SCIENTIFIC AMERICAN

DEADLY MUSHROOMS

$1.25

March 1975

Saving the Main Frame.

The COPE 1600 Remote Communications Processor

Some central computers get bogged down by nickel and dime jobs from distant plants and offices. With Harris "intelligent" terminals and controllers, remote locations can handle routine processing on the spot, saving main frame time for main frame work.

Write: Harris Corporation,
55 Public Square, Cleveland, Ohio 44113.

HARRIS

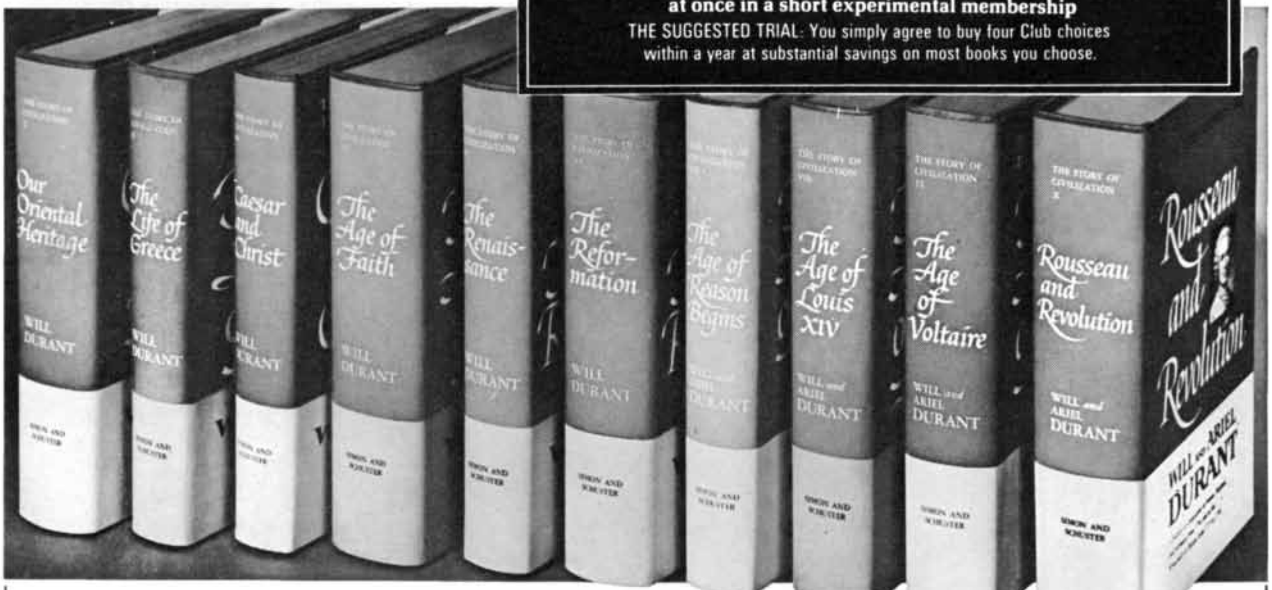
COMMUNICATIONS AND
INFORMATION HANDLING

As your introduction to membership in the BOOK-OF-THE-MONTH CLUB®

THIS GREAT $175 TEN-VOLUME HISTORY FOR ONLY $1.75 A VOLUME

at once in a short experimental membership

THE SUGGESTED TRIAL: You simply agree to buy four Club choices within a year at substantial savings on most books you choose.



Complete after forty years' research and writing

The Story of Civilization

by WILL and ARIEL DURANT

I. Our Oriental Heritage. Ancient Egypt and the Near East. The Far East, early to modern times.

II. The Life of Greece. Explores all facets of Greek life from prehistoric times to the Roman conquest.

III. Caesar and Christ. The rise of Rome and Christianity and the collapse of classic civilization.

IV. The Age of Faith. Christian, Islamic and Judaic civilizations, 325 to 1300, including the Crusades.

V. The Renaissance. Italy's golden age, 1304 to 1576. A turbulent world of intrigue and great art.

VI. The Reformation. Europe's religious conflicts, from two centuries before Luther to Calvin.

VII. The Age of Reason Begins. Europe, 1558-1646. The age of Shakespeare, Rembrandt, Galileo.

VIII. The Age of Louis XIV. The brilliant era of the "Sun King," Milton, Cromwell, Peter the Great.

IX. The Age of Voltaire. Europe from 1715 to 1756. The world of Frederick the Great, Wesley, Bach.

X. Rousseau and Revolution. Europe from the Seven Years' War to the storming of the Bastille.

EACH OF THE TEN VOLUMES IN THE SET LISTS FOR $17.50

THE TRIAL suggested here will prove, *by your own actual experience,* how effectually membership in the Book-of-the-Month Club can **keep you from missing,** through oversight, books you fully intend to read.

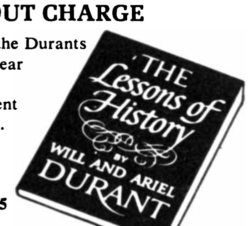
As long as you remain a member, you will be sent approximately every three and a half weeks the Book-of-the-Month Club *News,* a literary magazine published by the Club fifteen times a year, which describes the Selection as well as scores of other books, most of which are available at substantial discounts — up to **40%** on more expensive volumes. All these books **are identical to the publishers' editions in content, format, size and quality.** If you wish to purchase the Selection, do nothing and it will be shipped to you automatically. However, there is no obligation to purchase any particular volume. If you do not want the Selection — or you would like one of the Alternates or no book at all — simply indicate your decision on the reply form always enclosed with the *News* and mail it so we receive it by the date specified. If, because of late mail delivery of the *News,* you should receive a Selection without having had 10 days to decide whether you want it, that Selection may be returned at Club expense.

BOOK-DIVIDENDS: If you continue after your trial membership, you will earn at least one Book-Dividend® Credit for every selection or Alternate you buy. These Credits entitle you to obtain a wide variety of books, called Book-Dividends, at astonishing savings—**at least 70% of publishers' list prices.** Under this unique system, more than 100 Book-Dividends will be available to choose from every year, enabling you to acquire a well-rounded library at a small fraction of what you would otherwise have to pay.

INCLUDED WITHOUT CHARGE

A brilliant summary of all the Durants learned during their forty-year study of world history . . . and an invaluable supplement to *The Story of Civilization.*

A copy will be included, *free,* with each set sent to new members who enroll at this time. **Pub price $6.95**



BOOK-OF-THE-MONTH CLUB, INC.
Camp Hill, Pa. 17012 **5-A170-3**

Please enroll me as a member of the Book-of-the-Month Club and send me *The Story of Civilization* in 10 volumes, billing me $1.75 per volume. I agree to purchase at least four Selections or Alternates during the first year I am a member, paying in most cases special members' prices. My membership is cancelable any time after I buy these four books. A shipping charge is added to all shipments.

56

MR.
MRS. ....
MISS (Please print plainly)

Address.............................*Apt*.........

City..

State...........................*Zip*............

Dick Salisbury is down in the dumps...

He's a guy who spends a lot of time in the dumps, but he has good reason to be happy while he's there.

Dick Salisbury deals in steel cans, which we all know as tin cans. He is Vice President and Chief Operating Officer of Los Angeles By-Products Co., a firm that combs the garbage in major California cities for a half billion steel cans a year. In San Francisco, for example, his magnetic separators recover steel cans from garbage at the rate of 10,000 tons per year.

L.A. By-Products also operates steel can recovery facilities in Oakland, Sacramento and Contra Costa Counties. Since steel cans are magnetic, it is easy to pull them out of municipal garbage by the millions.

These reclaimed cans are sold to nearby copper mines for use in a chemical process that recovers copper from low grade ore.

Steel cans are magnetically recovered in more than 20 cities across the nation. In addition to their use in copper mining, they are remelted to make new steel, used in the production of ferro-alloys, or detinned. As a result of recent research, they will soon be used by the foundry industry as well.

Recovery and recycling of steel cans does more than simply preserve our natural resources. It saves valuable landfill space. It also conserves energy. And it puts cash in municipal treasuries.

Sure, Dick Salisbury and his men are down in the dumps, but they can manage a smile.

If you are down in the dumps about garbage, maybe we can help cheer you up, too.

For more information write, **Tinplate Producers, American Iron and Steel Institute, 1000 16th Street, N.W. Washington, D.C. 20036.**

but he's never been happier.

Tinplate Producers, American Iron and Steel Institute

Established 1845 SCIENTIFIC AMERICAN March 1975 Volume 232 Number 3

ARTICLES

DEPARTMENTS

THE COVER

The painting on the cover shows four mushrooms of the genus *Amanita,* all of them poisonous. They are, from left to right, the red- or orange-capped *A. muscaria,* the greenish *A. phalloides,* the pure white *A. virosa* and the brown *A. brunnescens.* Most deaths from mushroom poisoning are caused by *A. phalloides* or by *A. virosa* and a few other white amanitas that closely resemble it (see "The Most Poisonous Mushrooms," page 90). The other two species shown are not ordinarily lethal. The stalk and gills of the amanitas are white, and the deadly species are distinguished by a distinct, membranous "death cup" at the base of the stalk. Other identifying marks are more difficult to discern, such as the faint brown stains that develop on the stalk of *A. brunnescens* after it has been bruised by handling.

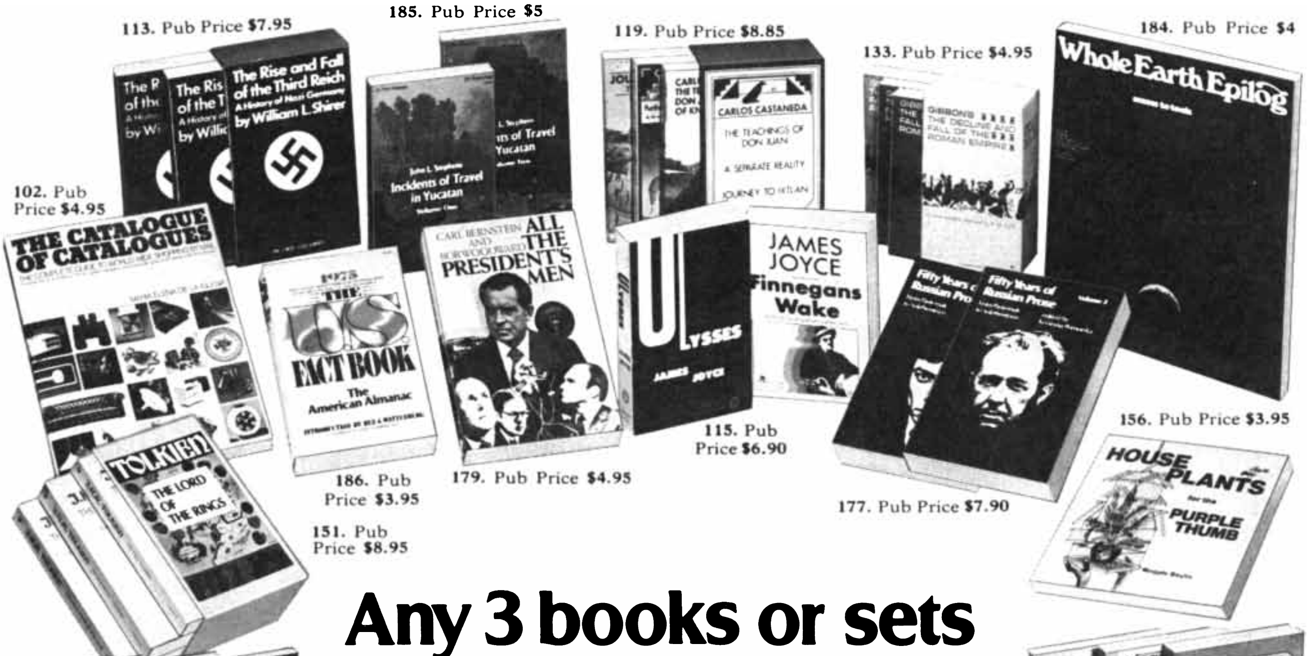
THE ILLUSTRATIONS

Cover painting by Bunji Tagawa

Page	Source
18–20	Laszlo Kubinyi
21	Atlas Copco Mining and Construction Technique, Inc.
22	John Eklöf
23	Ilil Arbel
25–35	Dan Todd
37	Laszlo Kubinyi
38	Ilil Arbel
39	Alphonse Chapanis, Johns Hopkins University
40–42	Ilil Arbel
51–55	Allen Beechel
56	Harry W. Green II, University of California at Davis
57	J. B. Dawson, University of St. Andrews
60–63	Allen Beechel
65–72	Alan D. Iselin
76	Fritz Goro
78	Andrew A. Benson, Scripps Institution of Oceanography (*top*); Lorelle M. Raboni (*bottom*)
79–83	Lorelle M. Raboni
86	Fritz Goro
90	Donald M. Simons
92–95	Bunji Tagawa
96	Anneliese M. Lengsfeld, Max Planck Institute for Medical Research
97	E. Weiss, Institute for Veterinary Pathology
98–100	Bunji Tagawa
103	Ben Rose
104	Ilil Arbel
105	National Central Library, Florence
106	Ilil Arbel
107	National Central Library, Florence (*top*); Ilil Arbel (*bottom*)
108	National Central Library, Florence
109	Ilil Arbel
112–116	Ilil Arbel
118–122	Jerome Kuhl

QUALITY PAPERBACK BOOK SERVICE

The common sense book club

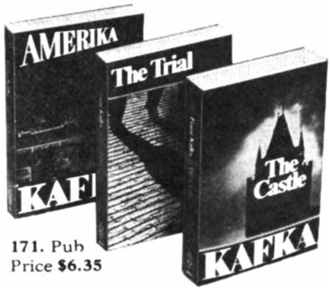


132. CAT'S CRADLE, SLAUGHTER-HOUSE-FIVE *and* BREAKFAST OF CHAMPIONS *by* KURT VONNEGUT, JR. (3 Vols.) Pub prices total $7.95

Any 3 books or sets for only $1 each

You simply agree to buy six more within the next year and as a member you'll immediately qualify for our Free Book Plan

Today, more than ever, fine paperbacks are the best values around in books, and the best way to buy them is by belonging to QPB. As a member, you will not only be kept informed about all the most exciting new books and the best of those already in print, but you will save 20% on all Main Selections and will immediately qualify for our Free Book Plan. It makes good sense to let the Quality Paperback Book Service help you find the best paperbacks to read and to own. There's no other book service like it.



HOW MEMBERSHIP WORKS

1. You will receive the *QPB Review* fifteen times a year—about every 3½ weeks. This informative catalog describes the Main Selection plus at least fifty Alternates. **The Main Selection is always offered at a 20% discount.**

2. If you want the Main Selection, do nothing. It will be shipped to you automatically. If you want one or more of the Alternates—or no book at all—just indicate your decision on the reply form always enclosed with the *Review* and return it by the date specified.

3. Free books. For every book or set you buy (exclusive of your three introductory choices for $1 each) you earn at least one Bonus Point, and the Bonus Points you accumulate entitle you to free books and sets. (You pay only shipping charges.)

4. Return privilege. If your *Review* is delayed in the mail and therefore you receive the Main Selection without having had ten days to notify us if you did not want it, you may return that Selection at our expense.

qpb
QUALITY PAPERBACK BOOK SERVICE, INC.
Middletown, Pennsylvania 17057

QUALITY PAPERBACK BOOK SERVICE, INC.
Middletown, Pennsylvania 17057 5-QB170-3

Please enroll me as a member of the Quality Paperback Book Service and send me the three books or sets whose numbers I've indicated in the boxes below. Bill me only $3 for all three, plus shipping. My only obligation is to purchase six more books or sets during the coming year, receiving a 20% discount on every Main Selection I take. My membership is cancelable any time after I buy the six additional books or sets. A moderate shipping charge is added to all shipments.

INDICATE BY NUMBER YOUR THREE BOOKS OR SETS ☐ ☐ ☐

Mr. / Ms. ______________________ 5
(Please print plainly)

Address ______________________ Apt. ______

City & State ______________________ Zip ______

O-20

LETTERS

Sirs:

The commentary entitled "Crack! Rumble! Boom!" in "Science and the Citizen" [SCIENTIFIC AMERICAN, January] asks and answers the question, "What makes thunder roll?" The thunder "signature" is of course related to the lightning signature: the irregular shape of the lightning channel. Professor Arthur A. Few, Jr., of Rice University, you point out, has ingeniously used arrays of microphones to record the thunder in such a fashion that he can reconstruct the lightning configuration. This work marks a major advance in the understanding of the acoustics of thunder.

By sheer coincidence an inverse effort was initiated in our laboratory several years ago. The objective was to reconstruct the thunder heard at a certain location, given the shape of the lightning stroke. The end result, which was attained in the summer of 1973, was a loudspeaker reproduction of the rumble and roll of synthetic thunder generated by a computer. Starting with a digitized photograph of a lightning stroke, the computer models the physics of the sound-generation process in terms of a sequence of overlapping "sonic booms." The result of the calculation is a time history of the sound pressure at the observer location. These digital data are converted into an electrical signal that varies with time. The varying signal is fed into an amplifier and loudspeaker to produce the sound of thunder.

Although not a perfect reconstruction (for one thing, the photograph is only a two-dimensional projection of the lightning stroke), this synthetic thunder was sufficiently realistic to fool an unsuspecting colleague: he called out in dismay, "Is it starting to rain?" During the Centennial Open House of the Faculty of Applied Science and Engineering at the University of Toronto on October 18–20, 1973, the thunder was demonstrated in a push-button-operated booth display by means of a tape recording. Outside Toronto the synthetic-thunder tape has been informally demonstrated at the AF/DOT Second Interagency Symposium on Transportation Noise, North Carolina State University, Raleigh, June 5–7, 1974, and further demonstrations at technical meetings are planned. Coworkers on this effort have been F. Lam, K. A. Leung, D. Kurtz and N. D. Ellis.

H. S. RIBNER

Institute for Aerospace Studies
University of Toronto
Downsview, Ont.

Sirs:

I should like to comment briefly on "The Fuel Consumption of Automobiles" in the January issue by John R. Pierce.

Professor Pierce wrote an excellent article. It should be read and studied by those who speak of improving fuel economy without a realistic recognition that trade-offs are involved.

