning all materials for a *complete refund.*

☐ Please enter my subscription for SCIENTIFIC AMERICAN *Medicine*
☐ I shall also enroll in the CME Program
☐ I enclose a check made out to SCIENTIFIC AMERICAN *Medicine* for $185*
☐ Please bill me

*Please add sales tax for California, Illinois, Massachusetts, Michigan, Ohio, and New York

Name____________________ MD Specialty__________

Address______________________________

City______________ State__________ Zip__________

Signature______________________________

Subscribers outside of the U.S. and possessions will be charged extra for shipping and handling; updates will be sent by surface routes unless airmail delivery is requested. Please allow 6-8 weeks for delivery.

10

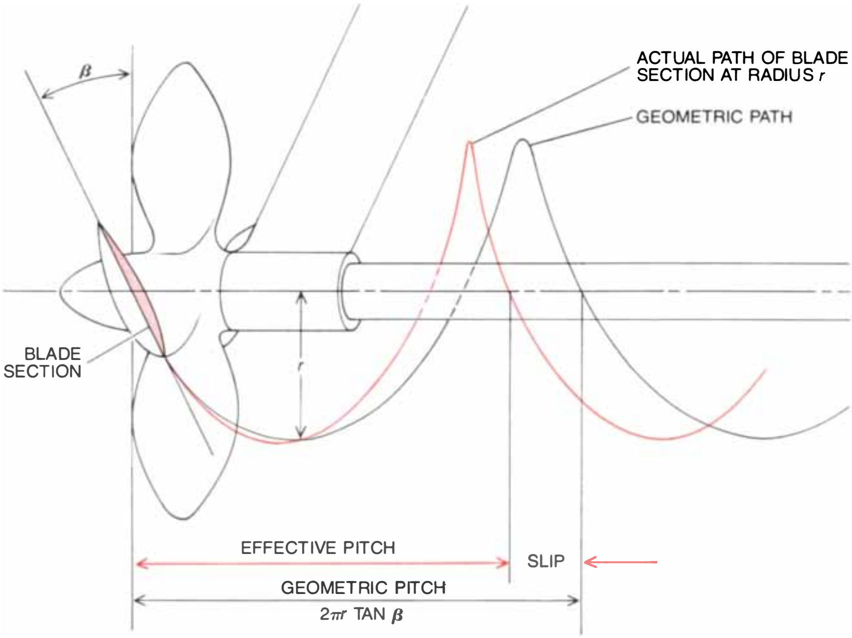
slip could be explained as the consequence of a simple increase in the momentum of the fluid in the propeller slipstream, an increase proportional to the thrust of the propeller.

In this theory the slip is half the ultimate increase in slipstream velocity (which depends on the thrust per unit disk area and the speed of the ship) multiplied by the time for one propeller revolution. The geometric pitch must be increased accordingly when one is trying to arrive at the effective (or desired) pitch, which is the ratio of the ship speed in feet per minute to the shaft speed in revolutions per minute. Although all of this is almost true experimentally, the actuator-disk theory gave no insight into why the blades should develop hydrodynamic thrust loads and torque loads in the first place. It did, however, give a precedent for the idea of "inflow" velocity, which would later be identified with the "induced" velocity of wing theory, and it associated an excessive slip (and a high disk loading) with reduced efficiency.

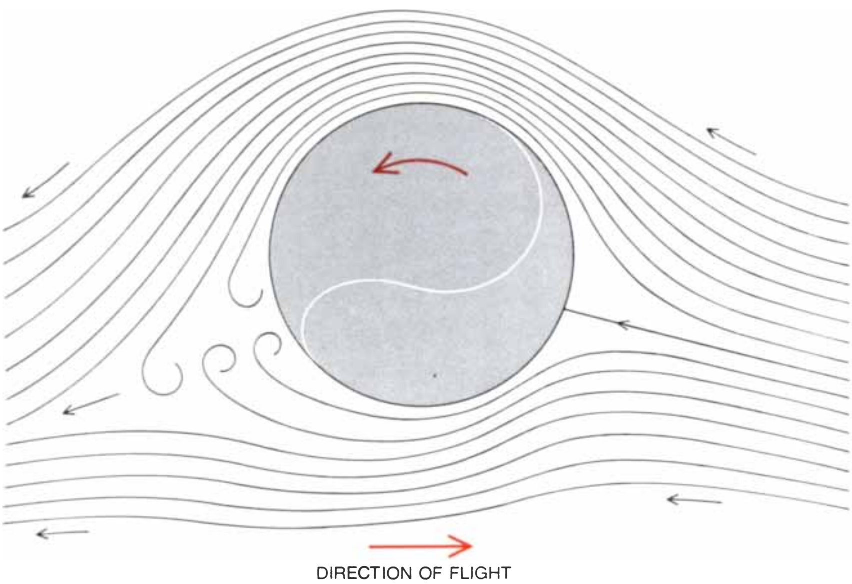
An explanation of thrust and torque loads awaited the development of the "vortex," or circulation, theory of lift by Kutta and Joukowsky in the period between 1902 and 1911. In Joukowsky's version of the theory the ideal flow of an inviscid fluid around a circular cylinder is conformally "mapped," or transformed, into the corresponding flow around an airfoil with a sharp trailing edge. Vorticity, or circulation, is added to the flow around the cylinder by introducing an idealized "bound," or "line," vortex of arbitrary strength coincident with the center of the cylinder. The strength of the vortex is adjusted mathematically until the flow around the cylinder stagnates at a singular point that coincides in the transformation with the trailing edge of the airfoil.

The hypothetical circulation of inviscid fluid around a cylinder of infinite length superficially resembles the flow of real fluid around a rotating sphere with a roughened surface such as a tennis ball or golf ball moving with backspin. Lift is produced as a consequence of the acceleration of the air flowing over the top of the ball and the deceleration of the air flowing under the ball. In accordance with a well-known theorem of the 18th-century Swiss mathematician Daniel Bernoulli, the fluid in the accelerated flow has a lower static pressure than the fluid in the decelerated flow, so that the ball is pushed upward from below. (A tennis ball hit with topspin receives a corresponding downward push from above.) Although ideal inviscid flow fields satisfy Bernoulli's theorem, there can be no lift without circulation. As a result the usual secondary-school explanation of airfoil lift in terms of Bernoulli's theorem is entirely inadequate.

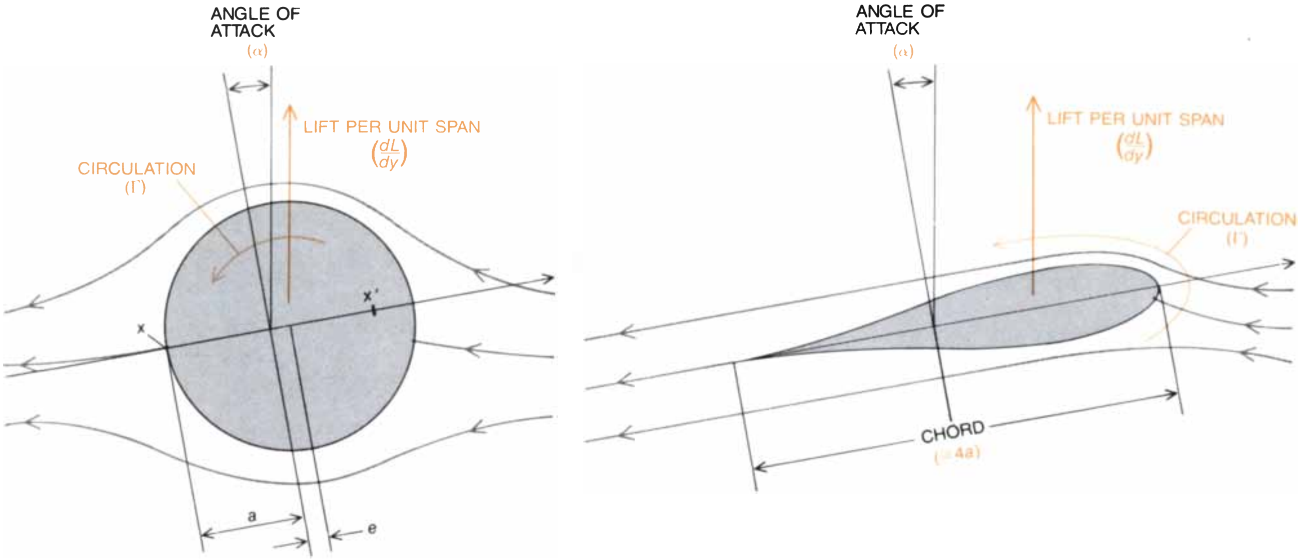
In that explanation lift is created because the air passing over the curved upper surface of a typical airfoil must travel farther, and therefore faster, than the air passing under the flatter lower surface. Hence the pressure above the wing must be less than the pressure below the wing, resulting in lift. Actually lift can be generated by a perfectly symmetrical airfoil that is tilted upward at a suitable angle of attack. How does the airstream know how to divide at the



SLIP OF MARINE PROPELLER was recognized early in the history of marine screws. The distance actually traveled by a ship during one revolution of the propeller, the "effective pitch," is less than the geometric pitch determined by the propeller's blade angle. William J. M. Rankine and William Froude explained that slip corresponds to the increase in axial momentum imparted to slipstream by propeller. They showed that excessive slip leads to low efficiency.

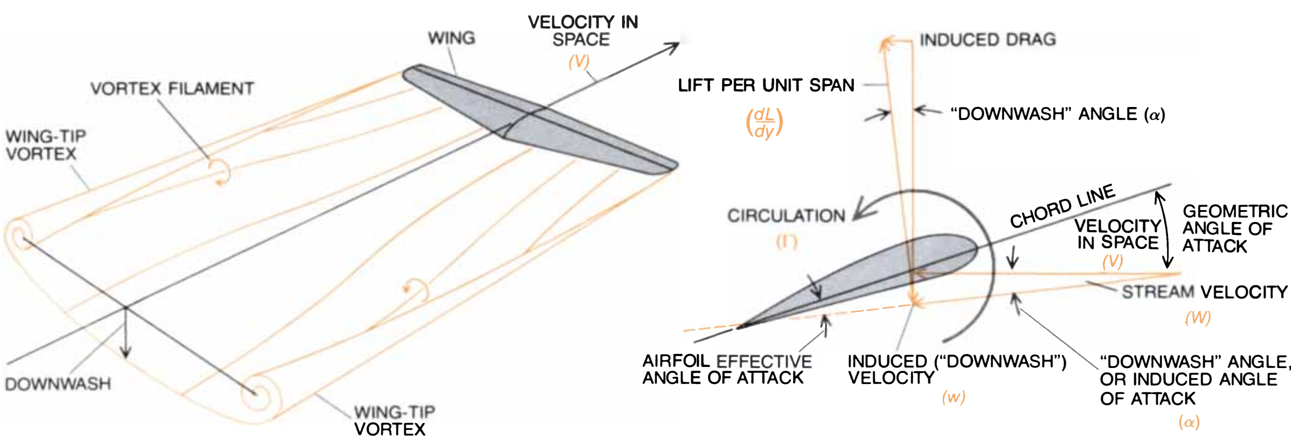


TENNIS BALL HIT WITH BACKSPIN mimics an airfoil by distorting the flow field in a way that creates aerodynamic lift. Perhaps the earliest explanation of the flight of a "cut" tennis ball is attributed to Lord Rayleigh. Because of backspin the air flowing over the top of the ball is accelerated to the rear and the air flowing under the ball is retarded. The effect is enhanced by the ball's fuzzy surface. According to Bernoulli's theorem, a statement of energy conservation, the pressure in the accelerated fluid above the ball, compared with the pressure in the retarded fluid below the ball, must drop. The resulting imbalance of forces pushes the ball upward.



LIFT PER UNIT SPAN = $\left(\frac{dL}{dy}\right) = \rho V \Gamma$

VORTEX THEORY OF LIFT was invented independently between 1902 and 1911 by W. M. Kutta of Germany and Nikolai Joukowsky of Russia. In Joukowsky's formulation, depicted here, an inviscid flow around a cylinder of infinite length is "mapped" mathematically into the flow around an airfoil. The flow can be symmetrical, as is shown. Joukowsky added to the inviscid flow an idealized vortex flow coincident with the center of the cylinder. The strength of this "bound" vortex, or circulation, Γ, is chosen so that it gives rise to a stagnation in the flow at the aft singular point x on the cylinder. When this circulation is mapped onto the airfoil, it aligns the flow field with the sharp trailing edge. Under these conditions the circulation creates a lift on both the cylinder and the airfoil. The lift per unit of span is equal to $\rho V \Gamma$, where ρ is the fluid density (in this case the fluid is of course air) and V is the velocity of the remote, or undisturbed, stream. The thickness of the airfoil is controlled by the value that has been assigned to the offset dimension e. The airfoil angle of attack, α, corresponds to the misalignment of the two singular points of transformation, designated x and x', with respect to the remote stream direction.



LIFT PER UNIT SPAN = $\left(\frac{dL}{dy}\right) = \rho W \Gamma$

CONCEPT OF TRAILING VORTEX SHEET was introduced by Ludwig Prandtl to account for the flow around wings where the lift and the bound vorticity must fall to zero at the wing tips. The vortex sheet (*left*) is made up of filaments, or infinitesimal line vortexes, aligned approximately with the direction of flight and increasing in strength toward the wing tips, where they roll up to form tip vortexes. The collective velocity field of the vortex filaments imparts a "downwash," or induced velocity, to the vortex sheet as part of the roll-up process. As is shown at the right, the downwash velocity (w) is added vectorially to the flight velocity in space (V) to produce a local stream velocity (W) that is misaligned with the flight velocity by the induced angle of attack (α). In order to produce a specified lift the geometric angle of attack of the wing section must be larger than the airfoil's effective angle of attack by an amount equal to the induced angle. As a result the lift vector (dL/dy) will be rotated aft by an amount equal to the induced angle, thereby giving rise to an induced drag.