Professor Pierce makes the valid point that there are a number of methods by which the fuel consumption of a vehicle can be improved. Some of the means available, such as the doubling of tire inflation in order to reduce rolling resistance, could receive a very adverse market reaction, since higher tire pressure seriously degrades a vehicle's riding characteristics.

However, Professor Pierce did not mention—most probably because it cannot be built into the "design" of a car—the very immediate benefit of routine car maintenance and prudent driving habits. Tests made on the General Motors city-suburban driving schedule, for example, show that fuel economy can vary by as much as two to one, depending on driving habits and whether the car is maintained to factory specifications. These are under the control of the driver who is driving a car today and do not have to wait for 1980.

Professor Pierce made one very important point that is often overlooked by those discussing fuel economy. After summarizing the advantages and disadvantages of diesel, stratified charge and hydrogen admixture to achieve higher efficiency through use of a leaner air-fuel mixture, he said: "As we have seen, they also reduce emissions of hydrocarbons and carbon monoxide in the exhaust, perhaps to the point of meeting present emission standards. Proposed standards on emission of oxides of nitrogen, however, are difficult to meet. The production of oxides of nitrogen can be reduced only if the mixture is made lean enough to lower the temperature of combustion substantially. It may be that the proposed standards on emission of oxides of nitrogen are unrealistically stringent."

We agree, and we have urged that Congress consider continuing the current oxides-of-nitrogen standards for the foreseeable future. Apparently the Administration agrees, since President Ford has called for a five-year extension of current emission standards in order to allow the automobile companies time to improve fuel economy.

ANTHONY DE LORENZO

Vice-President
General Motors Corporation
Detroit

Sirs:

A recent TRW advertisement in your magazine showed an artist's rendering of a black-hole accretion-disk model of Cygnus X-1. We erroneously attributed the calculations underlying this model to Kip Thorne of the California Institute of Technology. Professor Thorne informs us that this attribution is incorrect. The key calculations were carried out by research groups in Moscow and in Cambridge, not by him. We apologize for this error.

KEN MORITZ

Special Projects Manager
TRW Systems Group
Redondo Beach, Calif.

Scientific American, March, 1975; Vol. 232, No. 3. Published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017; Gerard Piel, president; Dennis Flanagan, vice-president; Donald H. Miller, Jr., vice-president and secretary; George S. Conn, treasurer; Arlene Wright, assistant treasurer.

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 Harry T. Morris, 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 30 cents.

Change of address (or other subscription correspondence) should be addressed to Circulation Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Please notify us four weeks in advance of change. If possible, kindly furnish an address imprint from a recent issue. Send both old and new addresses or complete and mail Post Office Form 3578.

☐ Please enter change of address.

☐ Please enter subscription.
U.S. and Canada:
☐ 1 year, $15 ☐ 2 years, $27 ☐ 3 years, $37
All other countries:
☐ 1 year, $18 ☐ 2 years, $33 ☐ 3 years, $45

NAME

NEW ADDRESS

OLD ADDRESS

What makes a CalComp tape drive better?

1. CalComp costs less.

2. CalComp has sales and service offices in over 30 countries around the world.

3. CalComp drives are compatible with IBM 360/30, 360/40, 360/50 computers and above. They're plug compatible with all IBM Systems 370 tape drives. And they're compatible with UNIVAC Series 9400 and 1100 computers.

4. All CalComp tape drives are self-loading and feature soft tape handling. (Only 5 rotating elements in the tape path.)

For a complete analysis of our ability to supply your tape drive needs, call your nearest CalComp office or contact: California Computer Products, Inc., SA-M3-75, 2411 West La Palma Avenue, Anaheim, California 92801, Telephone (714) 821-2011.

CALCOMP

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

MARCH, 1925: "One of the most remarkable astronomical announcements has recently been quietly made from the Mount Wilson Observatory, namely the first reliable determination of the distance of a spiral nebula from the earth. We have known for a decade that these extraordinary objects must be very remote; now Major E. P. Hubble has shown that they are even farther away than had been thought. Hubble discovered last year some faint, variable, starlike points on the edge of the Great Nebula in Andromeda. Further study was laborious, but it became plain that these objects are really stellar, that their variability is regular and that it is of the so-called Cepheid type. What makes this class of variable stars so useful is that, in every case tested, the real or absolute brightness of such a star depends only on its period of variation. The longer the period, the brighter the star. Hence if an astronomer can detect a Cepheid, can find out how long it takes to vary and can determine how bright it looks, he can figure out at once how far off it is. Hubble's newly discovered variables in the Andromeda nebula are perfectly good Cepheids. He has worked out the periods for 10 of them. When their distances are computed, they are found to be substantially the same for all 10 and at the enormous figure of a million light-years. This measurement leads to conclusions that are enough to make an astronomer gasp. The diameter of the nebula must be some 35,000 light-years, and over-all it must shine a billion times as brightly as the sun. Whether the nebula is actually composed of stars or whether something else may cause its light we do not yet know."

"The most important development of the past two years in radio is the increasing ease with which two or more broadcasting stations can be linked by telephone lines. Interconnection ultimately means national programs, nationwide utterances and more valuable subject matter. Nationwide communication has already been accomplished on several occasions, and the interconnection of stations within extensive areas is now a nightly practice. All of this has happened in the past year and it is transforming broadcasting from a local to a national service."

"Great changes in human affairs take place inconspicuously. The substitution of bronze for flint, of iron for bronze, of the printing press for the scribe and of mechanical power for human labor—each of those events occurred so gradually that probably even those directly concerned hardly realized what was going on or appreciated its significance. A case in point is a cultural change now in progress that promises to be as profoundly revolutionary as any that have preceded it. This change is the gradual abandonment of man's most ancient tool—fire. The first effective step toward a fireless future was the substitution of the electric lamp for a flame for illumination. Next came the use of the electric motor in the place of numberless small steam engines and their necessary boilers and fire boxes. The next step, and the one in which the electrical industry is at present particularly interested, is the substitution of electricity for fire in producing heat for industrial purposes."

SCIENTIFIC AMERICAN

MARCH, 1875: "We regret to announce the death of Sir Charles Lyell, a veteran scientist whose labors in the field of geology have gained for him universal renown. He was a native of Scotland and was born in 1797. His *Elements of Geology* and *Principles of Geology* are his two most valuable works, while his more recent volume, *The Antiquity of Man,* is perhaps the most important contribution yet made to that branch which connects Lyell's favorite science with the whole problem of the universe and its origin."

"What are bacteria? In their microscopic field of existence the great battle of biology—the problem of life's beginning—must be decided. So too does one of the greatest problems of pathogenesis hinge on their origins and effects. Are they or are they not the cause of contagious diseases? The bacteria are not animals, nor are all agreed that they are vegetables. For these and other doubtful organisms of the lowest rank Ernst Haeckel has proposed a new kingdom—the Protista, intermediates between and connecting the animal and vegetable kingdoms, and from the modification of which both animals and plants have been derived. Barring the last clause, the proposition bids fair to be generally adopted, as it relegates to a kind of no-man's-land a group of organisms in which animal and vegetable characteristics are so united that they cannot be assigned to either kingdom."

"Few books have more than a temporary life; like the daily newspaper, nine books in every 10 are read and thrown aside. (This leaves out of account the great mass of books that have no purpose and are never read.) Not one copy in a million is worn out by use, yet most books are printed and bound as though they were to be used forever. As a consequence few men can afford to buy many books, and those who do buy have to stand the excessive cost of small editions. We believe the successful book maker of the future will print for the million as well as for the few, and that good books, particularly scientific books, could be sold for a quarter the price now asked. Printed on thin yet clean white paper, the books could be sold, unbound, for the price of a magazine."

"After a century spent in spoiling our woodlands, we are, as a people, slowly awakening to the fact that the chief end of man is not to cut down trees. We are beginning to learn also that, so far from being incompatible with forests, permanent civilization is impossible without them, that the tree slayer's ambition to bring the whole land under tillage would result, if successful, in making tillage a waste of labor. Alternations of drouth and deluge, blighting heats and blasting colds have ever been the penalty for general forest destruction. In mining regions miles and miles of mountainous country have been stripped of timber and scourged by fire until nothing remains but blackness and desolation."

"It is one of the singular facts connected with modern literature that the deep and striking poetry of the rail and the locomotive has never yet inspired any man of genius to sing it forth to the world. Our poets go mad over the achievements of the Greeks and Romans, and seem blind to the everyday marvels of the age we live in. If Dr. Samuel Johnson held it the greatest joy of human life to travel in a post coach with four horses at the rate of 12 miles an hour, the man who rides the Flying Dutchman must surely feel a joy of superhuman intensity by getting through space five times as fast."

For the best performance, get a manual turntable. For the best manual turntable, get a Pioneer PL-71.

At Pioneer, we're convinced that a manual turntable is the best possible component for reproducing the sound of records. That's why we offer a complete line of manual turntables — topped by the truly superior direct-drive PL-71.

Strobe light to adjust and monitor for accurate speed.

Perfection in a turntable means just three things — (1) elimination of vibration or rumble, (2) accuracy of rotation, (3) unwavering consistency of speed, or elimination of wow and flutter. The PL-71 is magnificent in all three areas. Ultra-precise speed controls plus a built-in strobe help you adjust and monitor speed for absolute accuracy. The slow, stable direct-drive motor reduces vibration to the vanishing point and keeps wow ahd flutter well below an undetectable 0.05%.

Direct-drive for precision platter rotation.

Electronic speed adjustment for each speed.

As for record groove tracking, our S-shaped tonearm is tops, too, with its adjustable anti-skating, viscous damped cueing and direct-reading tracking force calibration.

PIONEER
when you want something better

S-shaped tonearm for optimum groove tracking.

The Pioneer PL-71 is the best of the manual turntables. And a manual turntable is best for your listening needs. With its extraordinary performance, design and engineering, the PL-71 is only $299.95. If you want the best of the best, you want the PL-71. Other Pioneer manual turntables are available from $99.95.

U. S. Pioneer Electronics Corp., 75 Oxford Drive, Moonachie, New Jersey 07074
West: 13300 S. Estrella, Los Angeles 90248 / Midwest: 1500 Greenleaf, Elk Grove Village, Ill. 60007 / Canada: S. H. Parker Co.

Another pocket-sized scientific achievement: The HP-21. $125*

We designed it for what we believe to be a growing class of calculator-oriented people who appreciate the very high quality that we put into HP pocket calculators but who don't require all the features and functions of our HP-45, 55, and 65.

The HP-21 is more compact than our other pocket-sized calculators. Closer to palm size. The minute you've grasped its keyboard's elegant simplicity, you'll see that it has considerably more capability than our pace-setting HP-35, presented in an easily comprehensible way.

- An addressable storage register with four-rule arithmetic in and out.
- Rectangular-to-polar conversion and vice versa.
- Degree/radian selection.
- Choice of floating-point or scientific notation, and control over the number of places displayed. When numbers get too large or too small for the floating-point, the display automatically switches to scientific.
- Single-keystroke calculation of a common antilog (10^x).

In addition, the HP-21 gives you

- Keys for all the basic trigonometric functions (sin, cos, tan, $\sin^{-1}$, $\cos^{-1}$, $\tan^{-1}$).
- Keys for e^x, y^x, Ln, Log, $\sqrt{x}$, 1/x, x⪌y, π.
- A 10-digit display and 2-digit exponent. Rounding to last displayed digit.
- Error displayed if you key in an impossible instruction.
- The HP logic system and four-register stack that save you keystrokes, show you all intermediate results, let you proceed naturally through the most complicated expressions without restructuring or keeping track of parentheses, and let you backtrack to correct errors.
- Function keys where you can see them without moving your hand (unless you're a southpaw).
- Execution of any function in less than one second.

In keeping the price down, we haven't skimped on traditional HP quality details, some of which you can see and feel, many of which are hidden.

- Double-injected keys. The legends go all the way through and will never wear off.
- Positive click on each key so you'll feel the connection.
- Moisture barrier under the keyboard to protect against spills.
- Single-unit battery pack that can only be inserted the right way, plus recharger.
- Gold-plated internal connections.
- Recessed display for better visibility outdoors.
- Rugged case with rounded edges. Withstands most tumbles to the floor.
- Soft vinyl zippered carrying pouch.

You will find the HP-21 at leading department stores and college book stores throughout the United States, along with our other hand-held models for science and business. For the name of your nearest dealer, write to Hewlett-Packard, or call toll-free, (800) 538-7922 (in California it's (800) 662-9862).

This new instrumentation recorder could be the only one you'll ever need. And it's so quiet all you hear is its purr.

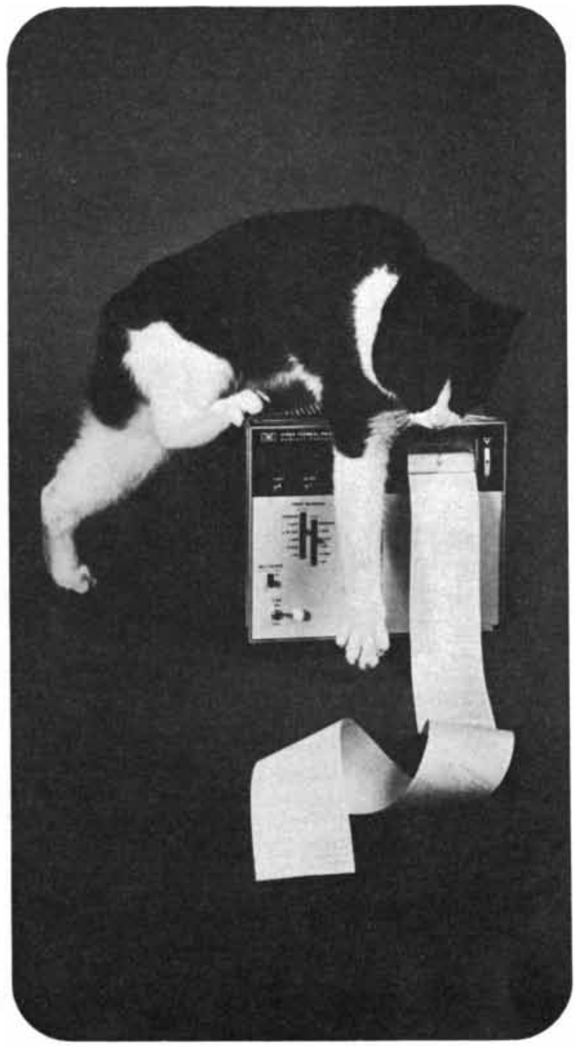
Now that digital measurements are the thing, many laboratories find that they must add a digital recorder to their list of basic equipment. On the surface, this looks like only a minor complication, but it can become a big problem. If the scientist is not careful in making his selection, he may well find that he needs a different kind of digital recorder each time he acquires a new digital instrument.

Our new 5150A Thermal Printer can help you avoid this costly equipment escalation because it satisfies the vast majority of digital recording requirements. Take the problem of different instrument output codes for example. The 5150A handles most cases: it has plug-ins for direct interfacing with ASCII-byte serial and BCD parallel outputs. It will even accommodate a combination of the two codes.

The 5150A has a full 64-character alphanumeric printing set and 20-column capacity when equipped with the ASCII plug-in. With the BCD plug-in, it offers a choice of 10 or 20 columns, each with a flexible 16-character set. If the standard set won't do, it can be factory-programmed to meet almost any desired character need at low cost.

With the scanner plug-in, the 5150A saves you systems design time and equipment cost by providing automatic data acquisition from as many as 13 ASCII-coded instruments on the Hewlett-Packard Interface Bus. Add the clock option and the 5150A records the time of each scan cycle—which can be set at selected intervals from two seconds to one hour.

Once you get the 5150A in your laboratory, you'll particularly appreciate the quiet of its thermal printing technique. It has no print wheels, hammers or inking system—no clank or rattle—just a quiet purr. Priced at $800* for the mainframe and $110* to $250* for the plug-ins, the 5150A is easily the new price/performance leader in instrumentation recorders.



For more information on these products, write to us. Hewlett-Packard, 1502 Page Mill Road, Palo Alto, California 94304.

*Domestic USA prices only

00542

Mail to: Hewlett-Packard, 1502 Page Mill Road, Palo Alto, CA 94304

( ) HP-21 Scientific Calculator

( ) HP 5150A Thermal Printer

Name__________ Title__________

Company__________

Address__________

City__________ State______ Zip______

HEWLETT *hp* PACKARD

Sales and service from 172 offices in 65 countries.

Palo Alto California 94304

THE AUTHORS

LARS E. BJÖRK ("An Experiment in Work Satisfaction") is a social psychologist on the staff of the Swedish Council for Personnel Administration. After completing his undergraduate studies in sociology in 1964, Björk worked for a time at the Tavistock Institute of Human Relations in London before returning to Sweden to attend graduate school at the University of Uppsala, where he received his master's degree in 1967. Since that time he has devoted himself to research and writing on the psychological and social consequences of the way in which work is organized in various Swedish industries.

HERBERT GURSKY and EDWARD P. J. VAN DEN HEUVEL ("X-Ray-emitting Double Stars") view the X-ray sky from both sides of the Atlantic. Gursky is at the Harvard College Observatory, where he went after working for a number of years at American Science & Engineering, Inc. As project director and later vice-president of that company's Space Research Division he was involved in the design and construction of the *Uhuru* X-ray-detecting satellite. He was graduated from the University of Florida in 1951 and went on to obtain an M.S. from Vanderbilt University in 1953 and a Ph.D. from Princeton University in 1959. He taught physics at Princeton and at Columbia University for several years before joining the staff of American Science & Engineering in 1961. Van den Heuvel is professor of astronomy at the University of Amsterdam and visiting professor at the Flemish Free University of Brussels. He majored in physics and astronomy as an undergraduate at the University of Utrecht and did his graduate work at both Brussels and Utrecht, earning a Ph.D. from the latter university in 1968. After a year as a postdoctoral fellow at the Lick Observatory he returned to Utrecht, where he remained until his appointment last year as professor at Amsterdam.

ALPHONSE CHAPANIS ("Interactive Human Communication") is professor of psychology at Johns Hopkins University. A graduate of the University of Connecticut, Chapanis received his M.A. and Ph.D. degrees from Yale University in 1942 and 1943 respectively. He joined the staff of the Systems Research Project at Johns Hopkins in 1946 and has been continuously associated with the university ever since. Chapanis is a past president of the Society of Engineering Psychologists and the Human Factors Society.