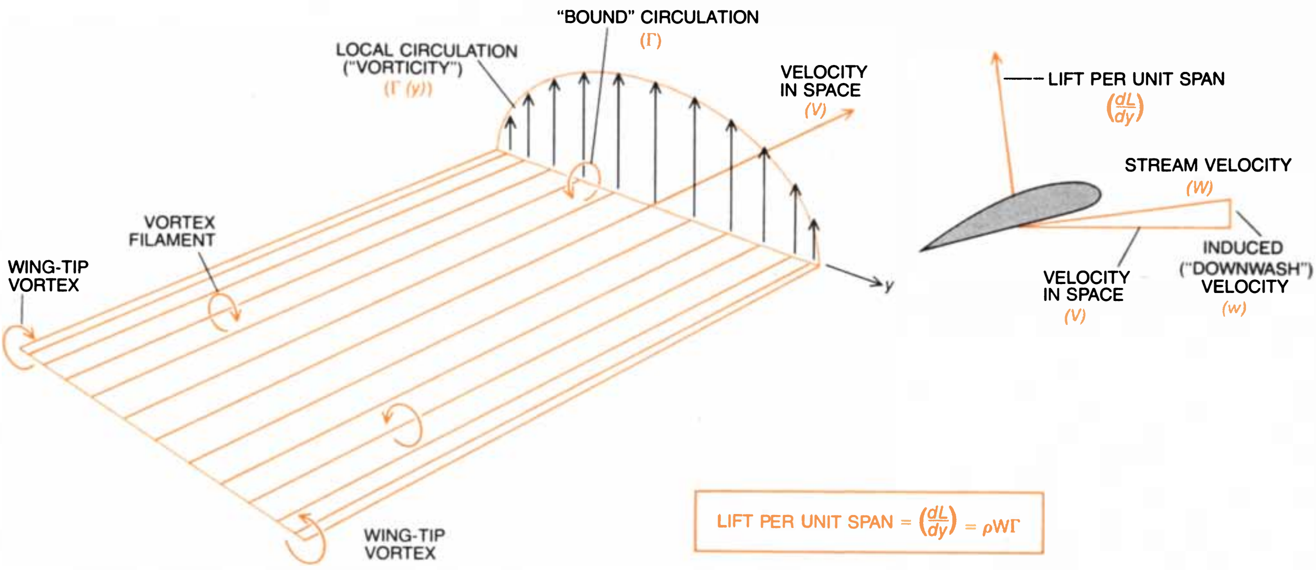
leading edge and reunite at the trailing edge so that the flow over the upper surface of the wing will on the average have a higher velocity and a lower pressure than the flow under it?

The angle of attack is therefore crucial, but there is no satisfactory physical theory to explain why the flow divides precisely the way it does. Joukowsky's adjustment of the vortex strength to produce stagnation at the aft singular point of the transformed cylinder (that is, the airfoil) has the effect of aligning the flow field with the trailing edge of the airfoil, a plausible assumption called the Joukowsky hypothesis or the Kutta-Joukowsky hypothesis. The effect of aligning the flow at the trailing edge is to misalign, or offset, the flow where it divides at the leading edge of the airfoil in such a way that circulation and lift result. The leading edge should be well rounded, as Joukowsky's mathematical airfoils are.

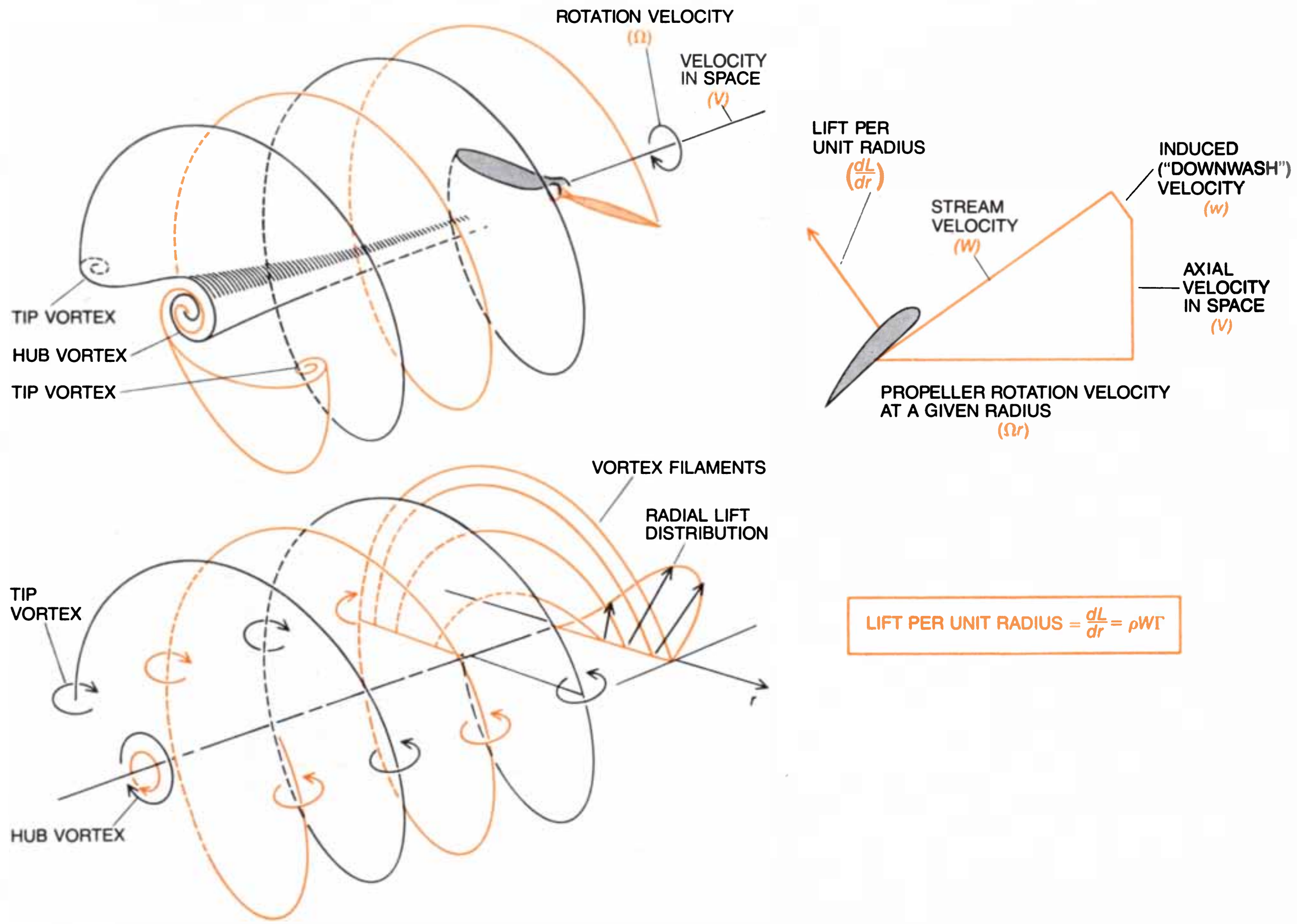
The angle of attack of the airfoil can be varied by tilting the axis connecting the fore and aft singular points of the transformation with respect to the direction of the remote flow field, that is, its direction before it has been influenced by the airfoil. The airfoil camber, or centerline curvature, can be varied by offsetting the radius joining the center of the cylinder to the aft singular point from the axis of the singular points in the transformation. Each of these misalignments involves a change in the strength of the circulation, or bound vortex, and gives rise to corresponding changes in the lift of the airfoil. Wind-tunnel tests of Joukowsky airfoils done by Joukowsky and his associates in Russia and by Prandtl and his associates in Germany just before World War I verified the predictions of the theory for small angles of attack and small amounts of camber. All modern airfoil theories incorporate Joukowsky's results as a special case.

Although Joukowsky's theory explains the lift of wing sections in two-dimensional flow (that is, in the absence of spanwise components in the flow velocity), it presents difficulties when it is applied to the flow around wings of finite span where spanwise flow exists and where lift and the bound vorticity must all go smoothly to zero at the wing tips. This problem was solved by Prandtl, Betz, Max Munk and C. Wieselberger at Göttingen shortly before or during World War I. They introduced the concept of a trailing vortex sheet made up of infinitesimal line vortexes roughly aligned with the direction of flight. The strength of each of the trailing vortexes is proportional to the local spanwise gradient of the bound vorticity associated with the local value of lift of the airfoil section on Joukowsky's model.

The trailing vortexes are particularly



OPTIMUM DISTRIBUTION OF LIFT, corresponding to the minimum induced drag, is achieved when the lift and the "bound" circulation vary elliptically from wing tip to wing tip. With that distribution the velocity field induced by the trailing-vortex system will convect the vortex sheet downward with a uniform downwash velocity across the entire span. In this condition a minimum amount of kinetic energy is transferred to the vortex wake for a given amount of lift. The minimum-loss condition was found by Prandtl's colleague Max Munk.



VORTEX THEORY OF PROPELLERS is like the vortex theory of wings except that the trailing vortex sheets of propellers are helicoidal surfaces (*top*). The sheets roll up into as many tip vortexes as there are blades in the propeller and into one central hub vortex equal in strength to the sum of the tip vortexes. The vectors of induced velocity (w) are perpendicular to the surface of the helicoidal sheets and correspond in magnitude to the slip in the Rankine-Froude momentum theory of marine screws. For minimum induced loss each helical vortex filament (*bottom*) should move perpendicularly to itself in such a way that the entire vortex-sheet array appears to move axially with a constant "displacement" velocity. The vortex-sheet motion for the minimum loss was discovered by Prandtl's colleague Albert Betz.

strong near the edges of the sheet, at the wing tips, where the gradient tends to reach infinity. The trailing vortex sheet rolls up around its intense outer edges, forming two distinct "tip vortexes" (which, however, were not accounted for in the primitive theory). The roll-up process is substantially complete several wingspans behind the trailing edge, where the tip vortexes are spaced less than a wingspan apart. When the humidity of the upper atmosphere falls within a certain range, the water vapor from the exhaust of an aircraft's engines will condense within the cores of the tip vortexes and make the vortexes visible as condensation trails.

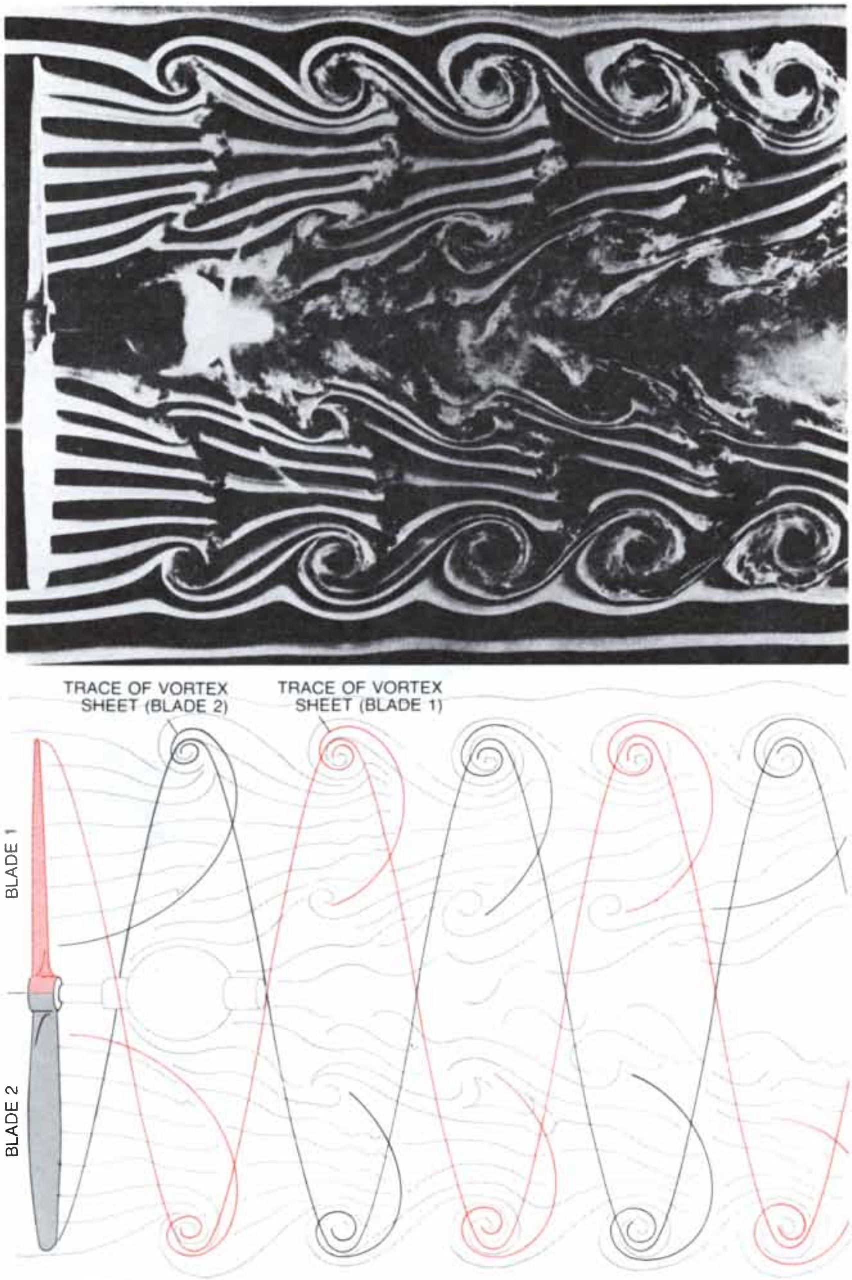
The collective velocity field of the vortexes within the vortex sheet gives rise to a "downwash," or induced velocity, that tends to convect the sheet downward at every stage in the roll-up process. In particular, if the lift of the wing is distributed elliptically in the spanwise direction, the vortex sheet (before it rolls up) is convected downward with a uniform velocity across its entire width. The kinetic energy associated with the trailing vortex flow field is unrecoverable, but the losses are minimized for a given lift, flight speed and wingspan if the lift is elliptically distributed. This important result is attributed to Munk.

Joukowsky's model for airfoil lift in two-dimensional flow can be incorporated into a vortex theory for wings of finite span by calculating the magnitude of the downwash induced by the trailing-vortex system. The vector sum of the downwash velocity and the flight velocity creates an induced angle of attack. The necessary geometric angle of attack with respect to the undisturbed air is obtained by adding the induced angle of attack to the theoretical angle of attack required for a wing of infinite span in two-dimensional (Joukowsky) flow.

The vectors of wing lift also are rotated to the rear through an angle equal to the induced angle of attack, thereby creating an induced drag. Finally, the rate of loss of energy corresponding to the product of the induced drag and the flight velocity exactly balances the kinetic energy continuously added to the vortex system. There is in addition a frictional drag on the wing sections due to the operation of viscosity in the boundary layers adjacent to the wing surfaces, an effect that was also described by Prandtl. The predictions of the vortex theory of wings were confirmed by wind-tunnel experiments at Göttingen during World War I.