PETER J. WYLLIE ("The Earth's Mantle") is professor of petrology and geochemistry at the University of Chicago. Born in London, Wyllie studied at the University of St. Andrews in Scotland, receiving his Ph.D. in geology there in 1958. He was originally attracted to geology, he writes, "because I wanted to see the world and work outdoors. At first this worked out well. I spent two years driving a dog sledge in geological survey work on an expedition to Greenland. Then I became involved in experimental geology, synthesizing rocks and reproducing deep-seated rock processes in high-pressure/high-temperature apparatus. Since then my work has been mainly in the laboratory, or in front of a typewriter trying to catch up with a backlog of research papers arising from the data supplied by the apparatus." Before moving to Chicago in 1965, Wyllie taught and did research at the University of Leeds and Pennsylvania State University.

W. A. H. RUSHTON ("Visual Pigments and Color Blindness") is Distinguished Research Professor in the Institute of Molecular Biophysics at Florida State University. For most of his professional career Rushton was associated with the University of Cambridge, first as a medical student and researcher at Emmanuel College in the 1920's and later as a Fellow and director of medical studies at Trinity College. In 1948 he was elected a Fellow of the Royal Society for his work on nerve excitation. Thereafter he became increasingly interested in the role of pigments in vision.

ANDREW A. BENSON and RICHARD F. LEE ("The Role of Wax in Oceanic Food Chains") began their collaboration on the work reported in this article at the Scripps Institution of Oceanography, where Benson is professor of biology and director of the physiological research laboratory and where Lee until recently was a graduate student and then a postdoctoral fellow. Benson joined the staff of Scripps in 1962 after teaching at the University of California at Berkeley, Pennsylvania State University and the University of California at Los Angeles. A Berkeley graduate, he obtained his Ph.D. in organic chemistry and neurophysiology from the California Institute of Technology in 1942. In 1950 he was the co-recipient (with Melvin Calvin) of the Sugar Research Foundation Award, and in 1962 he won the Atomic Energy Commission's E. O. Lawrence Memorial Award in Nuclear Science. Lee, who was born in Shanghai in 1941, studied and taught chemistry at San Diego State College before going to Scripps in 1966. After receiving his Ph.D. in marine biology there in 1970, he spent one postdoctoral year at Penn State and another back at Scripps before taking up his present posts as assistant professor at the University of Georgia and research biologist at the Skidaway Institute of Oceanography in Savannah.

WALTER LITTEN ("The Most Poisonous Mushrooms") has been an anonymous contributor to the pages of SCIENTIFIC AMERICAN since 1949, when he began writing the Eastman Kodak advertisement that appears regularly in this magazine. He plans to retire from full-time employment with Kodak in June and to move with his wife Mascha to Maine, where he contemplates "devoting myself to a floristic study of the fleshy fungi of the Maine coastal fog zone with the guidance and collaboration of Alexander H. Smith, who has just retired from the University of Michigan and is widely regarded as America's senior agaricologist." No mean agaricologist himself, Litten played a role in providing the reference to an antidote for *Amanita phalloides* poisoning in the case described in his article.

STILLMAN DRAKE and JAMES MacLACHLAN ("Galileo's Discovery of the Parabolic Trajectory") are historians of science at the University of Toronto. Both came to their present calling after first pursuing other careers. Drake worked for more than 30 years in the investment field, specializing in bond issues for public works, before joining the faculty of the Institute for the History and Philosophy of Science and Technology at Toronto in 1967. He obtained his B.A. in philosophy from the University of California at Berkeley in 1932 and returned to receive an LL.D. there in 1968. MacLachlan acquired his first degree (in engineering) from Toronto in 1950; after two years of radio research for the Canadian government he spent the next 15 years as a high school teacher of physics. His original interest in the history of science, he writes, "developed from a desire to provide effective and accurate illustrations in my teaching." That interest led him eventually to Harvard University, where he earned a Ph.D. in the history of science in 1972.

British Columbia, Canada.

These pictures are just a sample of what is waiting for you in British Columbia. 1. One of many intriguing shops that can be found throughout the Province. 2. Long Beach on Vancouver Island, 11 miles of unbroken beach on the Pacific Ocean. 3. The Williams Lake Stampede. Dozens of rodeos take place all summer long in British Columbia's cattle country. 4. An outdoor restaurant in Gastown, the original settlement of British Columbia's largest city, Vancouver. For more pictures and lots more information write: British Columbia Department of Travel Industry, 1019 Wharf Street, Victoria, British Columbia V8W 2Z2. Or see your local travel agent.

For the time of your life.

Choose one of these critically-acclaimed multi-record concert albums for up to 75% off!

It's your introduction to an exciting program that lets you audition important new classical recordings months before the general public!

Now you can enjoy "private debuts" of the greatest new classical recordings in your own home, months before the general public! How? By accepting our invitation to choose any one of the albums from the outstanding repertoire of past International Preview Society selections shown on these pages as your special introduction.

The International Preview Society—more than a record club!

The International Preview Society is a unique record program dedicated to bringing you the finest classical compositions in their most splendid performances. Only when the Society's Carnegie Hall Selection Committee—headed by Julius Bloom, Executive Director of Carnegie Hall—discovers a new album of extraordinary merit and importance will it be sent to you for your free audition. There are no minimum records you must purchase, nor will you receive a selection each month. In fact, there are often months when there are no albums available that meet the Society's high standards!

Enjoy advance audition privileges for 10 days free!

As a member of The International Preview Society you will be able to audition every Preview Selection long before it is released to retail stores, record clubs—even critics! You'll have the opportunity to listen to and enjoy the album for 10 full days . . . and then decide. You may keep any selection and enjoy substantial savings over the future suggested list price or . . . if you are not thoroughly delighted . . . just return the album and your obligation ends there.

All International Preview Society albums are superb, silent-surfaced European pressings and come complete with a handsome hinged storage case, illustrated libretto and *exclusive* Carnegie Hall Selection Committee Report detailing why that particular album was chosen as a Preview Selection.

A critically-acclaimed album . . . yours for only $7.98!

Choose any one of the outstanding concert albums shown on this page to see for yourself why The International Preview Society is consistently first with the best. Many of these recordings have won the highest critical acclaim and have been designated recordings of special merit by major music critics. But don't take our word for it (or even theirs)—you be the judge!

Mail the attached card or coupon at right indicating the album you'd like to receive as your special introductory selection today! You'll receive the album of your choice, and will reserve the right to audition future selections from The International Preview Society.

But please act quickly . . . the albums in this magnificent introductory selection will be available only as long as the supply lasts!

409. Handel—Concerti Grossi. Op. 3 (3 LPs) Raymond Leppard, English Chamber Orchestra

412. Mozart—Don Giovanni (4 LPs) Colin Davis, Royal Opera House, Covent Garden

411. Chopin—The Complete Works for Piano and Orchestra (3 LPs) Claudio Arrau, London Philharmonic

408. Schumann—Complete Piano Trios (3 LPs) The Beaux Arts Trio

404. Bruckner—Symphony No. 5 (3 LPs) Bernard Haitink, Concertgebouw Orchestra

405. Schubert—Winterreise (3 LPs) Hermann Prey, Wolfgang Sawallisch

407. Brahms—Piano Concertos No. 1 & 2 (3 LPs) Emil Gilels—Berlin Philharmonic

414. Schubert—Symphonies Nos. 3, 4 & 6 (3 LPs) Karl Böhm, Berlin Philharmonic Orchestra

415. Stravinsky—Three Great Ballets (3 LPs) Bernard Haitink, London Philharmonic Orchestra

401. Beethoven—Missa Solemnis (2 LPs) Eugen Jochum, Concertgebouw Orchestra

406. Schütz—Psalms Of David (3 LPs) Hanns-Martin Schneidt, Regensburger Domspatzen Chorus

413. Mozart—The Last Six Symphonies (3 LPs) Josef Krips, Concertgebouw Orchestra

416. Haydn—The Last Six Symphonies (3 LPs) Eugen Jochum, London Philharmonic Orchestra

410. Brahms—Four Symphonies (4 LPs) Claudio Abbado, Vienna Philharmonic, Berlin Philharmonic, Dresden State and London Symphony Orchestras

Special Introductory Offer!

Q80

The International Preview Society
175 Community Drive, Great Neck, N.Y. 11025

Yes! Please send me, for my free audition, the critically-acclaimed concert album I have chosen and indicated by the number below. I may return it after ten days and owe nothing, or keep it and pay only $7.98* plus a small postage/handling charge. This is up to $23.94 off the suggested list price. I will also receive, at intervals of one or more months, free ten-day previews of albums chosen by the Carnegie Hall Selection Committee far in advance of general release. I may keep those albums I wish always for far less than the suggested list price. *I am not obligated to buy any minimum number of albums and I may cancel this arrangement at any time.*

*Sales tax added for New York residents.

Print number of album you've selected here: ______

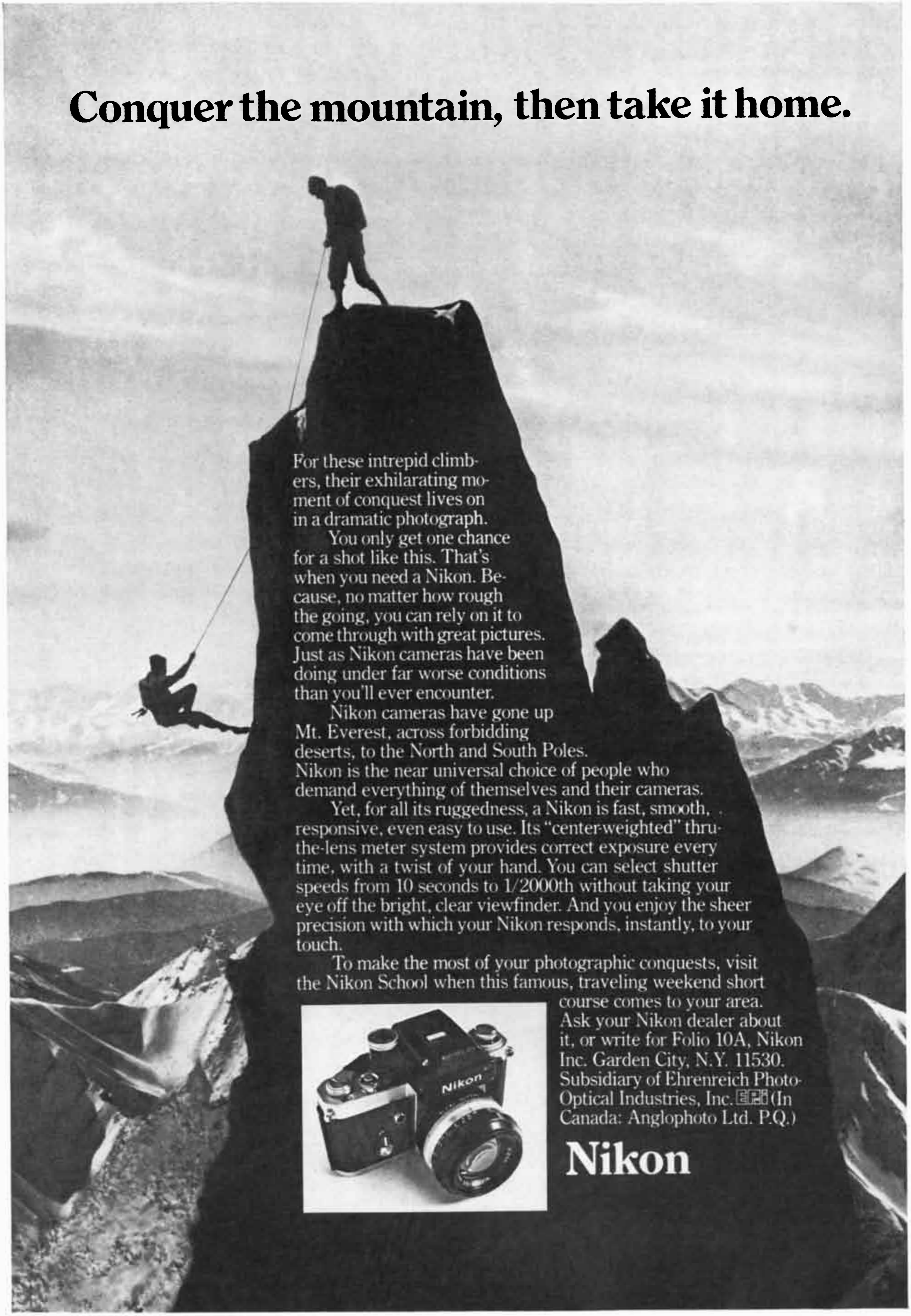
Name ______

Address ______ Apt. No. ______

City ______ State ______ Zip ______

If you are already a member, check here ☐ and include your account number ______. We will honor your order for this album at the special introductory price. **29728/Q80**

Offer limited to the connecting 48 States. Offer expires 9/1/75. Limit one membership per household. Only new members eligible. NOTE: All applications are subject to review, and we reserve the right to reject any application.


Conquer the mountain, then take it home.
For these intrepid climbers, their exhilarating moment of conquest lives on in a dramatic photograph.
You only get one chance for a shot like this. That's when you need a Nikon. Because, no matter how rough the going, you can rely on it to come through with great pictures. Just as Nikon cameras have been doing under far worse conditions than you'll ever encounter.
Nikon cameras have gone up Mt. Everest, across forbidding deserts, to the North and South Poles. Nikon is the near universal choice of people who demand everything of themselves and their cameras.
Yet, for all its ruggedness, a Nikon is fast, smooth, responsive, even easy to use. Its "center-weighted" thru-the-lens meter system provides correct exposure every time, with a twist of your hand. You can select shutter speeds from 10 seconds to 1/2000th without taking your eye off the bright, clear viewfinder. And you enjoy the sheer precision with which your Nikon responds, instantly, to your touch.
To make the most of your photographic conquests, visit the Nikon School when this famous, traveling weekend short course comes to your area. Ask your Nikon dealer about it, or write for Folio 10A, Nikon Inc. Garden City, N.Y. 11530. Subsidiary of Ehrenreich Photo-Optical Industries, Inc. (In Canada: Anglophoto Ltd. P.Q.)
Nikon
Nikon

Established 1845 SCIENTIFIC AMERICAN March 1975 Volume 232 Number 3

An Experiment in Work Satisfaction

Boredom on the assembly line, reflected in high labor turnover and absenteeism, has stimulated efforts to find new ways of organizing work. The author describes one such experiment in a Swedish plant

by Lars E. Björk

In 1913 Henry Ford combined two inventions of the 19th century, interchangeable parts and the subdivision of human labor, with 20th-century management practice and began to produce flywheel magnetos along a precisely timed conveyor belt. He thereby brought into being the era of mass production, which revolutionized the industrial world and reached its fullest expression some 50 years later in the intricately engineered high-speed automobile assembly lines of the U.S. and other advanced industrial countries. Now that era is itself being challenged. Its methods of production are criticized as being out of congruence with human needs, at least in some of the affluent and educated societies where those methods have been carried the furthest. Among the results cited are such symptoms of worker disaffection as high labor turnover, chronic absenteeism, poor workmanship and even sabotage. In Sweden industrial enterprises are experimenting with new ways to structure work that are more in accord with the motives and abilities of the people who do the work.

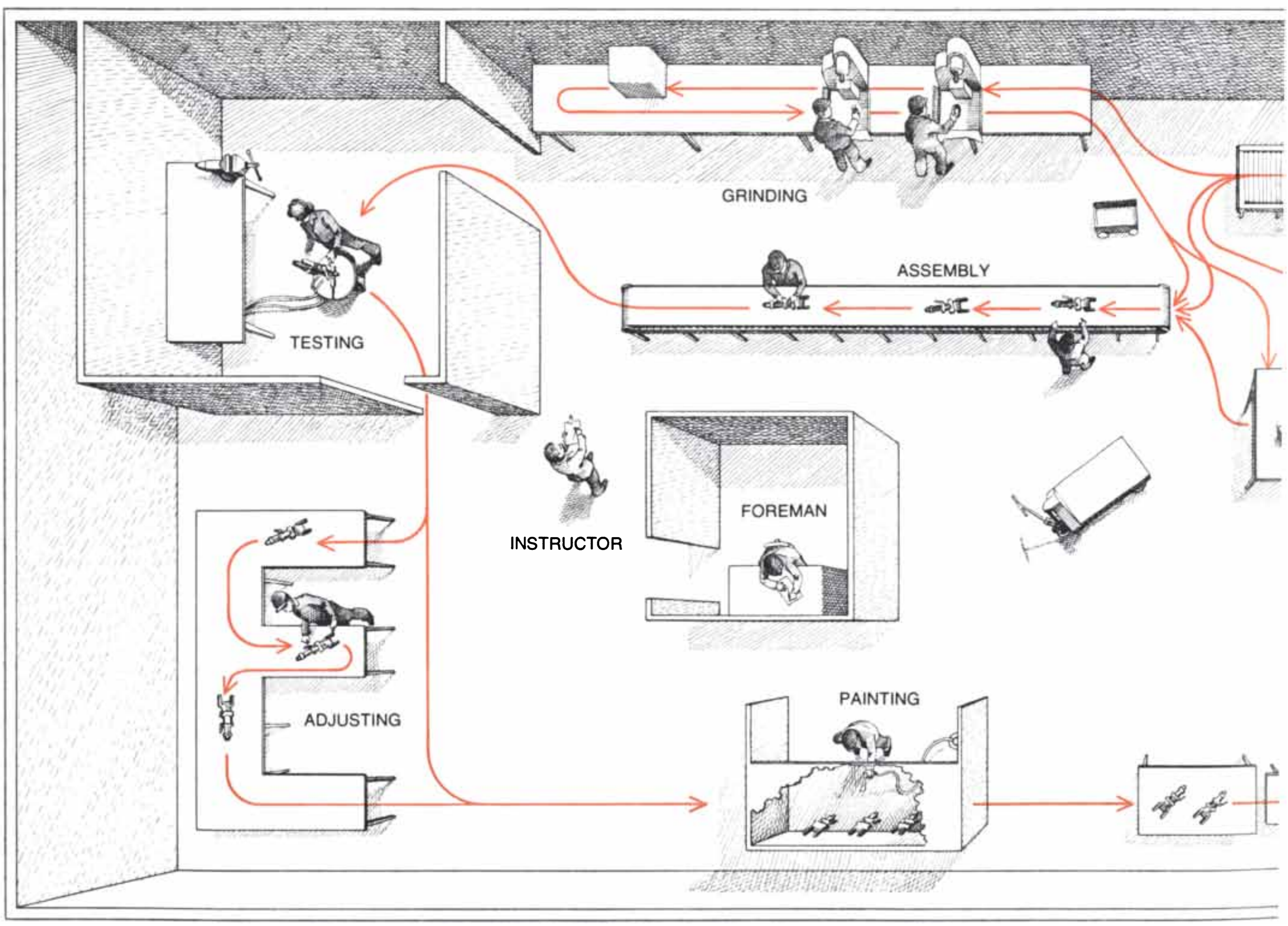
The woes of mass production are familiar, and I shall review them only briefly. The principles are job simplification, repetition and close control. The worker is viewed as one more interchangeable part, programmed to perform a small task that is precisely specified on the basis of time and motion studies. He is assumed to be a passive element in the production process, motivated primarily by his economic needs and characterized primarily by a predictable degree of strength, agility and perseverance; innovation and dealing with variations in the flow of production are considered beyond his scope and are left to specialists. In order to energize and coordinate some dozens or hundreds of atomized human "parts" in a plant, a rigorous and highly detailed control system is called into play, exemplified in its most extreme form by the balanced, intricately interwoven network of conveyors that constitutes an automobile assembly line.