From here it is only a short step to a vortex theory of propellers. The essential difference is that the trailing vortex sheets of a propeller are initially helicoidal. Eventually they roll up into as many tip vortexes as there are blades and into one central hub vortex of opposite rotation whose strength is equal to the sum of the strengths of the individual tip vortexes. The collective velocity field of the helicoidal vortex sheets is identical with the "tailrace," or slipstream, of the running propeller. And just as there is a single vortex-sheet motion that minimizes the induced losses of the lifting wing, so there is a single vortex-sheet motion that minimizes the induced losses of a thrusting propeller. This motion was discovered by Betz, and the corresponding (but approximate) radial distribution of bound vorticity for a Joukowsky model of optimum distribution of blade lift was calculated by Prandtl. They published their results in 1919. Goldstein joined the Göttingen group in the 1920's and defined the errors in Prandtl's approximate circulation distribution for his doctoral dissertation in 1929.

It is perhaps fair to say that even aero-



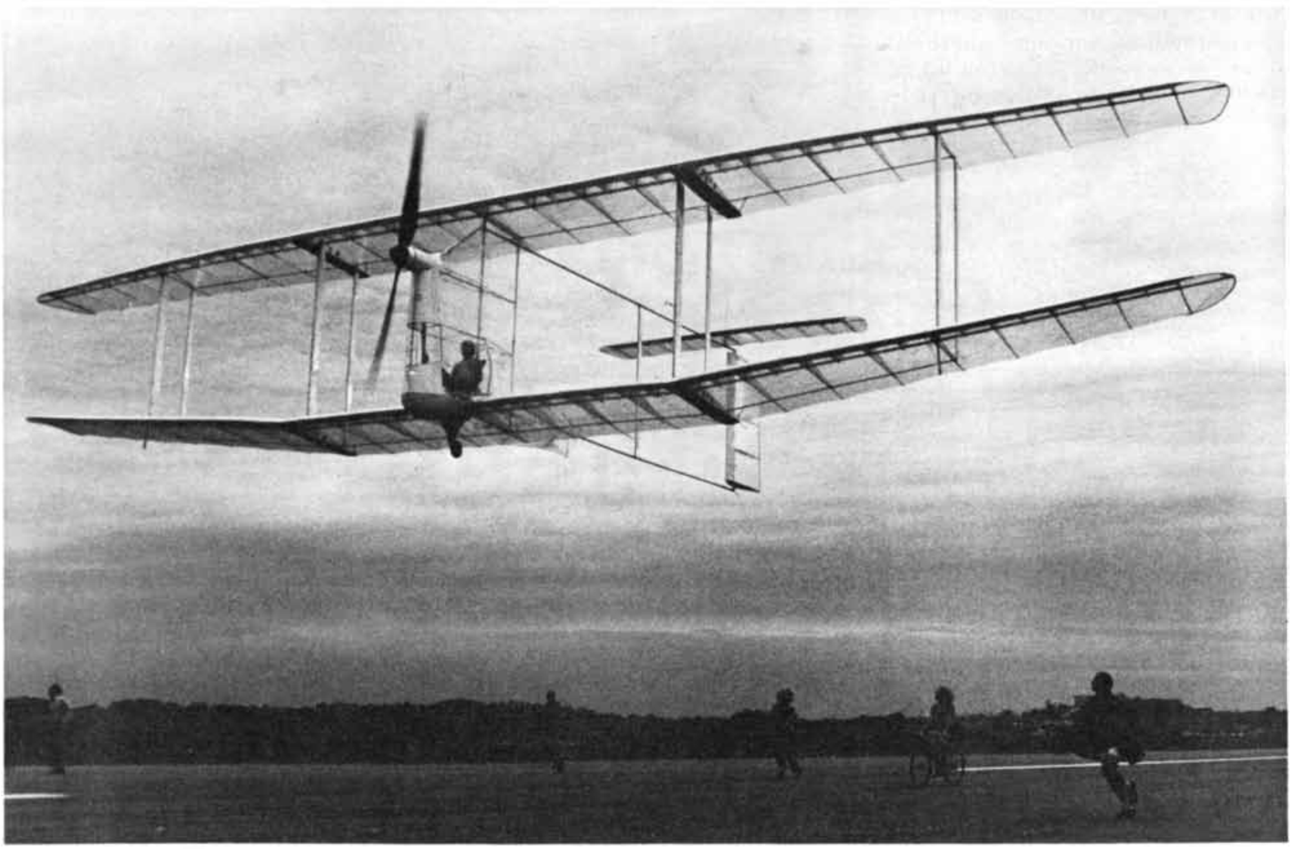
VORTEX WAKE OF TWO-BLADED AIRSCREW, a model-airplane propeller (*top*), was photographed in a special wind tunnel in the late 1950's by F. M. N. Brown of the University of Notre Dame. The photograph shows how smoke filaments in a vertical array are twisted up into a series of tip vortexes uniformly spaced at the pitch of the propeller. The vortex at the hub of the propeller is disrupted by the drive-shaft housing. Diagram at the bottom shows how curved discontinuities correspond to sections through helicoidal vortex sheets shed by blades.

nautical engineers have had difficulty assimilating the Betz-Prandtl-Goldstein contributions precisely because the vortex sheets of a propeller, and their geometric development in space, are hard to visualize in three dimensions. The propeller is, after all, a peculiar kind of wing. Unlike a wing, which in straightaway flight meets the air everywhere at the same average velocity and so has a common angle of attack from root to tip, the propeller meets the air at a different velocity at each radius. That is why a propeller blade is designed with a twist. Near the hub, where the axial, or forward, velocity of the blade section is high compared with the circumferential velocity, the chord line, or tilt, of the blade departs only slightly from the direction of flight. At the blade tip, where the circumferential velocity is high compared with the forward velocity, the chord line may be twisted almost at right angles to the line of flight.

As we have seen, the velocity field of the trailing vortex sheet creates a downwash that is perpendicular to the direction of the airfoil's motion through undisturbed space. Near the hub of the propeller, therefore, the downwash bends backward only slightly from the line of flight. In other words, the vector of induced velocity near the hub has a high rotational component and only a small axial component. Accordingly the blade sections near the hub contribute only a small part to the propeller's total forward thrust. Near the tip of the blade, in contrast, the induced velocity has a high axial component and a much smaller rotational component. As a result the thrust per unit radius for a blade of constant section would increase from hub to tip except near the tip, where it must fall to zero "elliptically."

It might be thought that since the outer part of the blade does more useful work than the inner part, the outer part should be emphasized in the design of the propeller. One can imagine, for example, a propeller whose blades expand in width from hub to tip. Such a geometry (even if it were structurally attainable) would be relatively inefficient. The propeller counterpart of the spanwise uniform downwash velocity for an efficient wing is achieved when the induced velocity at any radius of the propeller blade is equal to half the vortex-sheet velocity at the same radius, and when the vortex-sheet velocities, axial and rotational, vary with the radius in such a way as to present the appearance of a uniform axial "displacement" velocity. A rotating barber pole with helical stripes, for example, has a finite axial displacement velocity but no real physical axial velocity.

In general the propeller geometry of highest efficiency, known as the geometry of minimum induced loss, is that of a propeller whose blades have a maximum chord near 30 or 40 percent of their radius and taper to a pointed tip. Such blades were a common feature of the aluminum-alloy propellers made in the late 1920's and early 1930's. As engines became more powerful and were supercharged to reach higher altitudes it became commonplace to increase the blade area toward the tip in order to absorb the higher power without increasing the propeller radius. Larger diameters were ruled out because tip speeds were already transonic, that is, moving at roughly the speed of sound, a regime for which no satisfactory blade designs were known. In extreme cases the tips were even made square. Such



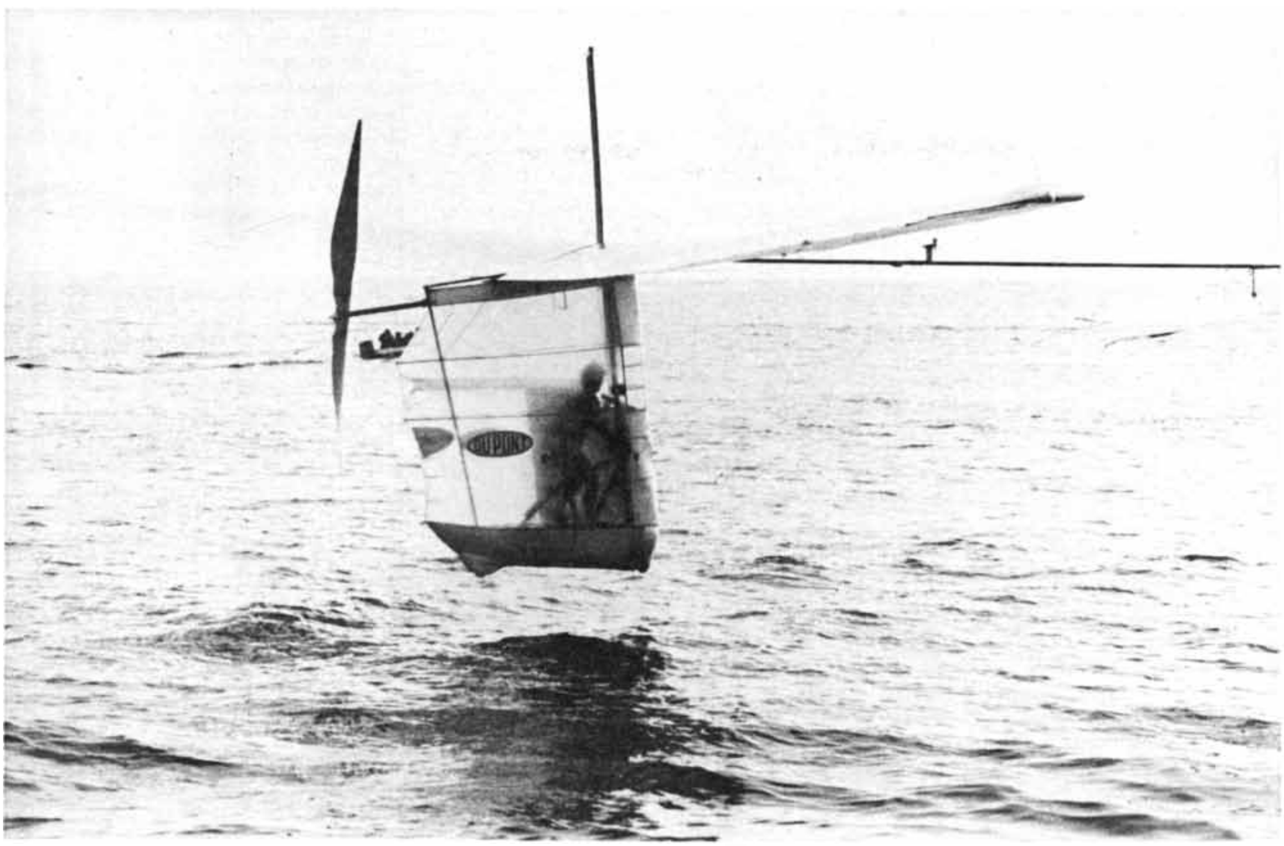
PEDAL-DRIVEN BIPLANE "CHRYSALIS" was built by students at the Massachusetts Institute of Technology between March and June of last year. The biplane configuration was chosen to minimize drag. The wingspan was limited to 22 meters by hangar space. Because of its comparatively short wingspan and its weight of 43 kilograms (30 percent more than that of *Gossamer Albatross*) *Chrysalis* required an input of at least .3 to .4 h.p. at the pedals, depending on the pilot's weight. The *Chrysalis* propeller, like that of *Gossamer Albatross,* was designed by a computer program written by Hyong Bang based on the author's algorithms. Photograph was made by Steven Finberg.

departures from the geometry of minimum induced loss led to small losses in propeller efficiency that were not noticed by the large majority of aeronautical engineers of the day, most of whom probably had not read, or perhaps had not understood, Goldstein's work (or the commentaries on it by Hermann Glauert and later by Theodore Theodorsen). Only when the designer needs every iota of efficiency is minimum-induced-loss design essential.

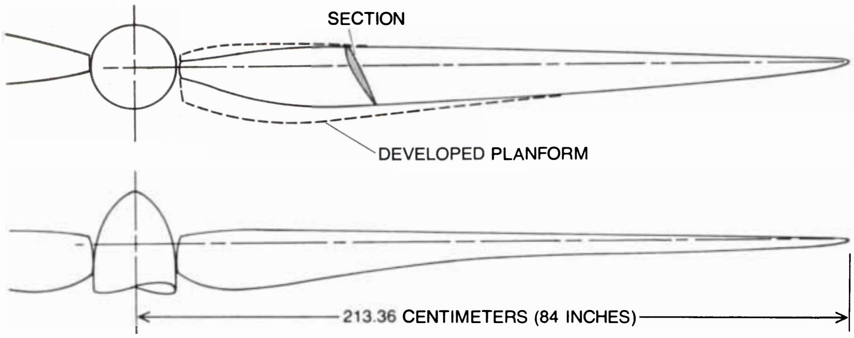
My own contribution to propeller theory has been to develop a "radially graded" momentum theory that lends itself to computer implementation and that is consistent with the Betz-Prandtl form of propeller vortex theory. All propeller theories seek to calculate the induced velocity of the propeller-blade elements that is due to the trailing-vortex system. In the case of propellers with minimum induced loss the half-displacement velocity corresponds to the slip velocity of actuator-disk theory of Rankine and Froude. As in wing theory, the axial and rotational components of the induced velocity (called the inflow velocity in propeller theory) require an increase in blade angle for a specified thrust together with a backward rotation of the blade-lift vectors, thereby increasing the torque needed to turn the propeller. The required increment corresponds to an induced torque. The vortex theory of propellers is hence entirely consistent with the early momentum theories of marine propellers and accounts for the details of blade loading that correspond to a specific propeller geometry.

Propellers, like wings, also have "profile," or friction, losses owing to the action of viscosity in the air flowing near the airfoil surface: the boundary layer. These losses are minimized by operating the blade elements at angles of attack where the ratios of drag to lift are low and, if possible, by choosing helix angles equal to 45 degrees minus half the angle whose tangent is the lift-to-drag ratio of the blade in two dimensions. Reducing the friction losses therefore calls for fairly high "advance ratios" (ratios of forward speed to propeller-tip speed of about .7) and a small number of blades, so that the solidity (or blade-to-disk area) can be concentrated in a few wide blades that will operate at high Reynolds numbers and correspondingly low skin friction. (The Reynolds number is a ratio of the momentum forces to the viscous forces in a fluid and helps to determine the character of the boundary layer and its resistance to separation.) Induced losses, on the other hand, are minimized by increasing the number of blades and lowering the advance ratio in order to decrease the spacing between individual vortex sheets and to make them more nearly perpendicular to the propeller axis. The objective is to make the slipstream velocity field as uniform as possible. The requirements for low induced losses and low friction losses in a propeller are therefore opposed to each other; propellers of the highest efficiency require an optimum balance of the two.