Mass production has had its advantages, better pay and a proliferation of consumer goods among others. In the 1960's, however, particularly in the affluent societies of northern Europe and the U.S., its disadvantages were increasingly perceived, first by workers and their unions and then by industry. The jobs were dull and tiring and destructive of the worker's self-esteem. Things got worse as monotony and exhaustion led to resentment and sloppy work; the response—simpler jobs, tighter control and more speed—made for further monotony, exhaustion and resentment. Turnover and absenteeism became major problems. The assembly line and the minute subdivision of labor made companies particularly vulnerable: the absence of only a few men could jeopardize production, and the line itself was inflexible in the presence of variations in demand or in the delivery of parts.

All of this was accentuated in Sweden. An implicit respect for engineering and engineers, a passion for efficiency and hard work, a highly competitive situation in industry and the persistence of piecework pay along with highly rationalized mass-production methods—all combined to heighten the pace of production and the consequent stress on the worker. Perhaps even more important, these developments in industry flew in the face of contemporary changes in Swedish society as a whole. Increased equality had become a primary goal. More young people were staying in school longer, and the educational system itself was dropping its old authoritarian methods and emphasizing initiative, a critical attitude toward authority and mutual cooperation in small groups. The unions, the powerful metalworkers' union in particular, responded to their members' demand for better working conditions and specifically for more direct influence over their own situation on the shop floor as a necessary step toward industrial democracy. Moreover, by the end of the 1960's many industries simply found themselves unable to recruit enough new workers, to keep those they had recruited from quitting and to keep those who stayed with them on the job every day. Turnover climbed as high as a third of the payroll per year for some companies; maintaining a full crew, at an estimated average replacement cost of $2,000 per man, became prohibitively expensive, and with almost full employment in Sweden it was often impossible.

The upshot was that many companies began to wonder if there was not a better way to design jobs and organize the flow of work. They were ripe for the ideas being articulated in Sweden by a number of social scientists, including Bertil Gardell and Reine Hansson, who were criticizing the increasing impoverishment of work and who felt that more satisfactory methods could be developed based on different assumptions about the nature of man, of workers and of workplaces. Human beings are not passive, according to such critics, but alive and curious. They learn, reason, evaluate and strive toward goals. When workers go home, they are grown-up men and women; regarding them as grown-ups on the job changes one's view of what they can do and of the principles of work organization. People need not be motivated by piecework rates; they can be self-propelling, motivated by the task itself as well as by pay. They can work in autonomous small groups shaped to fit the job. The groups can solve problems, learn from the problem solving and derive satisfaction from it. Instead of specifying each task in detail management can simply specify such essentials as quantity, quality and cost, what Philip G. Herbst of the Tavistock Institute of Human Relations in London called "minimum critical specifications." There need be no prescription of just how each job should be done; after a period of learning the workers will find the best solutions on their own, and they will be solutions based on how people really function and what the immediate shop-floor conditions really are. Technology is admittedly a limiting factor in any choice among various organizational possibilities. The model of man that is postulated by the people who make such choices, however, is even more limiting. In other words, given a specific technology there is more than one organizational choice, and the choice can take social requirements into account along with the workers' right to organize their work and design their own jobs.

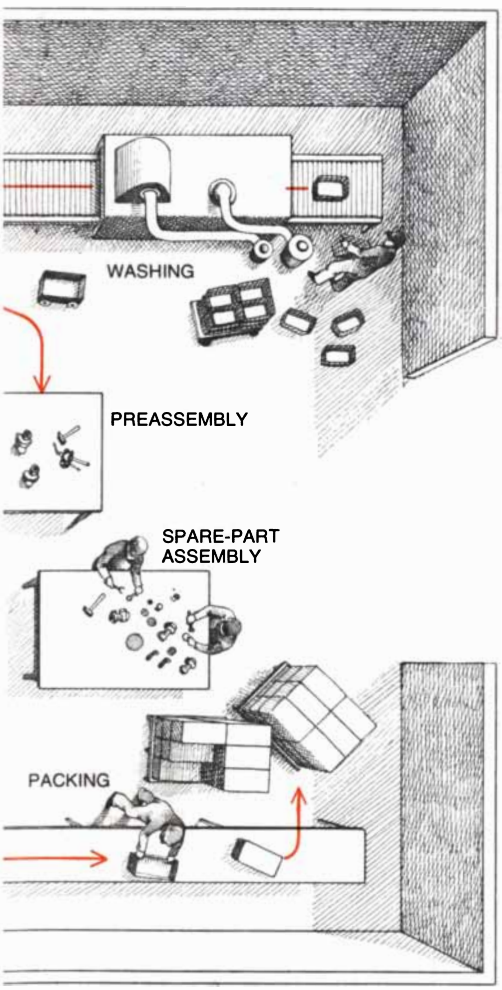
Such views as these stem largely from the "sociotechnical" approach to organizational renewal that was originated in the 1950's at the Tavistock Institute. These ideas were developed and put into practice first by Einar Thorsrud of the Work Psychology Institute in Norway. Experience in Norway, particularly with autonomous groups as the key element of change, soon attracted the interest of Swedish social scientists and manage-



DEPARTMENT 698 is pictured as it was before the experiment in work satisfaction began. The 12 men and their foreman, engaged in the final assembly of rock-drilling machines, were typically disposed as shown here, with each man working at a fixed station. Parts were delivered (*top right*), put through a degreasing machine and then transported to the next stage; some parts required grinding or preassembly before moving to the conveyor-belt assembly line. Once assembled, each drill was tested, painted and packed. The "instructor" handled difficulties in production and, along with the adjuster, was responsible for quality and the correction of

ment and union specialists in working conditions. At the same time demands for industrial democracy were becoming more familiar in politics and particularly in the union press. The process of change began, with efforts at reform initiated in several different ways. In industry quite a few manufacturing companies undertook pilot projects on their own, some of which have developed into major commitments to new methods of production. Perhaps the best-known of these involve autonomous-group production in automobile manufacturing: Saab's new division for engine assembly, opened in 1972, and Volvo's completely new automobile-assembly plant, opened last year.

Meanwhile a more deliberate effort had been organized by the three major Swedish labor-market organizations: the Swedish Employers' Confederation, the blue-collar Confederation of Trade Unions and the white-collar Central Organization of Salaried Employees. In 1966 the three set up a Development Council for Collaboration Questions, which in 1969 established a research task force, URAF (the initials of its Swedish name). URAF sponsored six independent research projects on various aspects of industrial democracy. One of the projects, on work organization and job design, was initiated by our research group at the Swedish Council for Personnel Administration. The other investigations deal with the supervisory function, production planning, office-job organization in a white-collar firm, personnel policy and such worker-representation systems as works councils. Some 30 investigators have been involved in the URAF projects. They are not looking for rules of thumb for increasing productivity or job satisfaction or even hoping for clear-cut, generalizable results as to the effectiveness of various pay systems, autonomous groups, job enrichment or other approaches. The hope is rather to understand the process of change itself: the causal relations, the forces that hinder or support change and the factors that slow the process or speed it up.



errors. Apart from these two men the workers were each assigned a single task. Their pay was based on some form of piecework, and so each man was primarily interested in keeping his own output at a high level.

When the URAF project came along, our group had just spent a year investigating organization and motivation in Department 698 of the Sickla Works, near Stockholm, of Atlas Copco Mining and Construction Technique, Inc. We had been interested in particular in how the motivation of the workers, who were engaged in the final assembly of rock-drilling machines, was affected by job design and work organization. The department had recently been reorganized and rationalized on the basis of time and motion studies. Its personnel, picked for their experience, skill and diligence, had more than met the goals set for the rationalized department, which involved a 25 percent increase in productivity. With the approval of Atlas Copco and the local chapter of the metalworkers' union, we sought a grant from a public foundation funded by the Swedish national bank and organized our new research project during the summer of 1969.

Together with management and the unions, we set up a development group as the decision-making body for the experiment. It consisted at first of representatives of management and the several unions in the plant and three researchers: Peter Hellberg, an industrial engineer, Reine Hansson, a psychologist, and myself, a social psychologist. There were two unusual things about our basic agreement. One was that the overall objective for the project would be to increase work satisfaction among the employees; productivity was not named as a primary goal, although management had made it clear that the current level should at least be maintained. The second was that in order to equalize the distribution of power each party in the development group was given the right to veto any action it interpreted as a threat to its interests. The objective of increased work satisfaction was intended to be achieved by changes aimed at giving the worker more opportunity to influence his own job, to take on responsibility, to solve problems and to advance his own development in the job. We were interested in learning whether such changes could in turn affect motivation to work and productivity. We assumed, however, that changes in these factors could not be measured precisely, as in a stimulus-and-response experiment in a laboratory; rather, we looked on the potential changes as elements in a slow and irreversible process.

The development group decided on a preliminary study that would begin with a survey and analysis of the department and end with proposals for change. We set up a working group (one man from management, one from the blue-collar union and two researchers) to meet weekly and carry out the study. Our analysis made it clear that the social system in Department 698 had become "hung up" on the technological system in ways that contributed neither to fulfilling the unit's objectives nor to the workers' satisfaction.

The department consisted of 12 men and a foreman. The input to the department consisted of the component parts of the rock drills, which came from elsewhere in the plant or were obtained outside. The department's products were some 40 different drill models, mostly variations on six basic types, together with spare parts. Production figures for each item were set by a planning department on the basis of actual orders and expected demand; the year was divided into eight six-week planning cycles (with a four-week vacation in July).

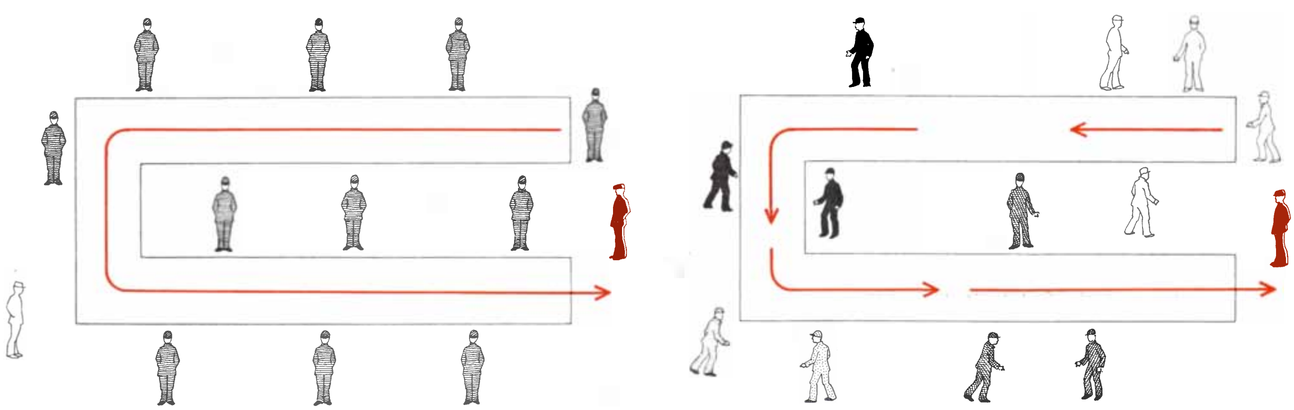
The department operated on a traditional "one man, one machine" basis [*see illustration on these two pages*]. Each of the 12 workers had one of the following tasks: degreasing the incoming parts in a washing machine, grinding and honing the air-throttle elements, preassembly, assembly, testing, painting and packing. There were two floaters: an "instructor" and an adjuster who between them handled quality control, rigged machine tools for new products, checked on unsatisfactory components and corrected errors in assembly. Most of the workers knew only their own regular job and perhaps one or two of the easier other tasks. The foreman directed the work and was also busy adjusting the complicated pay systems. These were of several kinds. The washer, grinders and assemblers, constituting a team, received piecework pay based on the department's output but at various base-rate levels; the tester and the painter were on individual piece rates, again at two different levels. The adjuster and the instructor were paid by the month. The pay levels were decided on the basis of time and motion studies, time studies alone or subjective evaluation by the foreman. There was a great deal of concern and mutual suspicion with regard to rates of pay, and an informal status hierarchy had developed based on skill, job difficulty and pay levels. The two workers on the assembly line felt themselves most exploited: underpaid and bound to one location and to a flow of work they could not control.

The analytical phase had been conducted primarily by the researchers. As we entered the stage of working out changes early in 1970 a need was felt for more participation by the workers in the department. The development group proposed that both it and the working group be expanded to include the foreman and two of the workers. The workers agreed and chose their own representatives. This in effect represented the experiment's first increase in worker autonomy. Through discussions in the working group and through a series of individual interviews that we held with all the workers (interviews that were repeated at intervals throughout the experimental period), we went through each job in the department. Each employee discussed his own job and suggested improvements in its design. Most of the suggestions involved small changes, aimed primarily at equalizing the various tasks and their rates of pay.

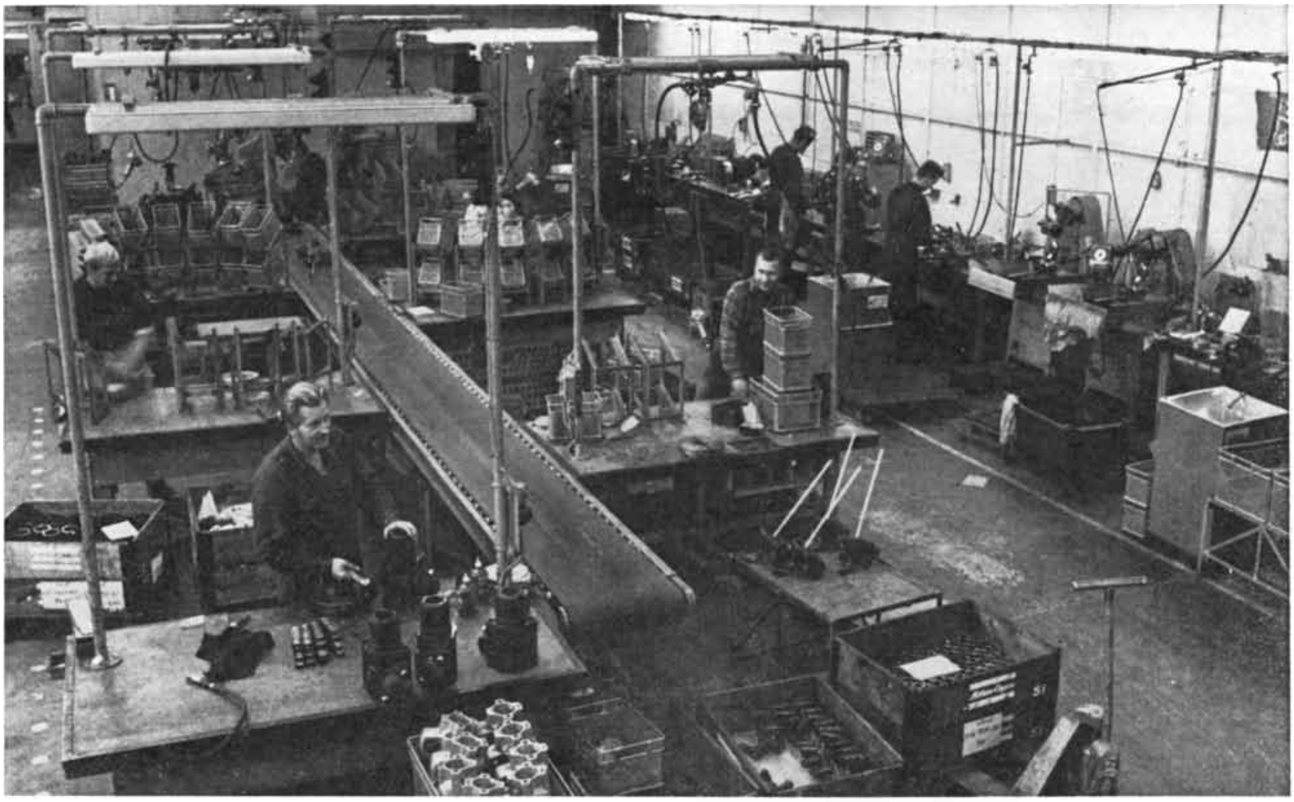
One major change was suggested by almost every worker, however. It was to do away with the conveyor-belt assembly line and substitute a large table on which the drills could be assembled at the men's own pace and in their own way. The company agreed to try it, and after some delay a specially built table designed by the workers was installed and several lots of drills were assembled on it. Rather surprisingly the drills were assembled as quickly as they had been along the production line, even without a breaking-in period. The table made for freer and less monotonous work. On the other hand, it was hard to handle the heavy steel components and move them around and to keep track of the many small parts on the table. Slowly a new point of view developed among the workers. The conveyor belt was seen not as simply a source of unvarying stress but as a tool: it could be used as a means of transportation, with its speed controlled by the workers themselves as they assembled drills on benches alongside the belt in the new ways they had developed by working around the table. Elements of the belt were brought back, having taken on a new meaning for the workers as a result of their own active experimentation; the belt technology could be useful to them because they had broken up the highly specified organizational pattern it had previously entailed. The experience with the table served also to show the workers that management would allow them to work out and test innovations they themselves suggested.

Meanwhile, however, the experiment had run into trouble over wages. The men were uneasy about the economic consequences of the experiment, fearing their pay might suffer in the long run as production methods changed. They asked for some guarantee that the proposals they made would not lead to lower piece rates and thus in effect a speed-up. Management proposed a new system under which 80 percent of their average current earnings would be guaranteed, with the remainder depending on production. The workers considered that worse than either straight piecework or a straight monthly wage and said no, and the working group suspended its activities. After the July vacation there was another unsuccessful proposal from management. At that point the union, anxious lest the entire experiment be ended by the wage problem, stepped in and played a more active role. After some negotiations the workers got their guarantee: Regardless of the level of production, their earnings would not, for the duration of the experiment, fall below their previous average pay.

Once the wage question was resolved it took less than a month for the working group to formulate a set of principles for changes in organization, largely on the basis of ideas developed at four spontaneous meetings of all the members of Department 698. The proposals were not specified in detail but rather outlined some possible and permissible areas of change. When the guidelines came before the development group in November, the management representatives asked for more precision. That was hardly possible, they were told, since the



FLOW OF PRODUCTION changed in the course of the experiment as shown in these highly schematic diagrams. Before the experiment (*left*) the flow was linear; the men were passively bound to specific tasks and locations and the drill parts moved past them. During the experiment the workers were divided into four groups of two to four men (*right*). Each group (*shown by different shading*) is responsible for a single lot of drills and, with each worker able to do most jobs, takes the lot through the entire process.



OVERALL VIEW of the department after the experiment can be contrasted with the drawing on pages 18 and 19. The major physical change is in the assembly system: the conveyor belt remains, but it now serves as a transport system; the drills are assembled on tables adjacent to the belt. Grinding and honing, primarily of the air-throttle parts for the drills, take place along the right wall.

whole idea was to let the systems change in pace with the employees' experience and not to establish detailed job specifications. A series of negotiations reached a new agreement on wages, and after that, in February of 1971, the development group approved the guidelines and a formal one-year experimental period began.

The basic idea was that the workers in the department would constitute a single team, with equal pay. The team would be split up into groups of from two to four men, and each group would move through the department, carrying out the entire sequence of operations for a single lot, or batch, of drills [*see illustration on opposite page*]. The allocation of individual tasks within each group would be left up to the group to decide, as would the exact method of performing each operation. It was to be more than job rotation among previously designed jobs, since the old designs could be scrapped in favor of whatever methods the men found most convenient and satisfying. The wage agreement provided for continuation of the existing guaranteed hourly rate and also an incentive system: a small hourly increment for each percentage point by which the team as a whole exceeded its established production level.

Things began slowly. When there was plenty of time, the foreman would let from two to four men take a drill lot through the entire production process. More and more of the men learned new jobs, and they sorted themselves into permanent groups that gradually came to handle all production. This meant not only the assembly process but also managing the flow of goods through the department. In consultation with the foreman the groups decided, on the basis of delivery schedules and production goals, which lots of drills to process.

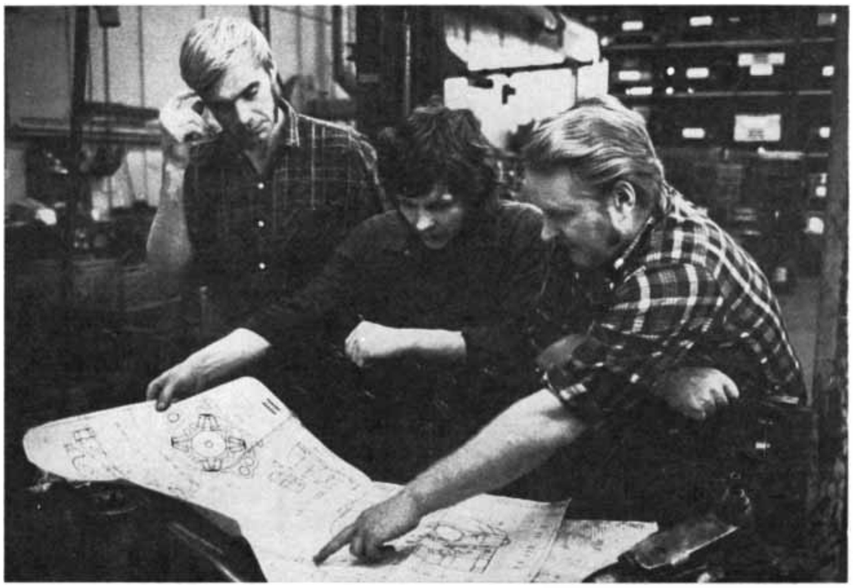
It was not an easy period. The workers were going through a time of intense learning, not only of new operations but also of new social relations: group cooperation in place of the individualistic and even antagonistic relations that had been shaped by the old one-man, one-machine organization. The mood was often one of irritability and tension as the men pressed to maintain the rate of production, partly for fear of losing their guaranteed earnings and partly to show that they could keep up the pace while working in their own way. At the end of the first six months management was pleased with the level of production but disappointed to see so little obvious change in production technique. The large changes in the patterns of work and social relations that the men were experiencing were not easily perceived by management.

Meanwhile two new conference groups, one for planning and one for rationalization, were established so that the men could deal directly with matters that had previously been handled by specialized management departments. The planning group consisted of two workers, the foreman and a representative of the planning department. It met once a week to discuss current production, establish priorities and deal with changes in parts or products. The rationalization group, consisting of two workers, the foreman and production engineers from the rationalization department, met about once a month so that the men could take part in any new technological developments. In these groups the specialists no longer prescribed the workers' actions but rather served their needs by supplying technical know-how and information.

During the second half-year of the experiment the department settled into a new steady state. There were fewer conflicts among the workers, who increasingly concerted their efforts and took on responsibility for the department's objectives and for one another. The foreman, who had been uncertain of his role at first, became more comfortable with his changed status. No longer required to maintain close control of every operation, he too was able to move to a higher level of competence, attending to budgetary and personnel matters and laying out machine tools and systems for assembling new models.

At a meeting of the development group early in 1972 the experiment was formally ended. In effect it was declared a success: management wanted the new system to continue as the ordinary way of doing things in Department 698. The other parties agreed. The working group was disbanded and the development group gave itself the assignment of evaluating the project.

That is not an easy thing to do. The most readily quantifiable change is in productivity, which is measured in Department 698 as the ratio between the man-hours allotted for a certain amount of production and the man-hours actually worked. Productivity has increased about 5 percent, but to report that figure is not to give the whole story. What seems more important is that a considerable amount of learning has taken place, with consequent benefits for long-term productivity. The men know one another's jobs and are anxious to cooperate, and so they can fill in when someone is absent. If several workers are off the floor simultaneously (at a meeting, for example), there is simply one less group in action, whereas the conventional in-line assembly process would have been seriously disrupted; the group system provides what the Tavistock theorists call "redundancy in function." Because the men are more oriented toward mutual goals and much more informed about the flow of materials and the logic of production schedules, they are better able to cope with variations in deliveries and with rush orders. The entire operation has become more flexible.

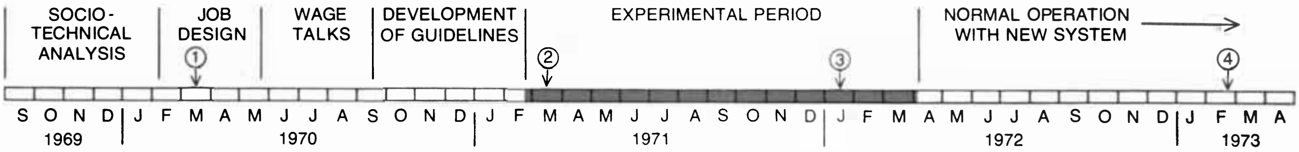


DUTIES ONCE RESERVED to the foreman are now taken on by the workers. For example, they study drawings of the drill they are about to assemble, check on the delivery of the component parts and decide how to lay out the work and when to begin each operation.

GROUP SYSTEM, in addition to relieving monotony, allows a worker to carry the product through the entire production process: to make a drill rather than to perform an operation.

Work satisfaction is even harder to measure than productivity. Absenteeism and turnover, its traditional indicators, were low before the experiment began and have not changed. Perhaps the most obvious sign of increased satisfaction is the fact that none of the men wanted to go back to the old system. Satisfaction is subjective, however, and means different things to different people. What the former assembly-line workers wanted most, for example, was freedom from the line; they got it, and they were pleased. In our interviews most of the men mention the opportunity to perform different operations as the best result of the experiment: "It makes the job more pleasant." Many of them say the mood of the department has changed for the better, something they had not expected: "We've got calmer. It's not as stressful as it was before." They like laying out jobs in their own way: "Makes you feel like you're your own boss." There were exceptions, however. One man told us: "I'd rather work alone than in a group. Then you worry about yourself and you don't have to depend on all the bull all over the department."

As for the workers' power to exert influence, that has manifestly expanded. The experiment was in effect a lesson in wielding influence. The workers' very reluctance to make suggestions at first made their influence felt. In the course of the experimental year it was the workers who increasingly influenced the design of their own jobs and then the movement of material through the department. Behavioral scientists often say that once a need is filled it no longer motivates an individual's actions, and we saw that illustrated in the department. When



WORK-SATISFACTION EXPERIMENT was preceded by a long preparatory phase, as is indicated in this chronological record. The circled numbers along the chart represent four series of interviews conducted by the researchers with each of the workers.

the men had got themselves an acceptable degree of influence in one area, they refocused on the next most important question and thus worked themselves from one issue to another. Their aspirations kept rising: "In the future we want to be in on decisions on the department's budget."

The concept of responsibility broadened during the experiment from individual concern for product quantity and quality to group responsibility for the entire department's production flow. The men assumed that taking on such responsibility was a prerequisite for continuation of the experiment and for an increase in their influence. It involved taking the initiative to keep production moving: "We take a lot of responsibility now. We untangle pileups ourselves." Social responsibility has also broadened. The men help one another with new jobs: "We have to be more considerate of each other now. Before the experiment we didn't have any desire to get together. Everyone worked for himself."

The entire experiment was also an exercise in problem solving, with the request for, and the modification of, the assembly table being perhaps the most obvious example. During the year all the workers were drawn into the solving of practical problems of job design and work organization. Once they had all learned one another's jobs and once the new organization had become set, however, there seemed to be no problems left to solve. Several of the men expressed a desire to keep making new changes and came to feel that their work situation had once again become monotonous. This raises a question about experiments such as ours: Does the active period of change arouse interest and expectations that bring on a kind of relapse into boredom? In Department 698 there is still one good outlet for problem solving, the rationalization conference group, which continues to deal with production technique and changes in the department. Its activities directly involve only two of the workers at a time, however.

As we analyzed the experiment we came to see it as a learning process not only for individuals but also for groups and for the major parties involved. The experiment freed the workers to change their own situation, but experience showed that the freedom to change was utilized only as the workers learned more about their situation. We had assumed that the conditions that would fundamentally affect changes in influence, responsibility and problem solving were management's traditional control mechanisms: work organization, job design, wages, supervision, planning and so on. This turned out to be true on the whole, but we found that the links among those mechanisms were also important. In other words, one cannot change a control mechanism such as wages without considering the effect of the change on organization, job design, planning and so forth. Those relationships also must be learned, by the researchers, the workers and management. Values and attitudes are deeply involved here. In many cases we found management had little understanding of how a control mechanism was perceived by the workers.

The fact that the experiment involved so much learning is perhaps the best explanation of why it took almost three years. Considering the amount of learning that went on at various levels, it was really not such a long time. One does not introduce a new way of working as one installs a new tool or even a new conveyor belt; behavior cannot be changed overnight by the promulgation of new rules. Actually the aim was quite different: to unfreeze the production technology and the grip it had on the social system and the people.

Two questions that need to be raised about an experiment of this kind have to do with reliability and diffusion. How much can be proved by an experiment involving 12 men? Not much, if the object is to measure a lot of factors before and after the experiment and see how they differed—to collect, in other words, what in a social context are incorrectly called "results." What we sought to do, however, was to understand the how and why of the differences, and in an effort to do that we followed the entire process of change, starting a year and a half before the experiment began and conducting our last interviews a year after it ended. Through our continuous contact with management, the unions and the individual workers we have been able to check dubious data in a way that is not possible in a large before-and-after survey. Our examination of the process of change has primarily confirmed the predictions we made, admittedly from a definite point of view: an assumption that people are capable of controlling their own work situation. What we learned about the process of change seems to us to "prove" that our initial assumptions were correct, but there should be additional evidence as the outcome of other URAF experiments becomes known.

Finally, does the experiment say anything about what can happen in the future and what can happen beyond the walls of Department 698? For the process of change to continue there will have to be changes elsewhere in the company. Horizontal diffusion is not important in this case, but the supervisory layers above the assembly department will have to modify their planning and control functions if the men in the department are to gain increased influence and autonomy. Any further diffusion through industry will depend heavily on the attitudes of the major labor-market organizations. As of now the position of the Swedish Employers' Confederation has been to promote practical cooperation at the local level between workers and management in a plant and not to talk about increasing influence. The Confederation of Trade Unions, on the other hand, is anxious to obtain legislation that will secure the worker's right to increased influence by law, even in situations where local cooperation breaks down in conflict. Broad diffusion of the new ideas will also depend, as the beginnings of change did in the 1960's, on the changing values of Swedish society as a whole.

In time perhaps such words as "experiment" and "pilot project"—which imply that one wants to leave the doors open for retrogression, to be able to sneak back to the old way if things go wrong—will seem old-fashioned. We may come to think of the delegation of power and influence to the worker as being quite normal and undramatic.

X-RAY-EMITTING DOUBLE STARS

An analysis of certain very powerful X-ray sources suggests that the radiation emanates from a binary system where a superdense collapsed star is orbiting closely around a massive normal star

by Herbert Gursky and Edward P. J. van den Heuvel

Picture a celestial body no bigger than New York City that spews out energy solely in the form of X rays at a rate equal to roughly 10,000 times the total energy output of the sun. The amount of energy emitted each second by such an object would be enough to propel all 300 million of the world's automobiles an average of 70 miles per day for the next 100 billion years—10 times the estimated present age of the universe. Imagine further that this gigantic outflow of energy shows large fluctuations within a thousandth of a second, and that every few days the radiating object describes a circular path in space with a diameter of some 20 million miles.

These facts should give some idea of the bizarre properties of a newly discovered type of double-star system. Eight of these X-ray binaries have been found so far; each is thought to consist of a fantastically dense, X-ray-emitting source orbiting closely around a much larger normal star. Six of the new double-star systems were identified with the aid of a single artificial satellite launched in 1970 from a site in Kenya. The satellite, named *Uhuru* after the Swahili word for "freedom," was built for the National Aeronautics and Space Administration by Riccardo Giacconi and his colleagues at American Science & Engineering, Inc. *Uhuru* has made so many discoveries that it has transformed X-ray observations into one of the most active branches of modern astronomy.

In the seven years before the launching of *Uhuru* dozens of celestial X-ray sources had been discovered with special detectors carried above most of the earth's X-ray-opaque atmosphere by rockets and balloons. Progress in X-ray astronomy was slow, however, since such detectors can operate usefully for only a few minutes per flight, and each research group can carry out only a few flights per year. Although some of the sources could be identified with familiar astronomical objects such as the Crab Nebula, the nature of most of the sources remained a mystery. Of the latter the majority are strong, pointlike sources that show a distinct concentration toward the central plane of our galaxy [*see top illustration on page 27*]. They must therefore be within the galaxy. According to the constellation in which they are located and the sequence in which they were discovered, they were given names such as Scorpius X-1, Scorpius X-2, Cygnus X-1, Cygnus X-2 and so on. Only after the launching of *Uhuru*, which could monitor a source for as many as 24 hours a day, could these sources be studied in detail. This close scrutiny has revealed that some are definitely double stars, and others show such similar characteristics that they probably belong in the same category.

The first evidence of the possible stellar nature of the X-ray sources came in 1966, when the strongest of them, Scorpius X-1, was identified optically with a faint blue starlike object that resembled an old nova, or exploded star. About a decade earlier it had been established that old novas are close binary systems in which one of the stars is a white dwarf. It is thought that in such a system matter is transferred from the normal component to the white dwarf, leading eventually to an explosion in the outermost envelope of the white dwarf. Hundreds of novas explode every year, of which a dozen or so can be seen, sometimes with the unaided eye.

The discovery of the optical counterpart of Scorpius X-1 gave rise to the suggestion that the X-ray sources are binaries and that the X rays are being emitted by a hot cloud around the white dwarf consisting of matter captured from the normal companion star. Other astronomers argued that it was a neutron star or even a "black hole" in a binary system that accounted for the X-ray production. A black hole is a star that has contracted to such a small radius that the escape velocity from its surface exceeds the velocity of light, with the result that light rays emitted from such an object will never be able to reach an outside observer [see "The Search for Black Holes," by Kip S. Thorne; SCIENTIFIC AMERICAN, December, 1974].