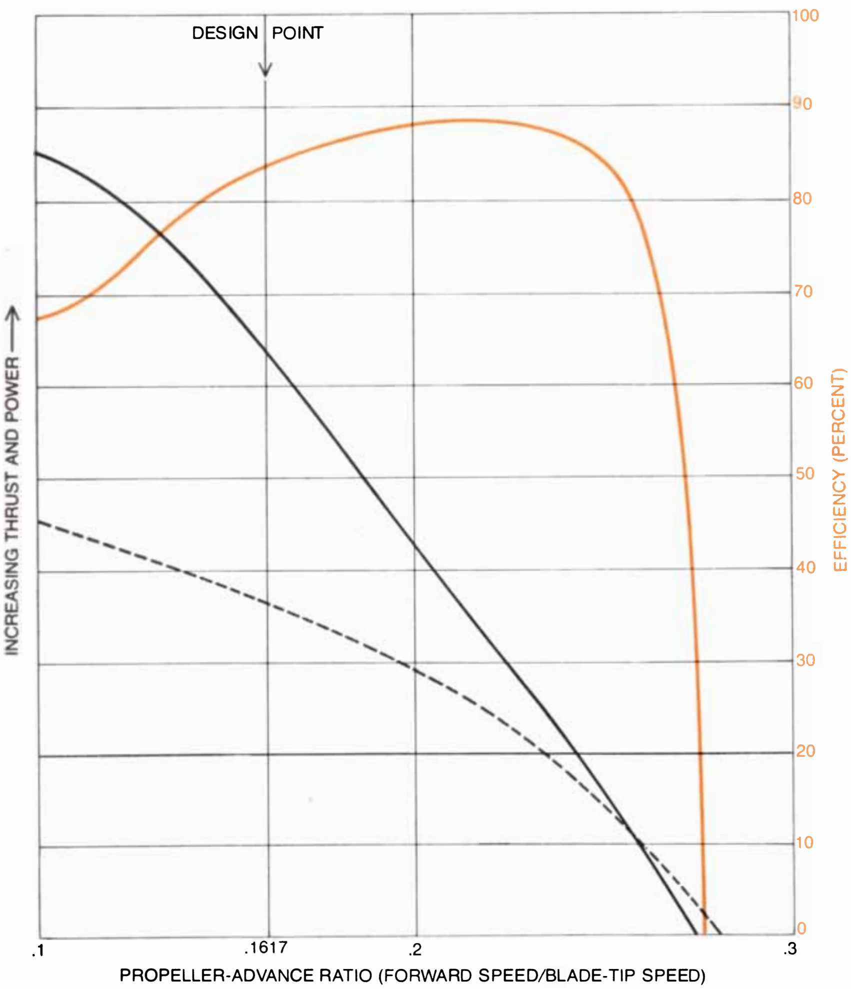
By the use of the approximate analytic Betz-Prandtl circulation distribution I have been able to write down simple algorithms, or computational procedures, that enable a propeller designer to determine the geometry of a propeller of minimum induced loss after accepting whatever design constraints may set a floor on reaching the lowest possible induced loss. (For example, if the design necessitates a two-blade propeller,



"GOSSAMER ALBATROSS" is shown approaching the coast of France last June to claim the prize of £100,000 that had been offered by Henry Kremer, a British industrialist. Allen, the 137-pound pilot, is a 27-year-old professional bicyclist and hang-gliding enthusiast. The flight of 22.26 miles, hampered by head winds, took two hours 49 minutes. The craft was never more than 15 feet above the water. Allen's instruments were an air-speed indicator and an altimeter incorporating the sonar range finder from a Polaroid SX-70 camera. His average pedaling rate was 75 r.p.m., corresponding to a propeller rotation of 95 r.p.m. The photograph was made by Don Monroe.



"CHRYSALIS" PROPELLER is shown from the front (*top*) and the side (*bottom*). The "developed planform," indicated by the broken outline, is the geometry the blade would have if it were untwisted. The maximum chord is at 30 percent radius. The propeller diameter of 4.27 meters is about 4 percent larger than that of *Gossamer Albatross.* Both propellers were designed according to the author's algorithms for minimum induced loss. *Chrysalis* was designed for an input of .5 h.p. (373 watts), a propeller-rotation rate of 135 r.p.m., a flight velocity of 4.88 meters per second (10.9 miles per hour) and therefore an "advance ratio" (forward speed divided by blade-tip speed) of .1617. Because of their high tip speed (6.26 times flight speed) propellers of this type would be unsuitable for aircraft flying more than 50 meters per second (112 m.p.h.).



CALCULATED PERFORMANCE of the *Chrysalis* propeller shows how the coefficients of thrust (*solid black curve*), power (*broken black curve*) and overall efficiency (*color*) vary with the propeller advance ratio. The minimum induced loss for the design disk loading occurs at the design point where the propeller efficiency approaches 83 percent. The efficiency continues to improve at higher advance ratios (equivalent to turning the propeller more slowly at a fixed flight speed) even though the induced losses are no longer minimized; they simply become less as the load on the propeller is reduced and the average axial velocity of the slipstream is decreased. Coefficients of thrust and power are higher when the propeller is turning faster than at the design point (*curves to the left of the design point*) and fall off rapidly with slower turning.

for whatever reason, the designer cannot explore the possible advantages of adding more blades.) These algorithms thus yield a propeller of the highest possible efficiency that will develop a given thrust or absorb a given amount of power for a specified diameter, number of blades, flight speed, shaft speed and air density. The algorithms also account for a wide range of lift-to-drag ratios for blade elements of different contours in order to achieve desired lift coefficients at specified design points. Since the calculations give the propeller efficiency not only at the design point but also under flight conditions away from the design point, they can be used by the designer to change his design point.

It has been my good fortune to take part in a practical demonstration of the validity of what may properly be called a classical theory of propeller aerodynamics, first published in 1919 (a year before I was born) and practically forgotten by 1942, when I began my professional career at the Buffalo, N.Y., plant of the Curtiss-Wright Corporation. I owe this demonstration to enthusiastic students at the Massachusetts Institute of Technology who wanted to build a pedal-driven airplane in January, 1979, and who called on me for propeller-design concepts I had organized mainly for presentation at a symposium.

The students were fired by the dream (remote, to be sure) of winning the prize of £100,000 offered by Henry Kremer, a British industrialist, for a human-powered airplane capable of crossing the English Channel. (An earlier Kremer prize of £50,000 for the first human-powered airplane capable of completing a figure-eight course around two pylons half a mile apart had been won in August, 1977, by *Gossamer Condor,* an ultralight craft designed by Paul MacCready, president of AeroVironment Incorporated of Pasadena, Calif.)

As a first step my students Hyong Bang, Robert Parks and Harold Youngren coded up a version of my propeller algorithms for machine calculation and applied them to the redesign of the propeller for an eighth-scale, free-flying, radio-controlled model of a biplane they planned to name *Chrysalis.* The model, with a wingspan of 2.74 meters, was powered by the motor from a Polaroid SX-70 camera that turned a propeller with an 11-to-one reduction gear. The model was constructed in one week in February. Compared with an intuitively designed propeller, with which the model would barely maintain level flight with fully charged batteries, the theoretically correct propeller, matched to the airframe and power plant, enabled the model to fly three and a half figure eights around pylons 20 meters apart in the M.I.T. gymnasium before the batteries ran down.

The second success for the revived

"'Less is more' seemed for a long time an exclusive pre-occupation of European car-makers...born of the necessities of life here. Suddenly 'less is more' has become the rallying cry of car-makers all over the world—even in America."

Pierre Tiberghien
Director, Automobile Research
and Development/Régie Renault

Pierre Tiberghien has spent all of his 29-year working career in some phase of engineering at Renault, the last five years as head of Renault's Research and Development Center. Despite his lifelong association with things mechanical and automotive, questions about horsepower, CX factors, precious metal shortages, front-wheel drive and inevitably, fuel economy, more often than not elicit answers that have to do with people and their ever-changing needs.

Though it's apparent Tiberghien has thought a lot about cars, it's even more apparent that what really fires him is the car's multi-leveled relationship to people and to the world. His talk is peppered with phrases—"social needs," "polyvalent vehicles," "cars as an invading phenomenon in everyday life"—that testify to the fact that his head and his heart are very close together.

When Tiberghien was graduated from the prestigious Ecole Centrale thirty years ago, European car-makers were responding to the immediate post-war needs of Europe. "Affordability" was the primary consideration; with the evolution of post-war Europe, European needs also evolved. The "two-box" concept is a case in point. Basic car configuration had traditionally been "three-box": motor "box" plus passenger "box" plus luggage "box." In the 60's, Renault engineers, responding to the changes in French and European life, began pioneer work on a "two-box" configuration. The passenger and luggage "boxes" were combined into one adjustable area; this new configuration—soon to be popularly called the "hatchback"—revolutionized basic car design. Here

Portrait by Marie Cosindas

was a car equal to the rigors of the city but, come Friday, ready to serve as a kind of mini station wagon. *Faire plus avec moins*, indeed.

"To do 'more with less' you must innovate not only on the object," says Tiberghien, "but also on the public's perception of an object. The 'something new' that you come up with can be shocking to people's expectations. But only by taking that risk can you succeed. The key is people's needs, not their expectations. Their expectations are based on the past; their needs, on the present and the future."

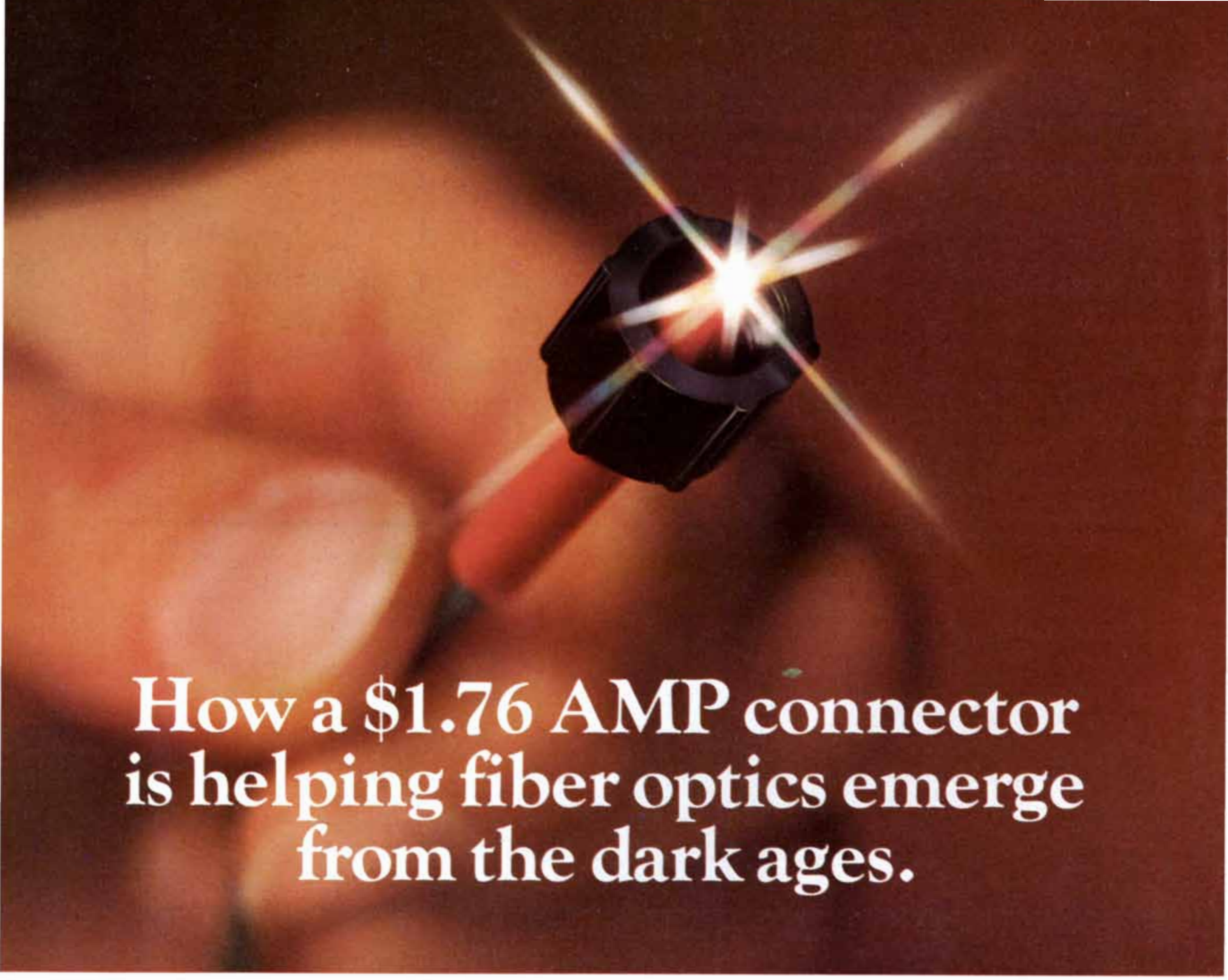
One of seven children, Tiberghien himself is the father of five, ranging in age from 10 to 23 years old. He is also a recent grandfather. Tiberghien and his wife live in a modern Paris apartment with a varying number of their brood. Long week-ends can find them in the small village of Muids in Normandy.

Quite a different holiday took place last summer when all seven of the Tiberghiens motored through the American West. Not only was it a unique vacation, it also enabled Tiberghien to experience American driving at first hand. He came away convinced that Renault cars uniquely answer many present-day U.S. driving needs.

The car is a phenomenon that has shaped the look and manner of the 20th century just as bronze tools and sailing ships shaped other eras. The other great 20th century phenomenon is the apparent shrinkage of the world: some occurrence in what once would have been a distant, exotic place now affects us right in our own backyard garage. Tiberghien is very sensitive to the new role of the car—and what a car should be—in this new "small" world. "We must be conscious of social responsibilities... to make cars that are adapted to man's needs... that include safety factors beyond the legislated required ones... cars that cautiously husband the use of increasingly more precious metals and that are economy-intensive. We reduce and reduce and then come up with something different that, surprisingly, is something 'more.'"