In these binary models the X rays are generated by the enormous acceleration

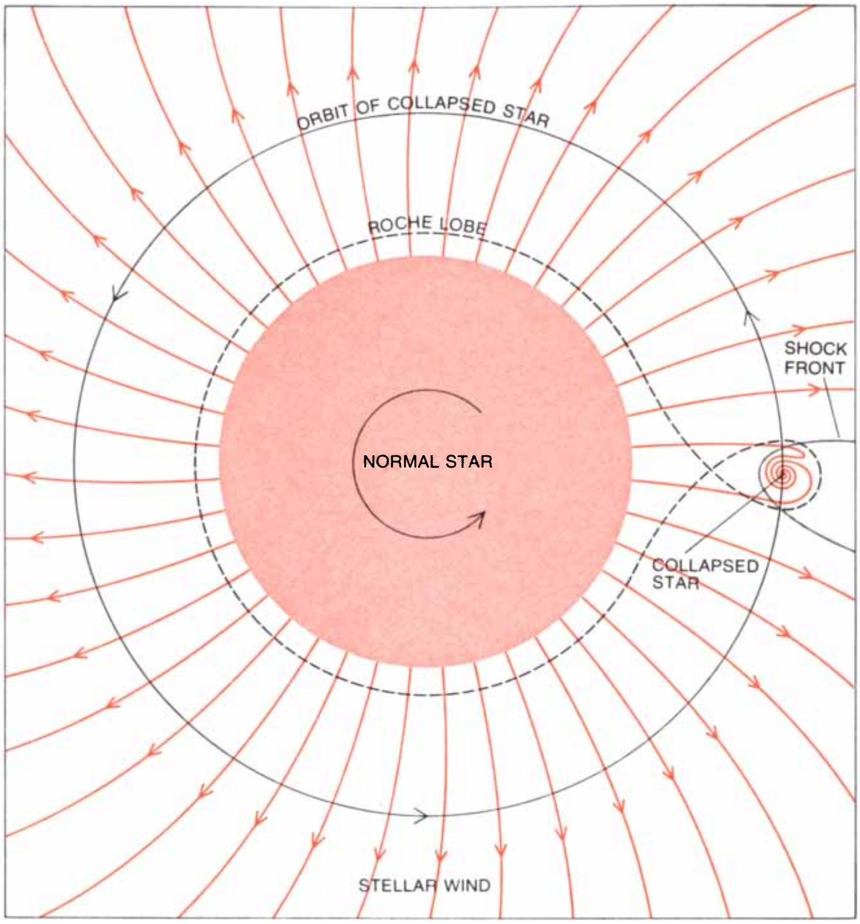
TWO POSSIBLE CONFIGURATIONS for a binary X-ray source are illustrated in the pair of schematic diagrams on the opposite page. The broken black line in each diagram represents that system's Roche lobe, the boundary beyond which an expanding star in a binary system will begin to transfer matter to its companion star. In the configuration at top the normal star is actually smaller than its Roche lobe, but it still loses mass in the form of a supersonic "stellar wind" at a rate equal to about a millionth of the sun's mass per year. The orbiting collapsed star, either a neutron star or a "black hole," is in this case merely an obstacle in the larger star's stellar wind, and a bow-shaped shock front is formed around it by the action of its gravitational field. Some of the material flowing through this shock front is decelerated enough to be captured by the smaller star. The bulk of the material in the stellar wind (some 99.9 percent of it) flows out into space without being affected by the compact star; the .1 percent that is captured is sufficient to turn the companion into a powerful X-ray

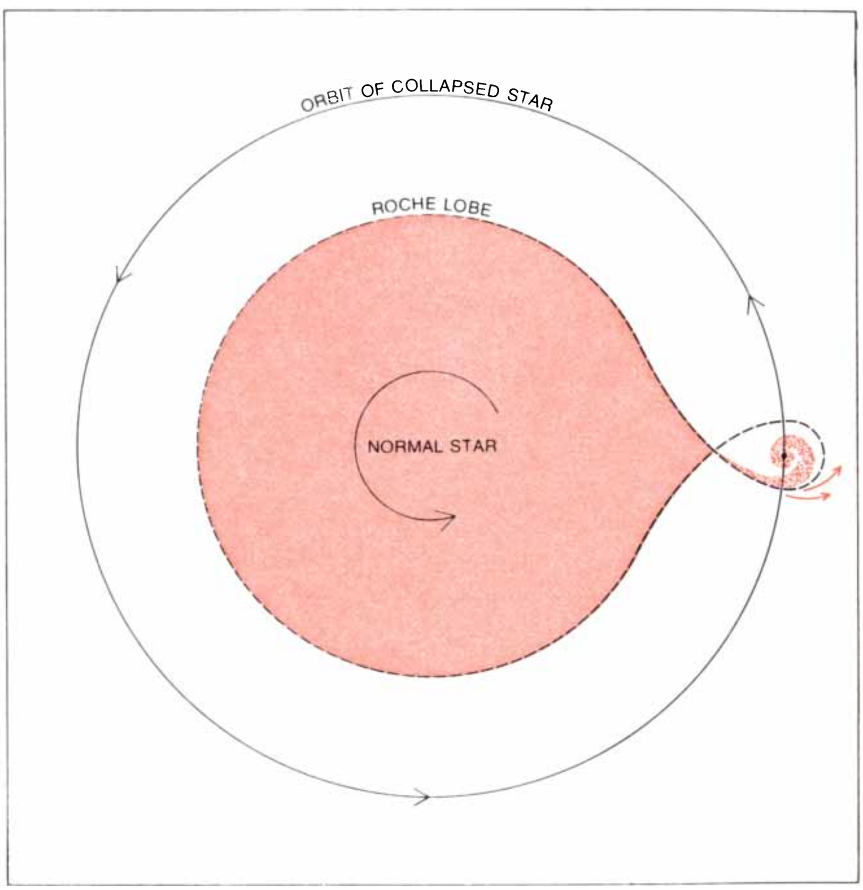
and heating of the accreted matter during its fall toward the compact star. When the falling matter arrives near the star, its temperature will have risen to several tens of millions of degrees Kelvin, and the object will have become a strong X-ray source [*see illustrations on next page*]. In the case of accretion toward a black hole one will observe only the X rays emitted by the matter that is still outside the limit known as the Schwarzschild radius, which is defined as the distance from the center of the black hole where the escape velocity just equals the velocity of light. For a black hole with a mass equal to that of the sun the Schwarzschild radius is about two miles.

At the time these suggestions were made there was in fact little evidence for the validity of such models. In the case of Scorpius X-1 only the spectrum of a hot gas cloud was observed, and in spite of many nights of observing with large telescopes astronomers could not detect any evidence of a binary character. The same turned out to be true for another novalike source, Cygnus X-2.

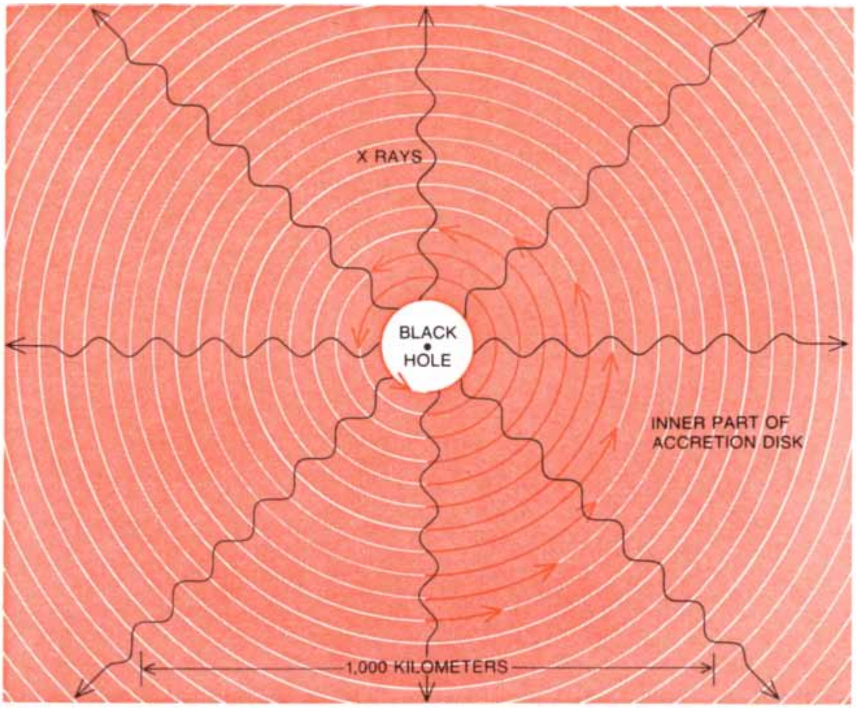
The revival of the binary model came in 1971 when the *Uhuru* team discovered that the X-ray emission from the source Centaurus X-3 varies regularly with a precise period of 4.84 seconds [*see top illustration on page 28*]. Comparison with the pulsating radio sources known as pulsars (already known to be neutron stars) was inevitable, particularly after a second regularly pulsating X-ray source was discovered: Hercules X-1, with a period of only 1.2 seconds. This period is so short that by itself it virtually eliminated the possibility that any stellar configuration less dense than a neutron star could be the X-ray source.

Periodic phenomena have a special importance in astronomy, not only because they often convey immediate information about the physical nature of an object but also because such periodicities can usually be measured with

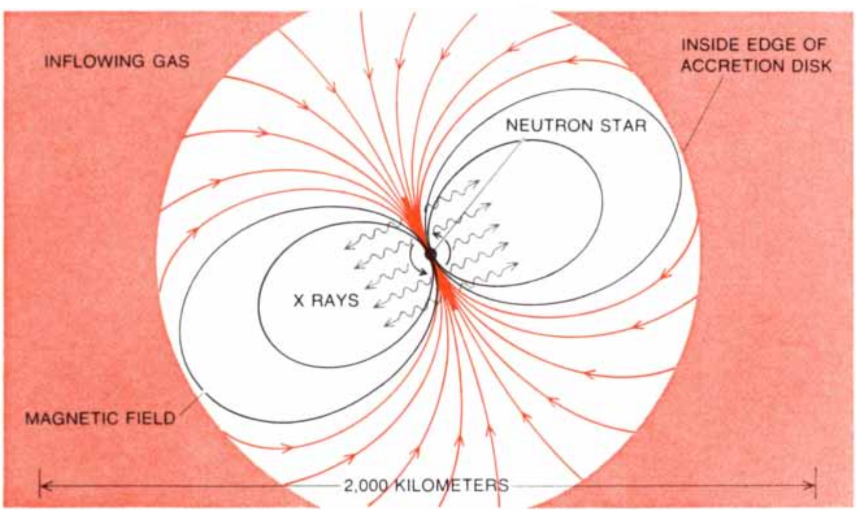




source (*see illustrations on next page*). In the configuration at bottom the normal star has expanded to the stage where it begins to overflow its Roche lobe. The material outside this boundary flows along the Roche lobe toward the compact star at a comparatively low velocity; hence the material lost by the normal star cannot escape from the system and is captured by the companion. This type of mass transfer can produce an X-ray source only if the normal component has a mass no larger than a few solar masses.



RAPIDLY ROTATING DISK of captured matter will form around the compact star in either of the binary systems depicted on the preceding page, owing to the relative motion of the pair of stars in each system. Within this accretion disk the orbital velocities of the individual captured particles of matter will increase toward the center (*colored arrows*). Collisions between the particles will release energy, heating the accretion disk and at the same time reducing the size of the particle orbits; as a result the particles will gradually spiral inward and eventually reach the surface of the star. X rays will be emitted from the innermost few dozen kilometers of the accretion disk, where the particles will attain velocities of some 100,000 kilometers per second and the temperature will reach as high as 100 million degrees Kelvin. Many astronomers believe the highly variable X-ray source designated Cygnus X-1 may be the site of just such an accretion disk around a black hole.



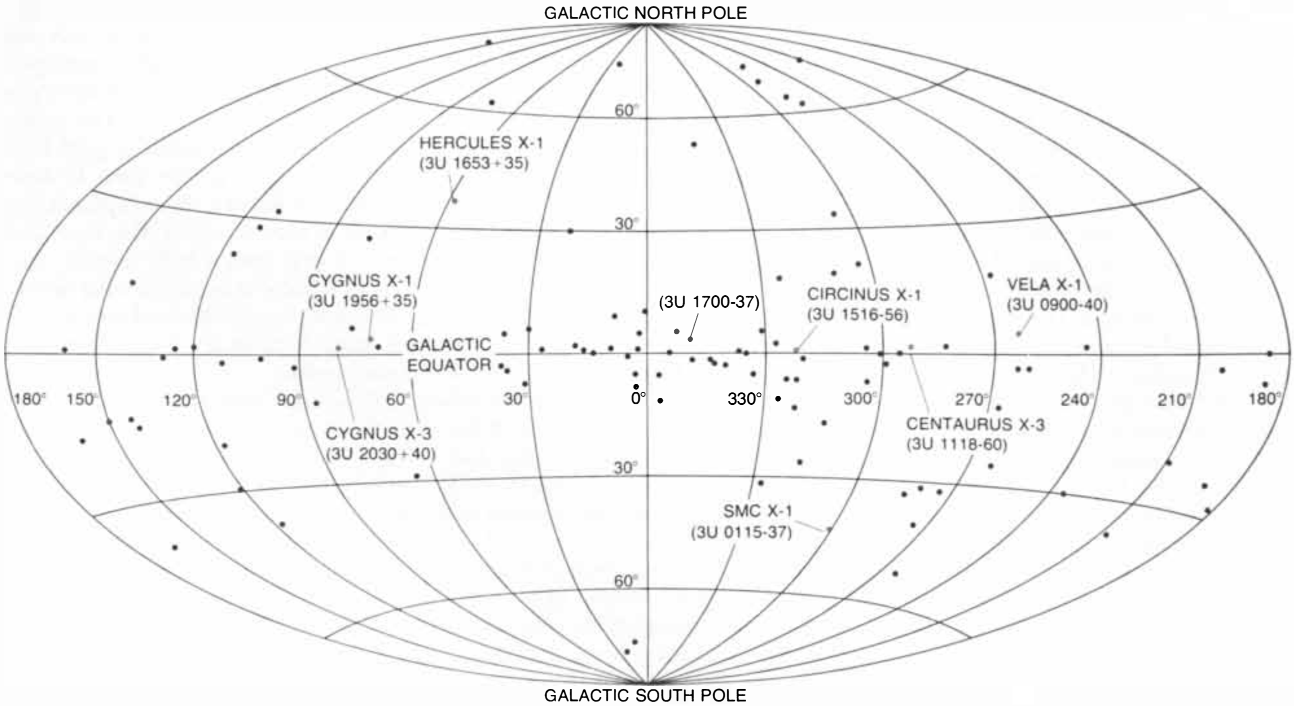
SPECIAL CASE in which the compact star in a close binary system is a neutron star with a strong magnetic field is thought to be accountable for the pulsed X-ray emission observed from such sources as Centaurus X-3. In this model the star's dipole magnetic field prevents the innermost few thousand kilometers of the disk from forming. From the inside edge of the disk matter is able to continue its inward flow only along the "open" magnetic-field lines found above the star's two magnetic poles. The infalling matter enters these two "magnetic funnels," which terminate on the stellar surface in two "hot spots," each with an area of about a square kilometer. The X-ray emission is believed to arise from the two hot spots and from the turbulent gas columns above them. The rotation of the neutron star around an axis that is inclined with respect to the axis of the magnetic field causes a periodic modulation of the X-ray intensity; as a result the source appears as an X-ray pulsar.

high precision, revealing subtle clues that can in turn lead to new discoveries. That, in fact, is what led to the discovery of the binary nature of Centaurus X-3. Ethan J. Schreier of the *Uhuru* team found that the 4.8-second period of the X-ray emission varied by as much as one part in 1,000 per day, a variation that is more than 1,000 times the variation observed in any other "normal" radio pulsar. At the same time the X-ray emission was found to vary by at least a factor of 10. After a few months it was realized that both of these variations repeat precisely every 2.087 days. At once the picture became clear: the regular increase and decrease of the pulsation period simply reflects the orbital motion of the X-ray source around an unseen companion star. The intensity variations of the source result from the fact that during each revolution it disappears for some time behind its companion.

The X-ray light curve of Centaurus X-3 is the curve of a typical eclipsing binary system. The correlation between the observed orbital velocities and the intensity variations is perfect, and the velocity variations can be made to fit with a pure sine wave, as one would expect of a circular orbit. These data establish without any ambiguity that the Centaurus X-3 X-ray source is in a binary system and that the source itself is a compact star. Furthermore, from the orbital period and the orbital velocity of the X-ray source it was possible to calculate that the mass of the unseen companion was at least 15.4 times the mass of the sun.

For a long time the companion of Centaurus X-3 could not be found. In the summer of 1973, however, Vojtek Krzemiński, a Polish astronomer working at the European Southern Observatory (ESO) in Chile, discovered a faint star that varies in brightness with the same period as the X-ray source does. On the basis of spectroscopic observations the star has been classified as an early-type giant with an intrinsic luminosity at least 100,000 times greater than that of the sun. The star is in the central plane of the galaxy, and its faintness is due to a combination of distance and the obscuring effect of interstellar matter. The distance, determined from interstellar absorption and reddening, is about 25,000 light-years; the star is almost as far from us as the center of the galaxy is. From the distance and the X-ray flux detected on the earth one finds that the power of the X-ray source is about 10,000 times the total luminosity of the sun.

At the time Cygnus X-1 was being ex-



EIGHT X-RAY SOURCES are identified in color on this map of the galaxy as seen by the X-ray telescope aboard the *Uhuru* satellite. The dots indicate the positions of all the X-ray sources listed in the Third *Uhuru* Catalogue (hence the "3U" designations in parentheses). The map is an equal-area projection in galactic coordinates, which means that the central plane of the galaxy is aligned with the equator of the coordinate system and the galactic center is at zero degrees latitude and longitude. The concentration of the X-ray sources along the galactic plane, particularly toward the center, indicates that most of them lie within the galaxy. The remaining high-latitude sources are probably outside the galaxy. The common names of the X-ray-emitting sources are derived from the constellation in which they are located and the sequence in which they were discovered; SMC stands for Small Magellanic Cloud.

NAME OF X-RAY BINARY	BINARY PERIOD (DAYS)	CHARACTERISTICS OF X-RAY EMISSION	X-RAY LUMINOSITY (x SUN'S LUMINOSITY)	CHARACTERISTICS OF THE VISIBLE COMPONENT	DISTANCE (LIGHT YEARS)	OTHER CHARACTERISTICS
CYGNUS X-1 (3U 1956+35)	5.6	IRREGULAR VARIATION ON TIME SCALE OF BETWEEN .001 SECOND AND ONE MINUTE	10^4	NINTH-MAGNITUDE BLUE SUPERGIANT; MASS LARGER THAN 20 SOLAR MASSES	10^4	SINCE MARCH 1971 ALSO A WEAK RADIO SOURCE. NO HARD X-RAY ECLIPSES
CENTAURUS X-3 (3U 1118-60)	2.087	X-RAY ECLIPSE: DURATION .488 DAY; PULSE PERIOD 4.84 SECONDS	10^4	13.4-MAGNITUDE BLUE GIANT; MASS LARGER THAN 16 SOLAR MASSES	2.5×10^4	EXTENDED LOWS
SMC X-1 (3U 0115-37)	3.89	X-RAY ECLIPSE: DURATION .6 DAY	2×10^5	13.4-MAGNITUDE BLUE SUPERGIANT; MASS LARGER THAN 25 SOLAR MASSES	1.9×10^5	EXTENDED LOWS; ONLY KNOWN X-RAY BINARY OUTSIDE OUR GALAXY
VELA X-1 (3U 0900-40)	8.95	X-RAY ECLIPSE: DURATION 1.7 DAYS; SLOW FLARES ON TIME SCALE OF HOURS	10^3	SEVENTH-MAGNITUDE BLUE SUPERGIANT; MASS LARGER THAN 25 SOLAR MASSES	$.8 \times 10^4$	WEAK RADIO SOURCE
3U 1700-37	3.412	X-RAY ECLIPSE: DURATION 1.5 DAYS	10^3	6.7-MAGNITUDE BLUE SUPERGIANT; MASS LARGER THAN 30 SOLAR MASSES	$.9 \times 10^4$	WEAK RADIO SOURCE
CIRCINUS X-1 (3U 1516-56)	PROBABLY LONGER THAN 15	X-RAY ECLIPSE: DURATION LONGER THAN A DAY; RAPID IRREGULAR VARIATION SIMILAR TO THAT OF CYGNUS X-1	?	?	?	
HERCULES X-1 (3U 1653+35)	1.7	X-RAY ECLIPSE: DURATION .24 DAY; PULSE PERIOD, 1.24 SECONDS	10^4	VARIES WITH 1.7-DAY PERIOD BETWEEN 13TH AND 15TH MAGNITUDE; MASS ABOUT TWO SOLAR MASSES	1.6×10^4	35-DAY VARIATIONS IN LIGHT CURVE
CYGNUS X-3 (3U 2030+40)	4.8	SINUSOIDAL VARIATION WITH 4.8-HOUR PERIOD	AT LEAST 10^4; PEAKS OF MORE THAN 4×10^4	NO VISIBLE STAR; INFRARED SOURCE WITH 4.8-HOUR PERIOD	AT LEAST 2.5×10^4	SOURCE OF INTENSE RADIO-WAVE OUTBURSTS

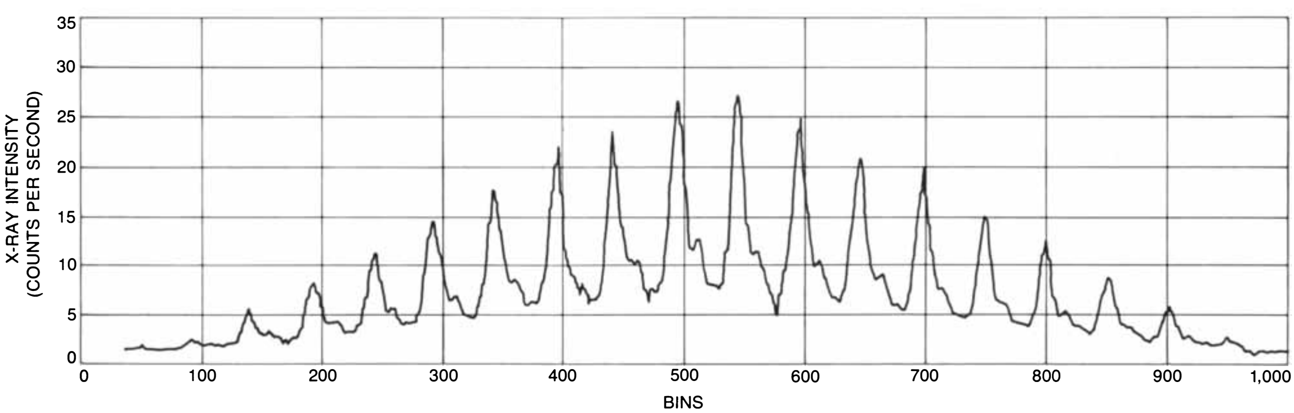
CHARACTERISTICS of the eight known binary X-ray sources are listed in this table. The fact that five of the eight double-star systems incorporate a massive blue supergiant star is considered remarkable, since in general such stars are very rare in our galaxy.

amined with *Uhuru*. This object is the second or third brightest X-ray source in the sky and was the first found to be variable in intensity. Initially it was hoped that an improved position would lead to an identification. It was also hoped that more could be learned about the intensity variation. What was discovered was that the intensity varied on as short a time scale as could be recorded with the *Uhuru* instrumentation: about a tenth of a second. The variation was then studied in more detail with rocket-borne detectors. These measurements indicated that the X rays were coming from an extremely small stellar object, a finding that provided a strong impetus to intensify the study of Cygnus X-1.

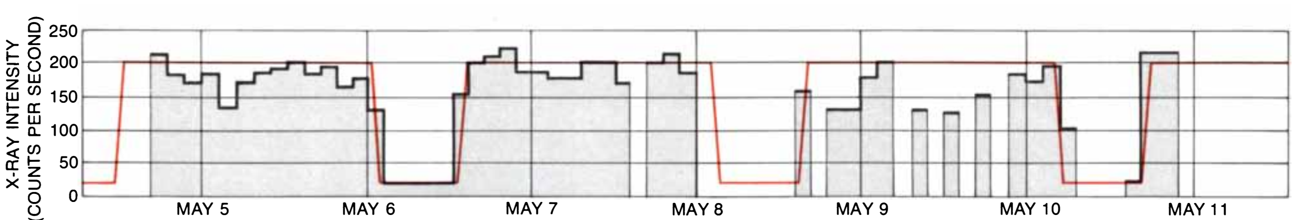
In 1971 Luc Braes and George Miley of the University of Leiden, working with the Westerbork radio telescope, discovered a faint radio source within the region of the sky known to contain the Cygnus X-1 X-ray source. Since no radio emission had previously been detected from this point in the sky, it meant that the radio source was variable. The observation indicated the presence of a peculiar object, possibly the X-ray source. Meanwhile a bright, hot supergiant star, very similar in type to the companion of Centaurus X-3, was found to be coincident with the radio source. After some time Harvey D. Tananbaum assembled a year's duration of *Uhuru* X-ray observations of Cygnus X-1, which, together with the radio data, revealed evidence of a simultaneous abrupt change in the emission of both X rays and radio waves [*see top illustration on page 30*]. The X-ray intensity was observed to decrease by a large factor at exactly the same time that the radio source first appeared. Radio waves and X rays have been emitted at a steady rate since then. This remarkable transition, which is still not explained, indicates that the X-ray source, the radio source and the bright star are one and the same object.

Meanwhile it was found by Louise Webster and Paul Murdin of the Royal Greenwich Observatory that the normal star associated with Cygnus X-1 is a single-line spectroscopic binary with a period of 5.6 days. The term spectroscopic here means that the presence of two stars is indicated by a periodic Doppler shift of the spectral lines of at least one of the stars as they revolve around a common center of gravity. In a single-line system the spectrum of only one component is visible. The supergiant therefore has an unseen companion, presumably the X-ray source. The orbital velocity of the supergiant is quite high; since its mass is greater than 15 times the mass of the sun, the companion star, which is able to swing the supergiant around with such a high velocity, must have a mass of at least four solar masses. (The best present estimates indicate that its mass is probably larger than eight solar masses.)

Remember that the rapid variations in the X-ray emission require a compact star. Remo J. Ruffini of Princeton University has demonstrated on very general theoretical grounds, however, that the mass of a white dwarf or of a neutron star cannot exceed about three solar masses. Beyond that limit gravitational forces are expected to exceed the internal pressure of the matter and the star would collapse. Since we know of no state of matter more compact than the neutron gas in a neutron star, the collapse presumably carries the star within its Schwarzschild limit; the object disappears from view and becomes a black

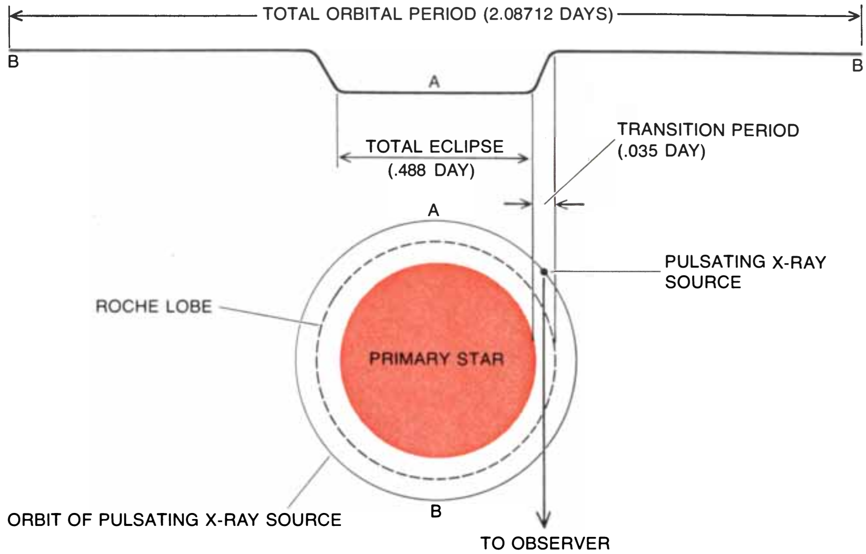


X-RAY EMISSION FROM CENTAURUS X-3 is characterized by short, regular pulsations with a mean period of only 4.8 seconds, as shown in this curve drawn from data recorded by the *Uhuru* satellite on May 7, 1971. The gradual variation in the height of the peaks from left to right is an effect of the satellite's rotation. A weak "interpulse" (*smaller peaks*) is evident between the main pulses.



EVIDENCE OF ECLIPSES in the Centaurus X-3 double-star system is revealed in this graph, in which the measured X-ray intensity (*gray bars*) is averaged over intervals of a tenth of a day and corrected to eliminate the effect of *Uhuru*'s rotation. The colored curve shows the average X-ray intensity predicted for Centaurus X-3 assuming that it is an eclipsing binary system. On the basis of these and other data one can deduce that the X-ray-emitting source in Centaurus X-3 is moving in an almost perfectly circular orbit at a velocity of some 415 kilometers per second; the orbital period is a little more than two days and the eclipse lasts about half a day.

hole. Thus Cygnus X-1 provides evidence not only for the binary nature of X-ray sources but also for the association of X-ray sources with compact stars and for the existence of black holes.



MODEL OF CENTAURUS X-3 is based on an analysis of the variations in X-ray intensity measured by *Uhuru*. The model assumes that the source of the X rays is a compact object orbiting a much larger central star. The numbers summarize the measured X-ray variations.

During this extraordinarily productive time additional X-ray binaries have been discovered and their companion stars have been identified. William Liller of the Harvard College Observatory made most of the optical identifications. (Recently Liller and his associates were able to show that the optical variation of Scorpius X-1 may be regular, with a period of .78 day. If this conclusion is confirmed, it would provide the long-awaited proof of the binary nature of the object.) The other regularly pulsing X-ray source, Hercules X-1, shows Doppler variations in its emission period and X-ray eclipses. Here the variations of both the X-ray source and its optical companion are correlated over a wide range, a finding that is not yet fully understood.

At least four other sources (Vela X-1, 2U 1700–37, Circinus X-1 and SMC X-1) show eclipselike variations in intensity. None of these sources emits pulses of X rays, so that it is not possible to observe the motion of the X-ray source. The optical counterpart of Circinus X-1 has not yet been found. The optical components of the other three sources are all blue supergiant stars quite similar to the companions of Cygnus X-1 and Centaurus X-3. Moreover, one of these objects, SMC X-1, is in a galaxy close to our own, the Small Magellanic Cloud, which means that its distance is known on independent grounds. This coincidence indicates beyond any doubt that the optical component is a very luminous blue supergiant star.

Finally, the source Cygnus X-3 shows periodic variations that may represent an eclipse. The period is only 4.8 hours, which is not particularly short for double stars in general but is much shorter than the period of any of the other X-ray binaries. Because of its short period, and also because of its enormous radio outbursts, Cygnus X-3 may represent a class of its own.

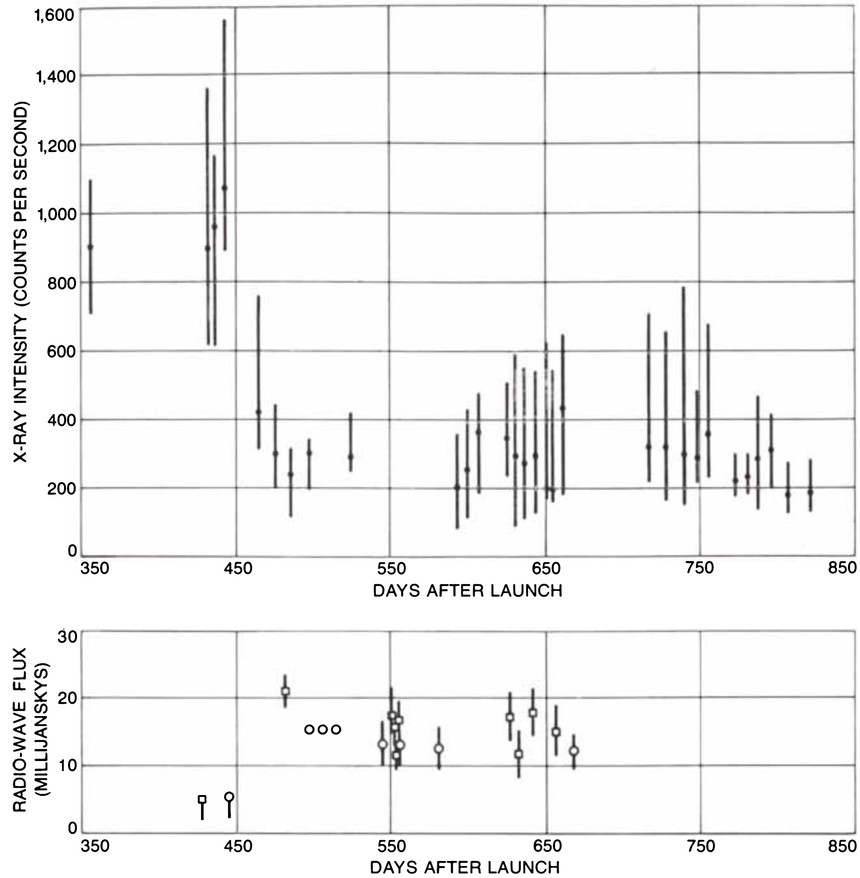
It is remarkable that in five of the eight known X-ray binaries the companion is a massive blue supergiant star. This cannot be an accident, because such early-type blue supergiants are rare in our galaxy. Their total number is estimated to be about 1,000, out of some 100 billion stars. Thus the probability of even one X-ray source being randomly associated with one of these stars is very small. Massive stars of this type are confined to the central plane of the galaxy, in which the spiral arms lie. Indeed, the massive X-ray binaries are all located close to the central plane. From the fact that it is known that three of these systems (Cygnus X-1, Vela X-1 and 2U 1700–37) are located less than 10,000 light-years away one can estimate the total number of such systems in the galactic disk to be about 50. This means that roughly 5 percent of all early-type supergiants in our galaxy are in massive X-ray binaries.

This extraordinarily high percentage is on the same order of magnitude as the percentage of close binaries among the normal massive stars in the spiral arms of our galaxy. It is known that among all stars more massive than about 15 solar masses the blue supergiants represent the evolutionary stage immediately following the star's exhaustion of its hydrogen fuel and its switching over to another nuclear fuel. Therefore the high percentage of X-ray binaries among the blue supergiants can mean only one thing: the supergiant X-ray-binary stage must be a normal one in the evolution of many of the massive close binaries. How can this be explained? And how were the compact stars in these systems formed?

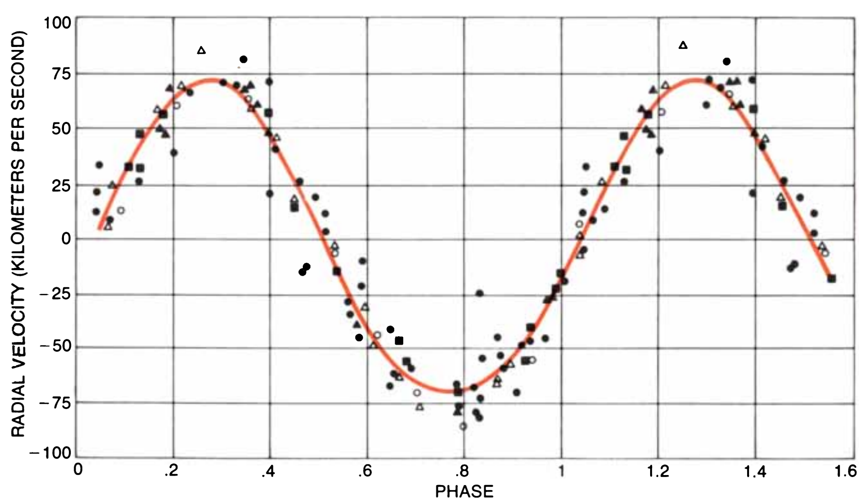
The Crab Nebula is the remnant of a supernova explosion. From the presence of a pulsar in the center of the nebula (and also in the Vela supernova remnant) we know that a neutron star is formed during a supernova event. According to the generally accepted view, such an event represents the final gravitational collapse of the core of a star more massive than about four solar masses that has exhausted its nuclear fuel, accompanied by the simultaneous ejection of the star's cooler outer layers. In the case of very massive stars (those that are more massive than about 10 solar masses) the collapsing core is expected to have a mass of more than three solar masses. Such a mass is too large for a stable neutron star; hence these massive supernovas are expected to give rise to black holes.

The presence of neutron stars (and possibly of black holes) in the X-ray binaries therefore forces one to the conclusion that in these systems a supernova explosion must have taken place. When did the star explode, and why was the entire system not disrupted by such a violent event?