We welcome your comments, questions or thoughts about any of the above. Write to: Renault, U.S.A., Inc., Corporate Group, Englewood Cliffs, N.J. 07632.

RENAULT
the car company of France



Many companies don't call AMP until they need electronic connectors or switches. But why wait till then? If you involve us early enough, we can help you save money. A lot of money.

Soon, tiny glass fibers transmitting light energy signals will be used where once hundreds of wires were needed.

They'll be adopted by many industries. And they'll touch our lives in many ways.

The future of fiber optics is bright. And one of the reasons is because we've already found a better way to connect the fibers.

Our connector precisely aligns each fiber automatically so that it can be polished and connected quickly. It's made a simple few-minute task out of a job that took skilled operators up to an hour.

Already, significant cost savings can be seen. Clearly.

Can we produce the same kind of savings for you?

If your company manufactures electrical/electronic products, the answer is probably yes. But it's important for you to call us early.

That's when our experience in developing connectors, switches and application equipment for hundreds of industries worldwide can help you the most.

Early Involvement

At AMP, we call this approach to solving our customers' problems "Early Involvement." It's our better way. And it's what makes a $1.76* AMP connector worth so much more than $1.76.

Ask for a copy of our brochure "AMP Has a Better Way." Call (717) 564-0100, Ext. 8420.

Or write to AMP Incorporated, Harrisburg, PA 17105.

*Quantity price for two connector halves and splice bushing

AMP

AMP has a better way.

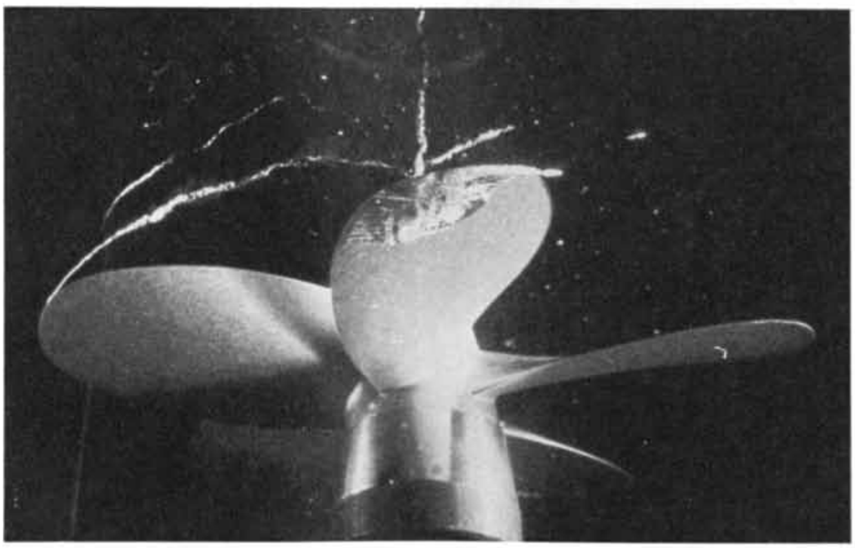
Barcelona • Brussels • Buenos Aires • Frankfurt • Harrisburg • Helsinki • s-Hertogenbosch • London • Luzern • Mexico City • Montreal
Paris • San Juan • Sao Paulo • Stockholm • Sydney • Turin • Toronto • Tokyo

Betz-Prandtl theory came when the same students designed a new propeller for the *Gossamer Albatross* monoplane, which MacCready's group had built for its own attempt at the cross-Channel prize. Bryan Allen, a young professional bicyclist who weighed 137 pounds, had been able to keep *Gossamer Albatross* aloft only 17 minutes when it was equipped with its original propeller. With the new M.I.T. propeller, in combination with improvements in the airframe of the plane, Allen believed he could keep the plane aloft for hours, or until his own "fuel supply" of glucose and water was exhausted. The actual Channel crossing, impeded by head winds, required Allen to pedal for two hours 49 minutes. The distance covered was 22.26 miles; the average ground speed was 7.9 miles per hour.

The third success for the Betz-Prandtl theory came with the design of the propeller for the full-scale *Chrysalis.* Constructed between March and June, the biplane made more than 320 flights during the summer before it was finally dismantled in September. If *Chrysalis* had not had a theoretically correct propeller, matched to its airframe and power plant (.5 h.p.), I doubt that with its short wingspan and substantial weight (10 kilograms more than *Gossamer Albatross*) it could have flown at all.

Compared with the problems of the designer of airplane propellers the problems of the designer of ship propellers are far more constraining. The screw diameter is limited on the one hand by the maximum permissible draft of the laden ship and on the other by its waterline when the ship is lightly laden. The propeller must ingest a heavily nonuniform flow field, distorted by the contours of the stern and full of velocity defects produced by the hull boundary layer. In spite of a favorable multiblade geometry, the sharp limitation on screw diameter leads to excessive disk loading and to large induced losses. The solidity of the propeller, already large because of the heavy disk loading, is further increased by the limits imposed on blade angle of attack and camber by the necessity of limiting cavitation in the flow. Cavitation results when the fluid velocity near the blades is raised so high that the local pressure falls below the vapor pressure of water. In effect the water boils, suddenly creating cavities in the flow field. Cavitation is the principal source of marine-propeller noise.

Further difficult problems are presented by resonant frequencies of excitation that can be set up between the turning blade and the hull of the ship. Marine propellers are usually given curved blades with swept-back leading edges in order to make the excitation less sharp as a function of blade-rotation angle. For all these reasons the analysis of ship propellers must take into account the variation in inflow velocity over the entire blade surface and account for the roll-up of the vortex sheet, which takes place very quickly with heavily loaded propellers. The simple approximate distributions of inflow velocity and circulation of Betz, Prandtl

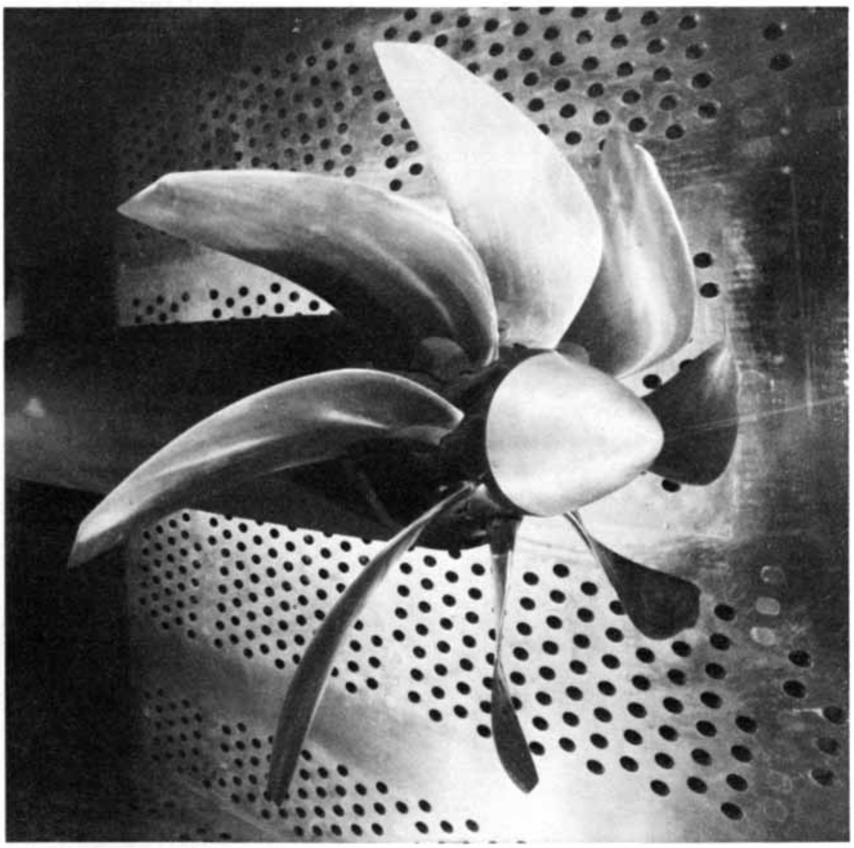


RUNNING MARINE PROPELLERS are studied in M.I.T.'s marine-hydrodynamics laboratory with the aid of high-speed photography by members of the Department of Ocean Engineering. In this photograph, which shows the screw from above, the white region at the tip of the middle blade is a large bubble of water vapor, evidence of abrupt cavitation caused by a nonuniform flow field that simulates the wake of a ship's hull. Low pressures created on the "suction," or front, surface of the blades induce instantaneous boiling of the nearby water. As the blade turns, the large bubble collapses and gives rise to a stream of smaller bubbles. Cyclic cavitation reduces propeller efficiency, causes vibration and generates a characteristic acoustic "signature." Cavitation is reduced by increasing blade area and by optimizing blade geometry.

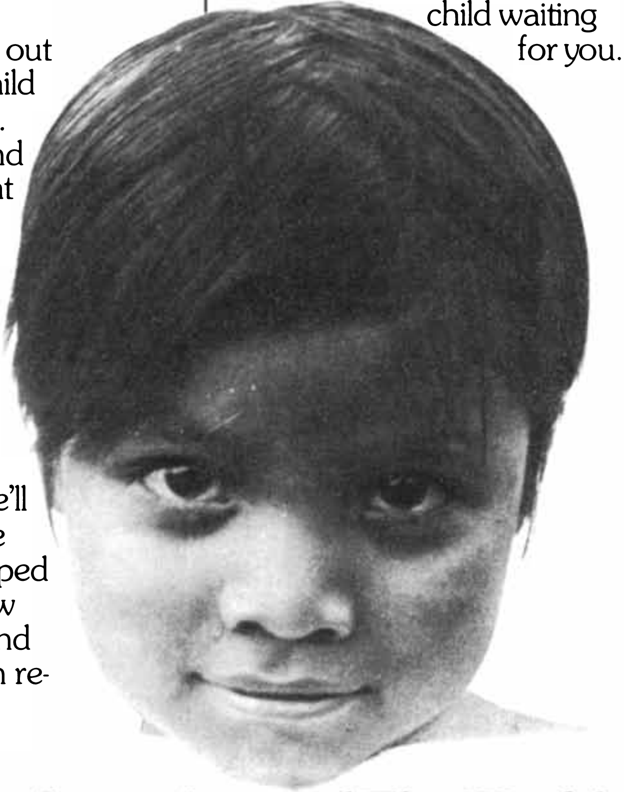
and Goldstein are not applicable to ship screws; the designer must fall back on semiempirical numerical methods of analysis. In the end the efficiency of marine propellers is usually about 70 percent. Naval architects accept this as the best that can be done given the nature of the strong practical constraints.

At present considerable study is being given to propellers for airplanes intended to fly at 80 percent of the speed of sound, which corresponds to about 850 kilometers per hour (about 530 m.p.h.) at cruising altitudes of from 10.7 to 13.7 kilometers (35,000 to 45,000 feet). The blades of such propellers will also have rotational tip speeds of 80 percent of the speed of sound, and when the forward component of velocity is added, the tip speeds will reach 113 percent of the speed of sound. At such speeds propellers will need to be multiblade in order to achieve good lift-to-drag ratios and will have high solidity. Hence they will strongly resemble marine propellers. Efficiencies of 80 percent should be achievable. The high-speed propellers will be noisy in operation (for reasons quite different from those that make marine propellers noisy). The noise, however, will be chiefly "near field" noise, which can be made tolerable by improving the sound insulation of the fuselage. The "far field" noise may actually be less than that of existing turbofan engines. The fuel efficiency of a turboprop engine, given a suitable high-speed propeller, should be about 20 percent higher than the efficiency of a high-bypass turbofan engine with a "gas generator" (the combination of compressor, combustion chamber and power turbine) of the same size as the one that would be needed for the turboprop engine. The propeller of the turboprop engine would be about twice the diameter of the fan in the turbofan engine.

After more than a century of application in hydronautics and aeronautics the screw propeller is very much alive and turning; I expect it will survive for as long as man uses water and air for transportation.



TRANSONIC "PROPFAN" is one of several designs being studied by United Technologies Corporation and the National Aeronautics and Space Administration for potential service in turboprop aircraft capable of matching the speeds of turbofan airliners. If the new craft are successful, they should consume about 20 percent less fuel than present types do. The propfan is termed transonic because the speed of the blade tips will exceed the speed of sound by about 13 percent. The aircraft will fly at 80 percent of the speed of sound (about 530 m.p.h. at 35,000 feet). For operation in the transonic regime an aircraft propeller must approach the high solidity of marine propellers. The solidity of this propeller is 70 percent. Broad blades supply high lift-to-drag ratios needed for transonic operation. The curved blade planforms are chosen to reduce "near field" (cabin interior) noise caused by focusing of the rotating pressure field.



THE AMATEUR SCIENTIST

In judo and aikido application of the physics of forces makes the weak equal to the strong

by Jearl Walker

Judo and aikido are martial arts that demand an intuitive understanding of the physics of forces, torques, stability and rotational motion. This month I shall examine a few of the basic throws of these two forms of combat. Although I cannot convey fully the grace each throw requires, I can break it up into components that can be examined in terms of classical physics. The experiments I shall describe call for actual performance of the throws, but you should do them only under expert supervision. Both types of combat can be dangerous to you and your opponent.

In judo the main goal is to overcome your opponent's stability. The skill lies in the anticipation of his movements and the timing of your responses. The idea is to avoid forcing your opponent into a firm resistance to your throw that would pit your strength against his. A small but skilled judo player has a distinct advantage over a larger but unskilled opponent if a contest of strength is avoided.

Probably the best example of this advantage is the basic hip throw, which is most effective against a taller and slower opponent. In the normal judo competition you face your opponent with your hands grasping the lapels or shoulders of his uniform. To execute the throw you step forward with your right foot to a point between his feet, pulling him downward and toward your right. The throw works well if you have caught your opponent just as he has stepped forward with his right foot. He is still stable against a pull directly toward you, but he is considerably less stable against a pull to your right because of the position of his feet.

During your step forward you curve your body forward so that your head is at your opponent's shoulder level. Next you rapidly turn your left hip backward while pulling him onto your right hip. This should be the first body contact during the movement. If you continue the pull with your hands and the rearward turn with your left hip until you are facing in the same direction as your opponent, he will be rotated over your right hip and onto the mat.

Since you do not want to hurt your opponent in the sport, you maintain your grip on his uniform during the fall so that he lands on his left side and can slap the mat with his left arm during impact. The slap spreads the impact force over a larger area so that the stress on his ribs is not enough to hurt him. Part of the early training in judo involves timing the slap on the mat to coincide with the impact. The only time I have been hurt in the sport is when I failed to slap properly.

Timing and smooth execution are essential to the hip throw, but an understanding of the physics, particularly of the torques and the center of mass, is also necessary. Your opponent's center of mass is the geometric center of his mass distribution. It can be regarded as the point where the gravitational force acts on the body as a whole, which is why it is sometimes called the center of gravity. Your opponent is stable as long as his center of mass remains over the support area outlined by his feet. When he stands upright in a normal posture, his center of mass is approximately between his spine and his navel. Therefore he is stable until you force or trick him into moving his center of mass or into losing part of his support area.

Suppose during a throw you manage to move your opponent's center of mass forward of his feet. Even if you then no longer aid the throw, the gravitational pull on his center of mass creates a torque that might make him fall. To calculate a torque one must multiply two items: the force acting to bring about a rotation and the lever arm between the pivot point and the force. The lever arm is the perpendicular line from the pivot point to a straight line through the force. If you have made your opponent unstable, the force that can rotate him about his feet to the mat is the gravitational pull on him. I represent this force, which is merely his weight, by a vector pointing straight down from his center of mass. Here the lever arm is the horizontal distance between the pivot point at his feet and an extension of a vertical line running through the weight vector. The torque on an unstable opponent is the product of his weight and the lever arm. When your opponent is upright, the lever arm for his weight vector is zero and so the torque is zero too. When he is caught with his center of mass forward of his feet, the lever arm is no longer zero and the resulting torque causes his rotation. The longer the lever arm (the more he is leaning), the greater the torque. One of the objectives of judo is to trick your opponent into an unstable position so quickly that recovery is impossible. Once he is unstable you can continue the throw by applying another torque to him, one that will bring him to the mat long before he can even attempt to regain his stability.

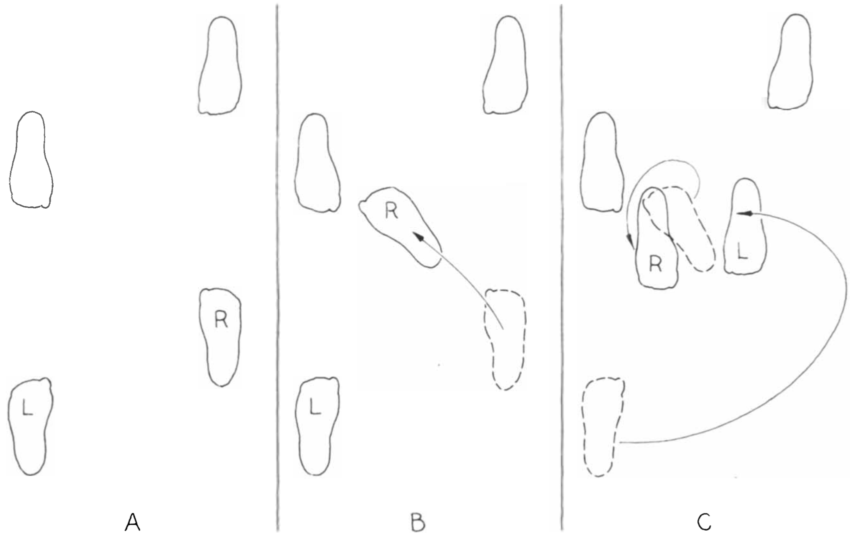
During the hip throw you initially pull on your opponent's uniform to make him unstable. If you pulled directly toward you, you could not easily cause this instability because his center of mass would be moved over his forward foot. He could then maintain his balance by bending that knee. To make him unstable you would have to move his center of mass a relatively great distance until it was beyond his forward foot, but the motion would require a strong and prolonged pull, which he could counter quickly.

An easier way to make your opponent unstable is for you to pull to your right, because in that direction his center of mass must be displaced only a relatively short distance before it no longer lies over the support area. He will probably not be able to counter such a pull before the instability is established. Thereafter he will not be able to counter the continued rotation involved in the hip throw.

Your pull has an additional purpose. It curves your opponent's body so that his center of mass is brought forward to his navel or just outside his body. This new position will aid you in rotating his body over your right hip. Once body contact is made a new pivot point is established at your hip and your pull creates a new torque on the opponent, one that will cause the rotation over your hip. As before the torque is calculated by multiplying the force on the opponent by the lever arm between the line of that force and the pivot point. This time the force is your pull and the pivot point is your hip. Thus the hip throw gives rise to two torques on your opponent, one torque due to his own weight and unstable position and one due directly to the pull you are exerting on him. The throw begins with the first torque so that you can set up the second one without resistance from him.

Suppose you do not curve your opponent's body forward and bring his center of mass out to his navel. Then when you attempt to rotate his body over your

The basic hip throw in judo



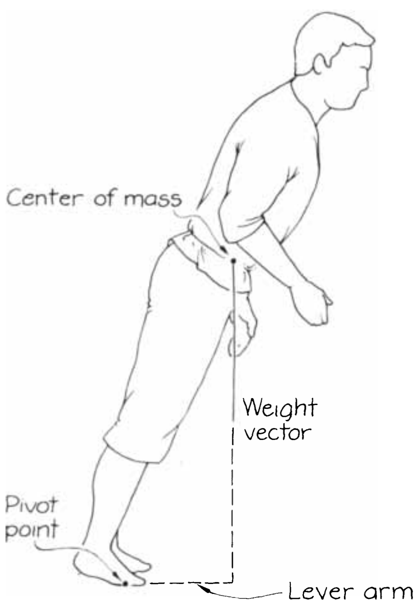
The footwork in the hip throw

right hip, a torque due to his weight will actually counter the torque from your pull on his uniform. Suppose he is still in an upright position when you make body contact and attempt to rotate him over your hip. The pivot point for the rotation is your hip; I shall apply it in determining the lever arms for both of the torques then acting on the opponent.

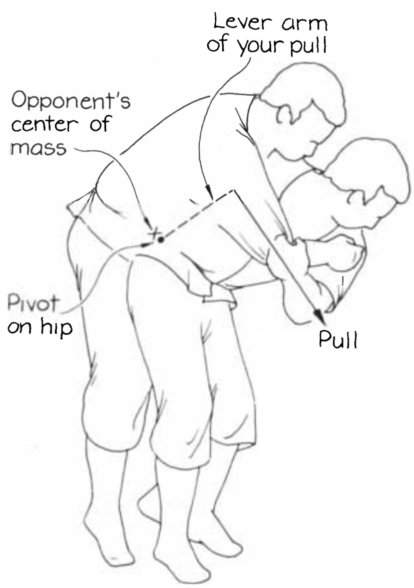
One of the torques is the product of your pull and the lever arm from your hip to the line through the vector of the pull. The other torque is the product of your opponent's weight and the lever arm from your hip to a vertical line passing through the opponent's center of mass. If your opponent is standing upright, the torque from his weight opposes the torque you are applying, since it attempts to rotate him in the opposite direction over your hip. To finish the throw you now must overcome the torque due to his weight, but the time required destroys your advantage in the surprise of the throw. Moreover, you must pit your strength against his.

When the hip throw is properly executed, you bring your opponent's body forward in a curve, move his center of mass out to his navel and so decrease or eliminate the lever arm associated with his weight. The torque due to his weight is therefore diminished and you have a comparatively easy time rotating his body over your right hip. The throw works better on an opponent who is taller than you are because you can pull him downward into the proper curved posture more easily than you could an opponent who is your height or shorter. You can also more easily slide your right hip under a taller opponent. The lever arm of your pull on the uniform of a taller opponent will also be larger, thereby providing more torque to bring him over your hip.

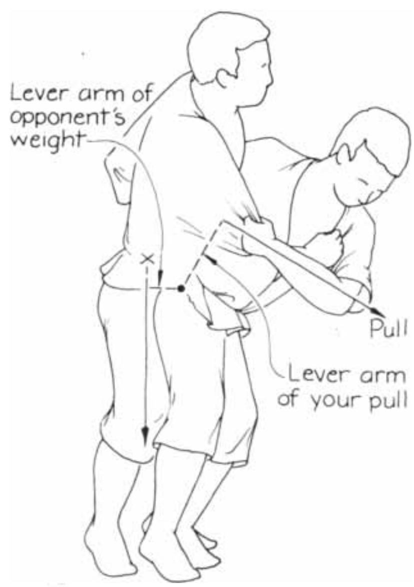
The "major outer reaping throw" (it is called *osotogari* in Japanese) is somewhat easier to understand in terms of the rotational motions. As your opponent steps backward with his left foot you step with your left foot just to the outside of his right foot and pull downward on his uniform to force his weight downward on that foot. Your pull will also be toward his right rear so that his body is curved backward. He is already in



An instability due to the weight vector



The forces in a hip throw



An improper hip throw

Lockheed knows how.

The aircraft you see above—the SR-71 and the L-1011 TriStar—are worlds apart in performance, yet each represents a significant aerodynamic advance.

The SR-71 flies for the Strategic Air Command—higher, faster than any other aircraft. Look at the chines on the front of the SR-71 at the right. Without them, most of the "lift" would be in the rear—and the aircraft would dive as it flew. With the chines, lift also is spread across the fuselage, helping the SR-71 to cruise over 2000 mph. The SR-71 was the first operational aircraft to use chines. Now you find them on advanced fighters.

The Aerodynamic Fuel Saver.

A unique aerodynamic advance on the L-1011 TriStar is helping airlines save fuel. Note those long wings. They've been lengthened nine feet since the L-1011 first flew.

An exclusive system of computer-driven Active Control Ailerons made this possible without costly, time-consuming redesign. Automatically, they work in flight to reduce wing structural loads, decrease drag, produce a smoother ride—and save enormous amounts of fuel over the life of each L-1011.

In the future, other jetliners will have Active Controls, long after they first appeared on the L-1011.

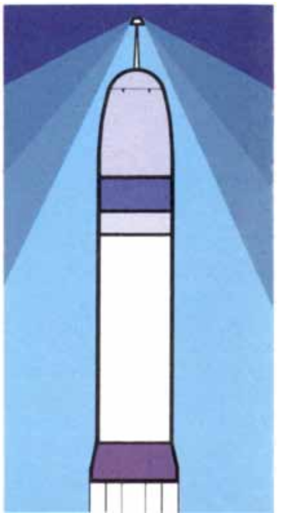
The Aerodynamic Range Stretcher.

The Navy told Lockheed that its newest fleet ballistic missile, Trident, must have twice the range of the earlier Poseidon. But Trident's overall physical dimensions could not be any larger than Poseidon's.

This meant Lockheed had to give Trident a blunt nose. (Poseidon had employed a sharp, aerodynamically efficient nose.) This also meant more aerodynamic drag—and drag works against range.

The solution: fool Mother Nature. Lockheed engineers developed an ingenious telescoping spike that extends more than four feet and locks into place at a precise moment after Trident is launched.

Instead of flying through the atmosphere with the blunt nose slamming into air, a small disc



# of flight.

at the front of the aerospike diverts airflow and creates the aerodynamic illusion that Trident is bullet-shaped. This spike, along with other advances, enabled Lockheed to meet the Navy's requirement of doubled range.

The Airlifter Afterbody Champion.

Airlifters are a special breed. To airdrop large equipment, they need wide rear doors that create a flat, two-dimensional surface. To provide fast, easy loading, their cargo decks must

almost hug the ground—and this means airlifters need a highly upswept fuselage afterbody to let the tail clear the ground on takeoff, particularly short field takeoff. Both the flat rear surface and upswept afterbody create aerodynamic drag that you do not encounter in passenger jetliners.

This caused the first great airlifter, Lockheed's C-130 Hercules, to have 10% of its overall drag in the afterbody. Shrewd linking of structural, mechanical and aerodynamic efforts enabled Lockheed to reduce afterbody drag on the second great airlifter, the C-141 StarLifter, to 3%. In the third great airlifter, the huge C-5 Galaxy, Lockheed engineers chiseled afterbody drag down to an incredible 1%.

The Aerodynamic Lift/Drag Champion.

In aerodynamics, lift is good, drag is bad. The more lift and the less drag, the more efficient the aircraft. And that's where the U-2 reconnaissance aircraft stands out. Those long, slender wings give it the highest lift-to-drag ratio of any operational powered aircraft in the world. But there's even more to the U-2's stellar record. The wings are so successful, in part, because they are extremely light in weight. Manufacturing those feathery wings was as great a triumph as designing them.

The U-2 story is far from over. An advanced version, the TR-1 recently entered production at Lockheed.

What's next? Lockheed's past and present record says new advances are coming, in many forms of flight.

Lockheed

What if you chose HP products to

For example: The HP 45C computer system can color your thinking (and vice-versa) 4913 ways.

Color clarifies. Graphics generated by the new HP 9800 System 45C can clarify your data with up to 4913 shades of color.

Designed for engineers and scientists who have complex design and analytical problems to solve, the System 45C is a completely integrated, powerful workstation with built-in color graphics, CRT display, light pen, operating system, read/write memory, enhanced BASIC language, keyboard, mass storage system, and a high-speed thermal printer.

But a mere list of attributes, impressive as they are, misses the real point: the System 45C can solve complex graphics/computational problems, provide three-dimensional representations in solid or wire frame form, and display results in vivid colors selected to enhance interpretation.

Central to this spectrum of capability is the System 45C's high-resolution CRT screen. Measuring 13 inches (330 mm) diagonally, it can display alphanumerics and vectors in eight colors; 4913 color shades are available for area fill. Colors remain crisp and clear over its entire area. Geometric figures such as circles, rectangles, and regular polygons can be drawn on the screen through simple graphic commands. Another such command quickly adds color fill, selected from standard color models, to any drawn figure. The system's firmware relieves the user from having to write application routines to utilize the graphic capability.

The standard system, including all internal peripherals, the powerful operating system, and 187 kilobytes of user-addressable memory, costs $39,500*. A minimum configuration system is available for $31,500*.

enhance your productivity?

For example: a new approach to making continuous, reliable measurements of a patient's respired CO_2, noninvasively.

Ordinarily, the amount of CO_2 in the blood is regulated automatically by the brain's respiratory center, which controls the rate and depth of respiration. But when a patient is anesthetized, or when disease prevents normal breathing, CO_2 can get seriously out of balance. And it is at just such critical times that the physician most needs to know the patient's CO_2 status, in order to control ventilation effectively.

CO_2 is often measured directly from arterial blood samples, in spite of the fact that significant time may pass between drawing the sample and analyzing it in the laboratory. In such a case, the physician must rely on information that may no longer reflect the state of the patient.

Nevertheless, the analysis of samples from an arterial tap is the favored method, simply because monitoring by other means has been cumbersome and unreliable.

Now, with the introduction of the HP 47210 CO_2 monitor (or capnometer), the physician can get stable, repeatable, continuous CO_2 measurements noninvasively. Microprocessor augmented, and using technology which departs radically from that of other CO_2 monitors, the HP capnometer makes direct measurements of inspired and expired CO_2 by on-line monitoring on a breath-by-breath basis.

The measurement is made optically by a small sensor secured to an airway adapter at the patient's mouth. A beam of infrared light from the sensor is transmitted through the adapter, and the absorption of this light by the breathing gases gives an accurate, continuous measure of the patient's CO_2.

The advantages of the HP capnometer are impressive:

- Simplicity of operation makes it possible to concentrate on the patient rather than the monitor.
- Built-in gas cells make it easy to check and recalibrate the instrument.
- The capnometer monitors itself continuously, and calls attention to itself if a malfunction occurs.
- It also monitors the ventilator or anesthesia machine, and alerts the staff in case of apnea or breathing system disconnections.

An innovative option makes possible the noninvasive *cutaneous* monitoring of CO_2 diffusing through a small area of the patient's skin. Its availability is subject to FDA approval, which is pending.

Price of the HP 47210 is $5800*.

For more information, mail the coupon to Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304. Or call the HP regional office nearest you: East (201) 265-5000, West (213) 970-7500, Midwest (312) 255-9800, South (404) 955-1500, Canada (416) 678-9430.

These are only two of many thousand product reasons for choosing HP.

HEWLETT PACKARD

Mail to: Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304
Please send me further information on

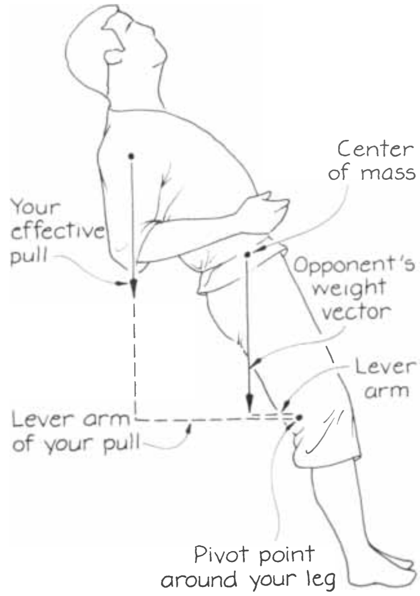
( ) HP 9800 System 45C computer
( ) HP 47210 capnometer

Name ____________________
Company ____________________
Address ____________________
City ____________ State ______ Zip ______

*Domestic U.S. prices only.

00044

The "major outer reaping throw"



The forces in the throw

an unstable position because your pull moves his center of mass to his right rear and away from the support area of his feet. He cannot escape by sliding his feet to the rear and regaining his balance because you have forced him downward. His instability results from the torque his weight creates around a pivot point at his feet, primarily his right foot. The lever arm runs between the pivot point and a line through his weight vector. Placing him in this position sets him up for the next part of the throw, in which you further remove his support area and apply a second torque to bring him rapidly to the mat.

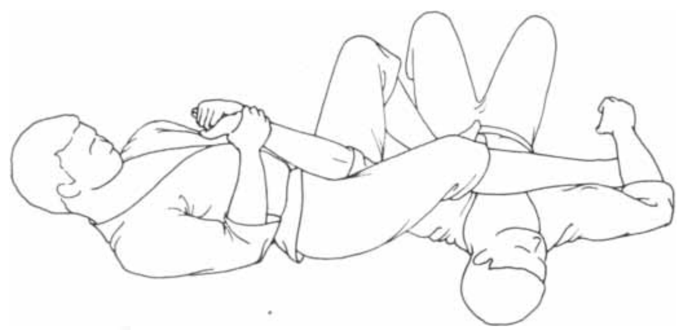
You continue the throw by stepping with your right foot around and behind your opponent's right foot. Then you sweep your right hip and leg to your rear while you force his right side downward and to his rear, leaving him with virtually no support base. The two torques cause him to rotate about a pivot point on your right leg. Even if you did not continue to pull on him after sweeping his leg, he would rotate about your leg because his center of mass is being pulled down by gravity. Your downward pull provides another torque to hasten his fall. In this throw the two torques complement each other.

The "sweeping ankle throw" (*okuriashi bari*) removes your opponent's leg support in a similar manner. As he is about to place his weight on his right foot in the course of a step forward or backward you sweep your left foot into that leg just above the ankle. Simultaneously you pull his uniform in his original direction of travel. Suppose he was moving forward. You pull forward (and so meet no resistance from him) as you sweep his right foot into his left one. Even if he manages to keep his left foot on the mat, his support area is greatly reduced and is swept from under his center of mass. His weight vector through the center of mass provides a torque that will take him to the mat. If you lower your body while maintaining your pull on his uniform, you will provide another torque that will rotate him to the floor. The pivot point is his left foot, and again the two torques complement each other.

Advanced judo classes teach methods of disabling an opponent on the mat. Most of these "hold downs" entail trapping your opponent on the mat with your weight positioned in such a way that he cannot roll over or rise even if he is stronger than you are. For example, in the "cross armlock" (*udehishigi-jujigatame*) you are positioned with part of your weight on the upper torso of your opponent on the mat. He not only is prevented from rising but also probably will not even move for fear of having his arm broken.

The maneuver originates when you are astride your opponent, who is on his back. When he raises his left arm to ward you off, you grasp his wrist with both of your hands, fall to his left side, throw your right leg across his neck and with your left knee raised drive your left ankle into his side. His left arm is pinned between your legs with the elbow downward. Even a gentle downward push on his wrist creates a tremendous torque on his arm around the pivot point where that arm crosses your right leg. He cannot sit up because your weight creates an overwhelming torque on him as he attempts to rotate his trunk about a pivot point at his hips. He also cannot free his left arm even if he is considerably stronger than you are. He could try to counter your torque on his arm by using his shoulder muscles, but they would

An ankle sweep



How to execute a cross armlock

pull on his arm at approximately the location of the pivot point and so their pull would have a short lever. As in most judo techniques, a person trained in creating the correct torques on an opponent has a tremendous advantage even if the opponent is much stronger.

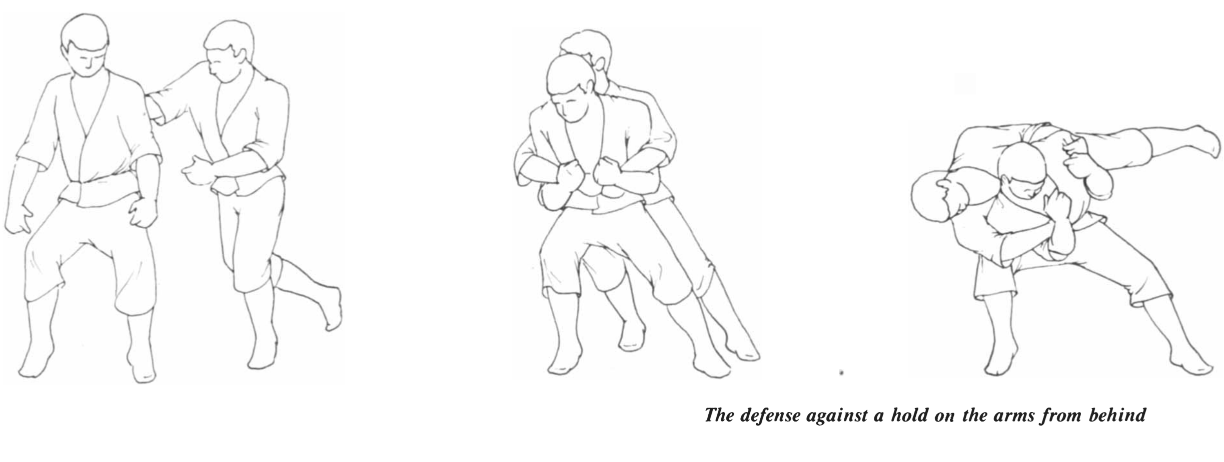
Aikido is a relatively modern form of martial art that incorporates techniques from many of the other martial arts. It is distinguished by its firm code of avoiding injury to the opponent. Hence it is a form of self-defense rather than a sport. It involves no techniques that can be regarded as attacks. I think it is the most difficult of all the martial arts to learn. Its demands for skill, grace and timing rival those of classical ballet.

Aikido employs many of the same principles of physics that are found in judo. Suppose your opponent grasps your wrists from behind. In one of the aikido maneuvers you smoothly lower your body while bringing your wrists upward and over your head toward the front. Your opponent hangs on to your wrists but is brought forward by your descent and slight lean forward. His position is therefore unstable because his center of mass is now slightly forward of his feet. You rapidly draw your right leg backward and drop onto your right knee. Your arms and torso are brought forward and downward in a large arc. Because the first part of the motion induced your opponent to hold tightly to your wrists, he is now thrown over your body in a front somersault.

As in much of aikido, your opponent actually throws himself. He cannot prevent your forward motion because of the unstable posture in which you initially place him. Even if he has superior body weight, he cannot stop the motion by pulling downward on your raised

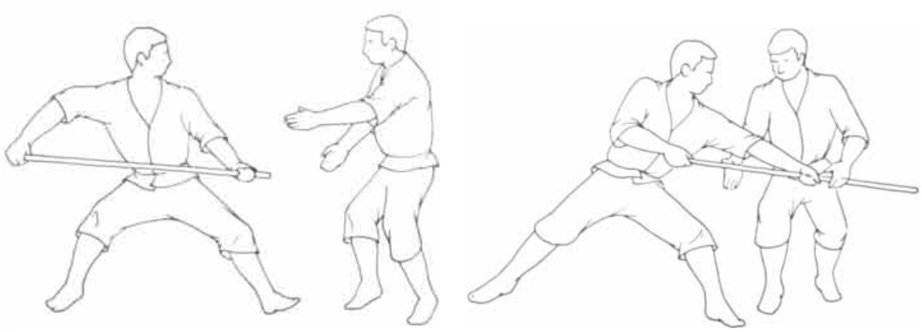
The aikido movement for escaping a hold on the wrists from behind

A defense against a hand slash



The defense against a hold on the arms from behind wrists. In such a position he can pull only along the length of your arms. The torque due to such a pull is zero because there is no lever. Remember, the lever is the perpendicular from the pivot point (in this case your shoulder) to the line through the force. The line through your opponent's pull passes through your shoulder and therefore has no lever arm. Even if he is heavy or strong, he cannot rotate your arms once you have them properly over your head.

Many of the techniques of aikido employ the deflection of a force directed at you. Suppose your attacker throws a



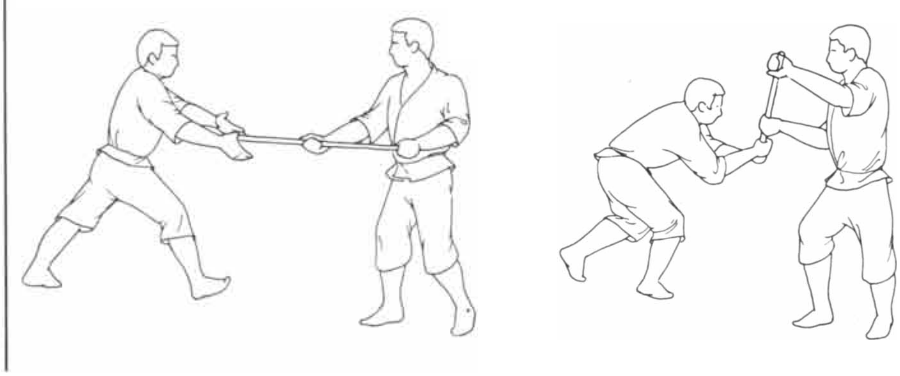
punch at your face. To stop the punch directly requires a large impact force, probably higher than 3,000 newtons. With such forces bones are likely to be broken. A wiser technique is to deflect the strike. Although a large force is necessary to stop the punch directly, only a small force is needed to deflect it. A force of 10 newtons may be enough to deflect a punch by a centimeter.

Although most followers of Western styles of fighting consider an attack to be an advantage, in aikido the attacker is at a distinct disadvantage because of the momentum of his strike. You can use his momentum to throw him to the mat. Suppose your attacker steps forward with his right foot and slashes at your face with the side of his right hand (a typical attack both in Western styles of fighting and in karate). You slide your left foot to the rear as you parry his slash with your left arm. The parry is meant to deflect the slash, not to stop it or even to slow it, since either effect would require strength from you. During the parry you guide your attacker's right arm downward into the grasp of your right hand. While still not fighting the forward momentum of his slash you pull him around in the circular motion you have begun with the withdrawal of your left foot. He was relatively stable against a pull directly forward because of his extended right foot, but he is highly unstable against a pull forward and to his left. In such a direction his center of mass does not have to be moved far before he becomes unstable against a fall. Therefore as you continue to circle you pull him in that direction. He now has two serious disadvantages. First, he is committed to a forward motion that would take a considerable force to stop, even from himself. Second, your pull and his motion are removing his center of mass from his base of support.

To complete the throw rotate your attacker's right arm downward while stepping to your left rear. Turn his wrist upside down and bend his hand around it. At this point it is impossible for him to prevent the throw. He is now off balance and completely unable to stop his own motion. He also cannot pull out of your grip because you have bent his arm at the wrist. Although his arms may be strong, he cannot prevent the torque you create when you push his hand around his own wrist. You bring him to the mat.

How would such a strike to the head be handled in karate? In the Korean karate style of *tae-kwon-do* I was taught to parry a slash with a powerful strike across the opponent's arm. Deflection was important but so was countering the slash with a large force. Force was working against force, and usually the stronger person won. (I was rarely that person.)

Circular motion is employed in aikido both for deflection and to aid in throwing an opponent off balance. Suppose someone approaches you from behind,



How to contend with an attacker who lunges at you with a stick

reaches around your body and pins your arms to your sides. You should reach upward and hold his hands tightly to your chest while sliding your foot forward. The timing is critical because you want to move your torso forward at a rate matching the speed of your opponent. If you delay, you will lose the advantage of exploiting his momentum. If you move too fast, you will have to drag him forward. You must slide your right foot forward at the correct speed and then suddenly lean forward and rotate your body to your right.

The combination of your opponent's initial momentum and your rapid rotation throws him off to the right. He cannot prevent your throw because your lean forward brings his center of mass forward of his feet. He cannot release himself from the forward motion because of his established momentum, because you have pinned his hands and because of his grip around your arms. The centrifugal force on him during your rapid rotation is too large for him to deal with in his unstable position. Hence he essentially throws himself to the floor.

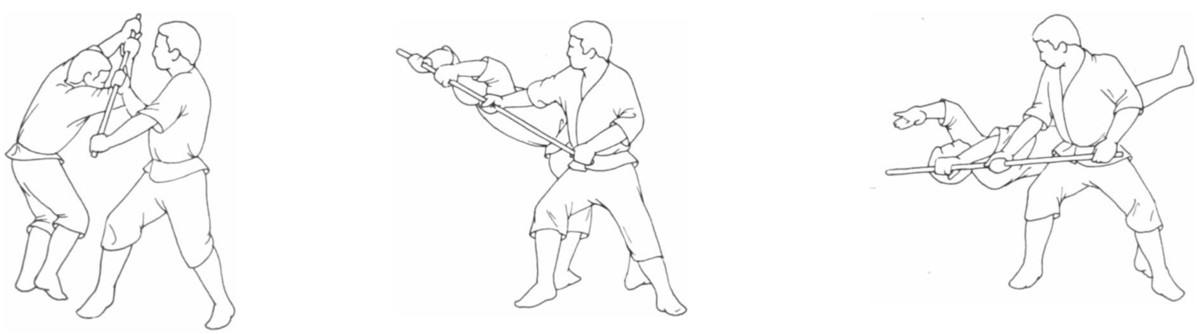
Two more examples of how aikido employs a small force to bring an attacker off balance entail stick fighting as it is taught in advanced classes. Suppose an attacker thrusts a long stick at your midsection, advancing with his left leg during the lunge and thrusting the stick horizontally, holding it with the palms of both hands down. It would be futile to try to stop the end of the stick. You rapidly step forward with your right foot so that the stick passes you on your left. (The agility to do this comes only with long practice.) As the stick passes you turn your body to face it so that you can grab it with both hands. Your left hand is forward of your attacker's outermost hand. Your right hand is between his hands.

In grabbing the stick your objective is not to stop its motion, which would require considerable force. Rather it is to deflect the lunge upward, around to your left in a circular motion and then up and over your attacker's head. Once he has committed himself to the forward lunge he can do little to prevent the deflection. He would need a large force to stop his momentum, and he cannot thrust horizontally to your midsection while pulling downward to prevent your deflection.

Once you have the stick over your attacker's head he is easily thrown. With his left foot forward he is highly unstable against a pull to his left rear because in that direction his center of mass must move only a short distance before it is no longer over his support area. When you have the stick over his head, you pull it downward over his back in that direction. He falls to the mat on his back and probably releases the stick.

Suppose you have a stick and a determined attacker rushes forward to grab its forward end. Allow him to grasp it but lead him with it (as if it were a carrot in front of a donkey) so that he continues his rush. Also lower your end to trick him into bending downward. Once he has committed himself to this awkward motion and is about to pass to your right you bring your end of the stick upward over his face and then downward over his back. If this motion is executed rapidly, he still has a strong grip on the stick and therefore is bent backward by your pull downward over his back and by the continued forward motion of his torso. The torque due to his own weight rotates him to the floor around the pivot point of his feet. His grip on the stick also provides a torque that rotates him. He actually throws himself to the mat because of his initial forward rush and a bit of trickery on your part.

Aikido has hundreds of techniques for employing such trickery against a determined opponent. In nearly all of them a small deflection force parries an opponent's thrust and then guides it so that he throws himself down. When I watch an aikido master defend himself, the motion seems so fluid and effortless that I am inclined to suspect the opponent of faking when he falls to the floor. The fall is not faked. It looks that way because the master has spent years developing an intuitive feeling for the basic physics of forces, rotation and torques.



What to do when your opponent tries to seize your stick

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

MATHEMATICAL GAMES

AN EPISODE OF FLATLAND. C. H. Hinton. Swan Sonnenschein & Co., 1907.

FLATLAND: A ROMANCE OF MANY DIMENSIONS, BY A. SQUARE. E. A. Abbott. Dover Publications, Inc., 1952.

SPHERELAND: A FANTASY ABOUT CURVED SPACES AND AN EXPANDING UNIVERSE. Thomas Y. Crowell, 1965.

A BAN ON THE PRODUCTION OF FISSIONABLE MATERIAL FOR WEAPONS

THE LAST CHANCE: NUCLEAR PROLIFERATION AND ARMS CONTROL. William Epstein. The Free Press, 1976.

NUCLEAR PROLIFERATION: PROSPECTS, PROBLEMS, AND PROPOSALS. *Annals of the American Academy of Political and Social Science,* Vol. 430; March, 1977.

NUCLEAR ENERGY AND NUCLEAR WEAPON PROLIFERATION. Stockholm International Peace Research Institute. Taylor & Francis Ltd, 1979.

INTERNATIONAL NUCLEAR FUEL CYCLE EVALUATION: INFCE SUMMARY VOLUME. International Atomic Energy Agency, 1980.

SYSTEMIC LUPUS ERYTHEMATOSUS

IMMUNOPATHOGENESIS OF SYSTEMIC LUPUS ERYTHEMATOSIS. David Koffler in *Annual Review of Medicine,* Vol. 25, pages 149–164; 1974.

THE IMMUNOLOGIC APPROACH TO SYSTEMIC LUPUS ERYTHEMATOSUS. H. G. Kunkel in *Arthritis and Rheumatism,* Vol. 20, pages S139–S147; 1977.

RHEUMATOID DISEASES. Ken Fye, Haralampos Moutsopoulos and Norman Talal in *Basic and Clinical Immunology,* edited by H. Hugh Fudenberg, D. P. Stites, J. L. Caldwell and J. V. Wells. Lange Medical Publications, 1978.

RECENT PROGRESS IN THE STUDY OF AUTOANTIBODIES TO NUCLEAR ANTIGENS. R. M. Nakamura and E. M. Tan in *Human Pathology,* Vol. 9, No. 1, pages 85–91; January, 1978.

THE IMMUNOLOGY OF RHEUMATOID DISEASES. David Koffler in *CIBA Clinical Symposia,* Vol. 31, No. 4; 1979.

GAMMA-RAY-LINE ASTRONOMY

GAMMA-RAY ASTRONOMY. E. L. Chupp. D. Reidel Publishing Co., 1976.

EVIDENCE FOR STRONG CYCLOTRON EMISSION IN THE HARD X-RAY SPECTRUM OF HERCULES X-1. J. Trümper, W. Pietsch, C. Reppin, W. Voges, R. Staubert and E. Kendziorra in *The Astrophysical Journal Letters,* Vol. 219, No. 3, Part 2, pages L105–L110; February 1, 1978.

GAMMA-RAY LINES: A NEW WINDOW TO THE UNIVERSE. Richard E. Lingenfelter and Reuven Ramaty in *Physics Today,* Vol. 31, No. 3, pages 40–47; March, 1978.

MEGALITHIC MONUMENTS

STONEHENGE. R. J. C. Atkinson. Macmillan and Company, 1956.

THE MEGALITH BUILDERS OF WESTERN EUROPE. Glyn L. Daniel. Praeger Publishers, 1958.

BRITTANY. P. R. Giot. Praeger Publishers, 1960.

ANCIENT EUROPE FROM THE BEGINNINGS OF AGRICULTURE TO CLASSICAL ANTIQUITY. Stuart Piggott. Edinburgh University Press, 1965.

BEFORE CIVILIZATION: THE RADIOCARBON REVOLUTION AND PREHISTORIC EUROPE. Colin Renfrew. Alfred A. Knopf, Inc., 1973.

CONVECTION

HYDRODYNAMIC AND HYDROMAGNETIC STABILITY. S. Chandrasekhar. Oxford University Press, 1961.

THE AMATEUR SCIENTIST. C. L. Stong in *Scientific American,* Vol. 216, No. 1, pages 124–128; January, 1967.

THE AMATEUR SCIENTIST. C. L. Stong in *Scientific American,* Vol. 223, No. 3, pages 221–234; September, 1970.

BUOYANCY EFFECTS IN FLUIDS. J. S. Turner. Cambridge University Press, 1973.

CONVECTIVE INSTABILITY: A PHYSICIST'S APPROACH. Christiane Normand, Yves Pomeau and Manuel G. Velarde in *Reviews of Modern Physics,* Vol. 49, No. 3, pages 581–624; July, 1977.

THE AMATEUR SCIENTIST. Jearl Walker in *Scientific American,* Vol. 237, No. 4, pages 142–150; October, 1977.

THE PHYSIOLOGY OF THE KOALA

THE ECOLOGY OF ARBOREAL FOLIVORES. Edited by G. G. Montgomery. Smithsonian Institution Press, 1978.

THE KOALA: PROCEEDINGS OF THE TARONGA SYMPOSIUM ON KOALA BIOLOGY, MANAGEMENT, AND MEDICINE. Edited by T. J. Bergin. Zoological Parks Board of N.S.W., 1978.

METABOLISM AND HEAT BALANCE IN AN ARBOREAL MARSUPIAL, THE KOALA (*PHASCOLARCTOS CINEREUS*). Robert Degabriele and T. J. Dawson in *Journal of Comparative Physiology,* Vol. 134, pages 293–301; 1979.

SUPERCOILED DNA

SELF-LINKING AND THE GAUSS INTEGRAL IN HIGHER DIMENSIONS. James H. White in *American Journal of Mathematics,* Vol. 91, No. 3, pages 693–728; July, 1969.

THE WRITHING NUMBER OF A SPACE CURVE. F. Brock Fuller in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 68, No. 4, pages 815–819; April, 1971.

THE BEHAVIOR OF THE TOTAL TWIST AND SELF-LINKING NUMBER OF A CLOSED SPACE CURVE UNDER INVERSIONS. Thomas F. Banchoff and James H. White in *Mathematica Scandinavica,* Vol. 36, Fasciculus 2, pages 254–262; 1975.

LINKING NUMBERS AND NUCLEOSOMES. F. H. C. Crick in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 73, No. 8, pages 2639–2643; August, 1976.

STRUCTURE AND REACTIONS OF CLOSED DUPLEX DNA. William R. Bauer in *Annual Review of Biophysics and Bioengineering,* Vol. 7, pages 287–313; 1978.

DNA TOPOISOMERASES: ENZYMES THAT CATALYZE THE CONCERTED BREAKING AND REJOINING OF DNA BACKBONE BONDS. James C. Wang in *Molecular Genetics,* edited by J. Herbert Taylor. Academic Press, 1979.

IS DNA REALLY A DOUBLE HELIX? F. H. C. Crick, J. C. Wang and W. R. Bauer in *Journal of Molecular Biology,* Vol. 129, No. 3, pages 449–457; April 15, 1979.

DNA GYRASE AND THE SUPERCOILING OF DNA. Nicolas R. Cozzarelli in *Science,* Vol. 207, No. 4434, pages 953–960; February 29, 1980.

THE SCREW PROPELLER

AIRPLANE PROPELLERS. Herman Glauert in *Aerodynamic Theory,* edited by William Frederick Durand. Springer-Verlag, 1934–36.

AERODYNAMICS: SELECTED TOPICS IN LIGHT OF THEIR HISTORICAL DEVELOPMENT. Theodore Von Kármán. Cornell University Press, 1954.

MARINE HYDRODYNAMICS. J. N. Newman. The MIT Press, 1977.

THE AMATEUR SCIENTIST

JUDO ON THE GROUND. E. J. Harrison, W. Foulsham & Co., Ltd., 1954.

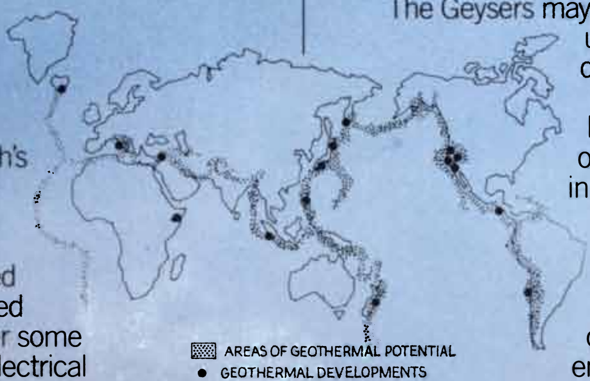
AIKIDO AND THE DYNAMIC SPHERE: AN ILLUSTRATED INTRODUCTION. A. Westbrook and O. Ratti. Charles E. Tuttle Company, 1970.

ADVANCED TECHNIQUES. Morihiro Saito in *Traditional Aikido: Sword, Stick and Body Arts.* Minato Research & Publishing Co., 1973–74.

MOTOROLA HELPS GET ENERGY FROM

In northern California, the natural heat of the earth is being harnessed to serve the energy needs of man. Natural steam, heated by the molten rock below the earth's crust, is being tapped, extracted and piped to power some of the electrical generators that help light up San Francisco, 90 miles to the south.

The place is called The Geysers, and it's part of a pioneering effort by Union Oil Company of California to make geothermal energy a practical alternative to expensive imported oil.



The Geysers may be an unorthodox power source, but it has one thing in common with every other branch of the energy industries: the need to maintain good, reliable communications, no matter how remote or primitive the site.

At The Geysers, as in many places around the world, the solution to that problem is Motorola communications.

ELECTRONIC PROBLEM-SOLVING.

One problem at The Geysers, for instance, was the rolling terrain and steep, narrow canyons, among which conventional high-frequency radio signals could get diffused and lost. Motorola solved that one with ingenious simplicity: a low-band two-way radio system that, as one engineer put it, "gets into the nooks and crannies."

ELECTRONICS EVERYWHERE.

But this is merely the tip of the iceberg of Motorola's experience in energy-industry communications.

One of America's largest oil refineries has 35 Motorola

THE TASTE BEYOND
12-YEAR-OLD PREMIUM SCOTCH
Discover more.
Let your mature taste lead you from the finest premium Scotch on up to the most expensive 12-year-old Scotch in the world. The Glenlivet. The ultimate in Scotch.
Most premium Scotch is blended and depends on several whiskies for taste and smoothness.
Yet the unblended character of The Glenlivet is noble enough to stand alone. In this distinctive Scotch whisky you'll experience an exceptional smoothness and full-bodied richness, unmatched in all other premium Scotch.
Try the taste beyond premium Scotch.
The Glenlivet
Unchanged since 1824.
The GLENLIVET
AGED 12 YEARS
The GLENLIVET
12 YEARS OLD
Unblended all malt Scotch Whisky
Distilled and Bottled in Scotland
GEORGE & J. G. SMITH LIMITED
U.S. IMPORTERS
THE GLENLIVET DISTILLING COMPANY
NEW YORK, N.Y.