Fairly accurate answers to these questions can now be given. Much progress has been made during the past decade in our understanding of the evolution of close binary systems, largely as a result of the work of such theorists as Rudolf Kippenhahn and Alfred Weigert in Germany, Bogdan Paczyński in Poland and Miroslav Plavec in Czechoslovakia. We now know, for example, that during the evolution of a close binary large amounts of matter can be transferred from one star to the other. This exchange of mass is a consequence of the fact that a component of a close binary has only a



SIMULTANEOUS ABRUPT CHANGE in the emission of both X rays and radio waves from Cygnus X-1 was found to have taken place in March, 1971. The *Uhuru* data show that at that time the intensity of "soft" (low energy) X rays from the source decreased suddenly by a factor of three; it has remained at a lower level since then (*graph at top*). At exactly the same time weak radio waves, measured at two different frequencies, were first detected from roughly the same position in the sky; they have been observed at a steady rate ever since (*bottom graph*). The position of the radio source could be very accurately determined; this in turn led to the discovery of the visible counterpart of the Cygnus X-1 X-ray source. The vertical bars indicate the range of variability observed in the received signals.



RADIAL-VELOCITY CURVE of the visible component of the Cygnus X-1 binary system, plotted by workers at the Royal Greenwich Observatory, shows that the radial velocity of this blue supergiant star (designated HDE 226868) varies regularly by about 70 kilometers per second in a period of 5.6 days. These figures imply that the X-ray source has a mass more than four times that of the sun, which in turn suggests that the source is a black hole rather than a neutron star. Symbols denote observations by different groups of investigators.

limited space available for its evolution, owing to the presence of its companion. This pear-shaped space is called the Roche lobe (after the 19th-century French mathematician Edouard Albert Roche, who first recognized its importance). In the evolution of a typical binary, as soon as one of the components expands to a volume larger than that of its Roche lobe the matter outside the lobe will begin to flow toward the companion star. Such an overflow will occur at some time in any close binary system, since all normal stars are known to evolve eventually into red giants. The evolutionary change in radius of an isolated normal star of 20 solar masses, for example, has been computed by Camiel de Loore and Jean-Pierre de Greve of the Flemish Free University of Brussels [*see top illustration on opposite page*]. This picture is characteristic of stars more massive than about 12 solar masses.

During the prolonged stage of hydrogen "burning" the radius of a star remains very small. The convective core in which the hydrogen nuclei fuse to form helium nuclei represents between 15 and 40 percent of the stellar mass, depending on the total mass [*see bottom illustration on opposite page*]. The larger the mass of a star, the shorter its lifetime. In massive stars the higher pressure and temperature in the core drive up the rate of nuclear fusion, which shortens the lifetime of the star.

When all the hydrogen in the core is consumed, the core, which is now largely helium, begins a stage of rapid contraction and heating, accompanied by a rapid expansion of the cooler, hydrogen-rich stellar envelope. The star then becomes a supergiant with a radius of up to several hundred times the radius of the sun. In this supergiant stage helium nuclei fuse to form carbon in the star's core.

The majority of the massive close binary systems have orbital periods of between three and 20 days. In these systems the components will overflow their Roche lobes during the stage of rapid expansion of the envelope. The more massive component will evolve faster and hence will be the first to overflow.

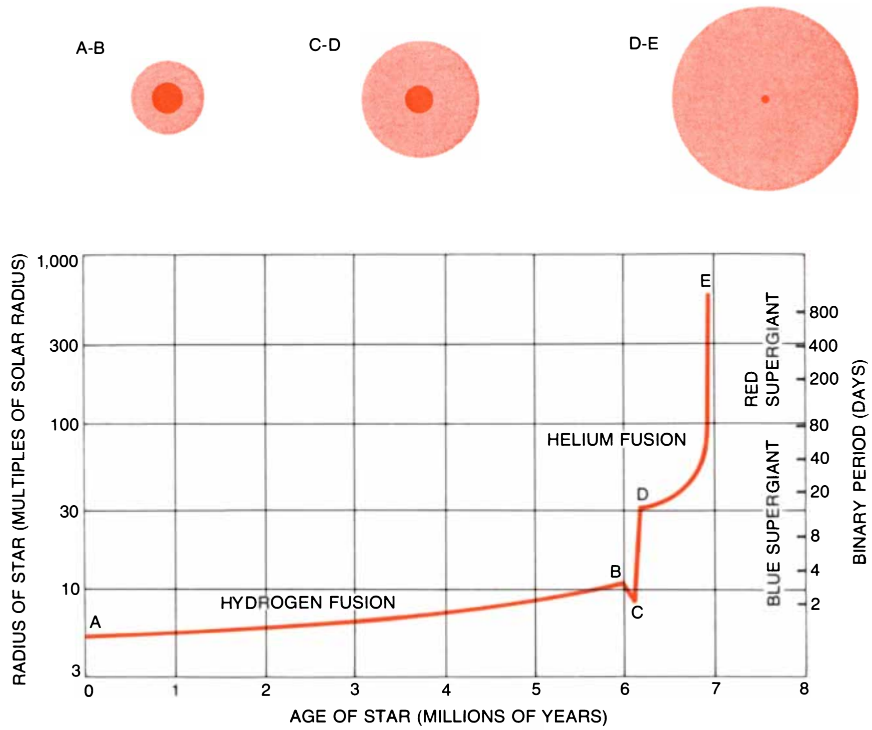
The computations of Paczyński and others show that the ensuing loss of mass tends to accelerate the rate of expansion of the envelope of the mass-losing component. Because of this effect the loss of mass does not end until almost the entire hydrogen-rich envelope is lost to the companion, leaving behind only the helium core. The resulting helium star then begins helium fusion, and its radius subsequently shrinks to only a few times the radius of the sun. Such helium stars are very hot and luminous, and they resemble the Wolf-Rayet stars, which are found in many close binaries.

The evolution of a system of this type, leading finally to the formation of a massive X-ray binary, has been worked out in some detail by de Loore and de Greve on the basis of a hypothetical model proposed by one of the authors (van den Heuvel) and John Heise of the University of Utrecht [*see illustration on next page*]. From this picture the following evolutionary stages emerge. About 6.16 million years after the birth of the system the 20-solar-mass primary star overflows its Roche lobe and transfers (in only 18,000 years) 14.6 solar masses to its six-solar-mass companion, which becomes a 20.6-solar-mass star. The computations show that the latter star, which had not yet finished its hydrogen-burning stage when the mass exchange began, will continue hydrogen fusion for another six million years or so after the mass exchange. The prolonged stage of hydrogen burning is attributable to the large amount of hydrogen-rich material received from its companion. This "fuel injection" rejuvenates the binary component on the receiving end and causes it in effect to "live twice."

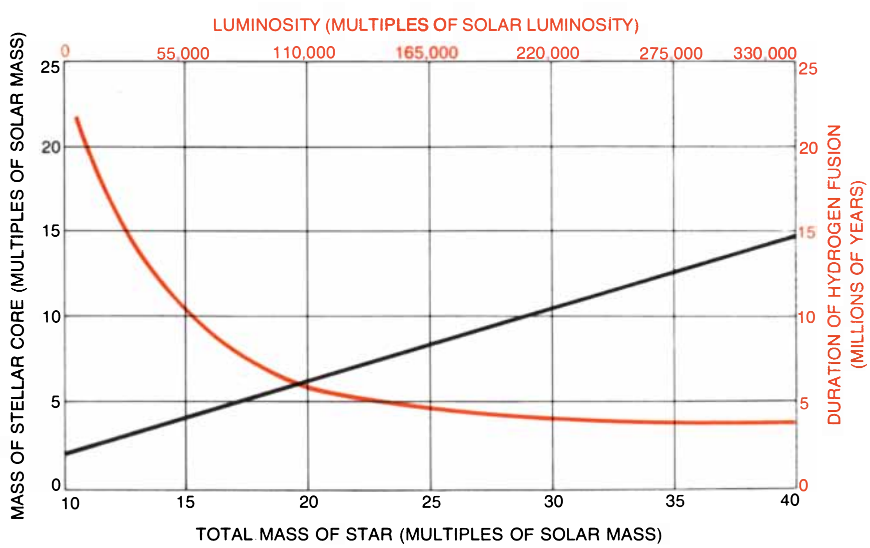
After the mass exchange the orbital period has changed to 5.2 days. This is a consequence of the law of the conservation of angular momentum, which requires that if no mass is lost from the system, its total angular momentum cannot change.

How will the system evolve further? Helium stars evolve very rapidly, since they have only a relatively small amount of nuclear fuel available to supply their enormous luminosities, which are equal to those of hydrogen-rich stars four to five times more massive. The 5.4-solar-mass helium star finishes burning its helium and carbon within about 600,000 years after the mass exchange. During this time its radius remains much smaller than that of its Roche lobe; computations by W. David Arnett of the University of Illinois show that the same will be true during the subsequent few thousand years, as the stellar core evolves through the burning of neon, oxygen and silicon toward its final collapse. Therefore the helium star will not lose any more mass, and it will continue to evolve as if it were an isolated helium star.

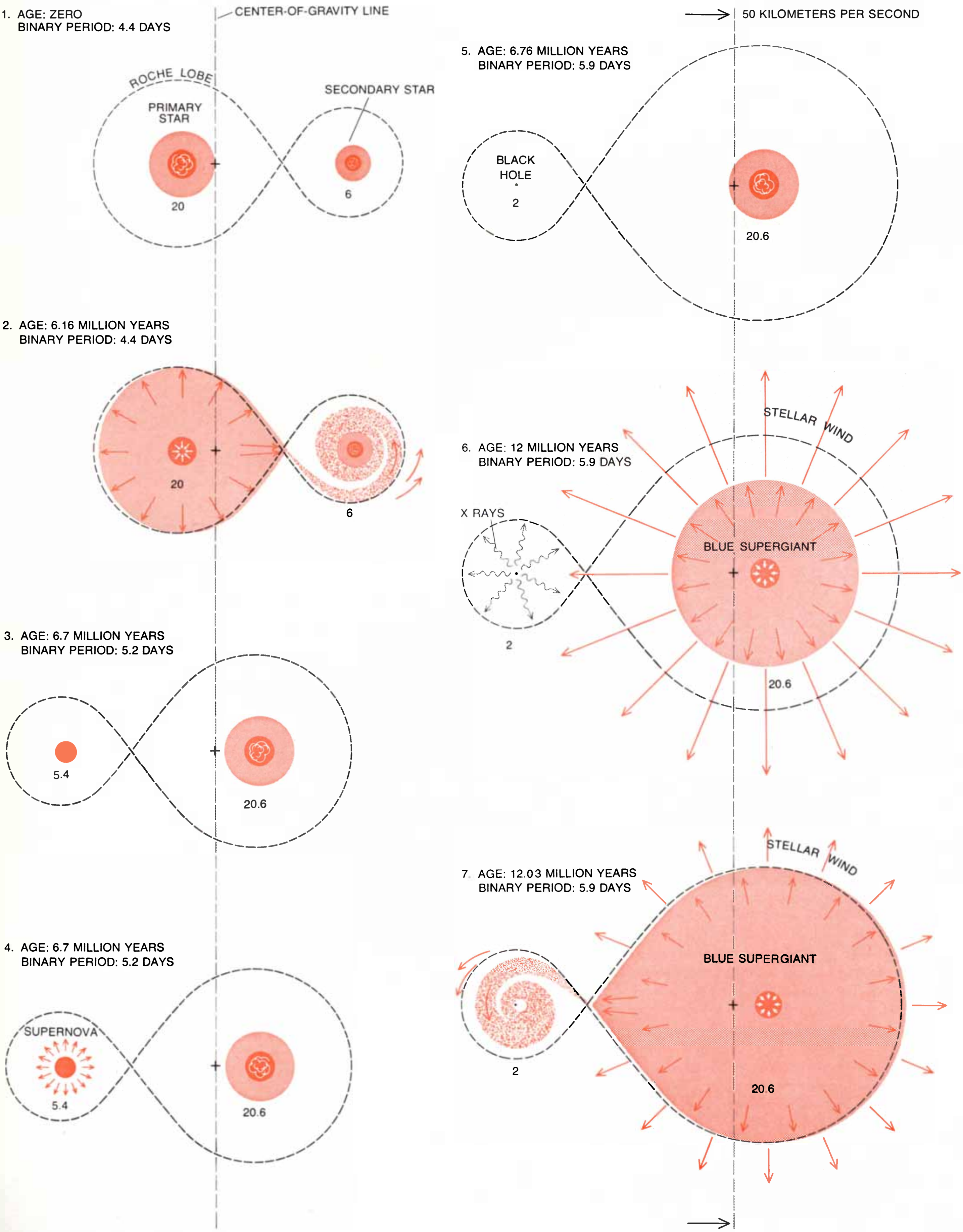
This type of evolution holds for all helium stars more massive than four or five solar masses. Such computations suggest that the core of such a star will finally collapse into a neutron star or a



EVOLUTION OF NORMAL STAR with a mass 20 times that of the sun is depicted here on the basis of computations by Camiel de Loore and Jean-Pierre de Greve of the Flemish Free University of Brussels. At its birth the star consists of 70 percent hydrogen, 27 percent helium and 3 percent heavier elements. Between the points *A* and *B* hydrogen nuclei fuse to form helium in the core of the star, where the temperature is about 35 million degrees K. During this hydrogen-fusion stage the radius of the star increases very slowly. By point *B* all the hydrogen in the core has been consumed. The subsequent contraction and further heating of the core cause the stellar envelope, which is still rich in hydrogen, to expand (*C–D*). At point *D* the temperature in the core reaches some 100 million degrees K., and helium nuclei begin to fuse to form carbon. From this point on helium fusion becomes the main energy source of the star, which thereafter passes through several giant phases (*D–E*). In most close binary systems the normal star will overflow its Roche lobe during the stage of rapid expansion of the envelope, that is, between points *C* and *D*. The scale at right indicates the period of a binary system in which this normal star just fills its Roche lobe (assuming that its companion star has a mass equal to approximately six times the sun's mass).



MASS OF HELIUM CORE contains an increasing fraction of the total mass of the star, as is depicted here for stars with masses of between 10 and 40 times the sun's mass. Since in a close binary system the normal component, after completing its hydrogen-fusion stage, transfers almost all its hydrogen-rich envelope to its compact companion, its total mass after the mass-transfer stage is over becomes approximately equal to the mass of its helium core.


1. AGE: ZERO
BINARY PERIOD: 4.4 DAYS
CENTER-OF-GRAVITY LINE
ROCHE LOBE
PRIMARY STAR
SECONDARY STAR
20
6
2. AGE: 6.16 MILLION YEARS
BINARY PERIOD: 4.4 DAYS
20
6
3. AGE: 6.7 MILLION YEARS
BINARY PERIOD: 5.2 DAYS
5.4
20.6
4. AGE: 6.7 MILLION YEARS
BINARY PERIOD: 5.2 DAYS
SUPERNOVA
5.4
20.6
50 KILOMETERS PER SECOND
5. AGE: 6.76 MILLION YEARS
BINARY PERIOD: 5.9 DAYS
BLACK HOLE
2
20.6
6. AGE: 12 MILLION YEARS
BINARY PERIOD: 5.9 DAYS
STELLAR WIND
X RAYS
BLUE SUPERGIANT
2
20.6
7. AGE: 12.03 MILLION YEARS
BINARY PERIOD: 5.9 DAYS
STELLAR WIND
BLUE SUPERGIANT
2
20.6

black hole, giving rise to a supernova. Because a four-solar-mass helium star in a binary results by mass loss from a 15-solar-mass normal primary star, this type of evolution (toward a supernova) is expected to occur preferentially in the most massive close binaries. In the less massive systems the helium star is likely to continue losing mass and might finish as a white dwarf with a radius much larger than that of a neutron star.

How does the supernova explosion affect the binary system? On the basis of simple celestial mechanics one finds that when the component with the smaller mass loses mass explosively, the effects of the mass loss itself can never disrupt the system. The only effect of the mass loss is that the orbit of the exploded star will become elliptical, and the motion of the center of gravity of the system will be accelerated to a velocity of several dozen kilometers per second. This result of the mass loss is called the sling effect [*see illustration on next two pages*]. In this way a "runaway" binary system can be formed, a system that is moving away from its place of birth with a velocity of dozens of kilometers per second. Other factors influencing the shape of the orbit are the impact of the supernova shell on the unexploded component (which may be considerably accelerated) and possible asymmetries in the explosion. In the case of a helium star exploding in a close binary neither of these two effects appears capable of disrupting the system. Therefore it would seem that most of these systems will remain bound after the explosion.

For a system in which the mass of the collapsed remnant is assumed to be two solar masses the effects of mass loss and shell impact cause the orbital period to increase suddenly from 5.2 to 5.9 days and accelerate the center of gravity to a velocity of some 50 kilometers per second. The eccentricity of the orbit after the explosion is quite pronounced; the strong tidal forces in such a close system are likely to make the orbit circular again within a few million years.

The evolution of the system after the explosion would be expected to proceed as follows. Let us assume that the remnant is a black hole. Remember that after the mass exchange the companion needs six million years to finish burning its hydrogen. Accordingly it will not expand outside its Roche lobe within the first 5.4 million years after the explosion of its companion, and there will be no mass exchange during that time. If the remnant is a neutron star, it would—if it were single—be observable as a rapidly pulsating young pulsar for perhaps the first 10,000 years following the explosion. Later it would still be observable for several million years as a slower and weaker radio pulsar. The tenuous outer atmosphere of the companion star, however, is likely to inhibit all radio emission from a pulsar in a binary system. Hence a binary pulsar of this type is probably undetectable with conventional radio-search techniques. During the first 50,000 to 100,000 years after the explosion an expanding supernova shell may be visible around the system. Afterward, for the next 5.3 million years, the system will appear as a normal, single-line spectroscopic binary with no clear peculiarities. This period is what we have called the "quiet-collapsar-binary stage."

When will the system become an X-ray source? At first one would think that this would happen when the companion has expanded to its Roche lobe and has begun to transfer matter to the collapsed star. It turns out, however, that the rate of mass transfer in this case (about a thousandth of a solar mass per year) will be too large to lead to the formation of an X-ray source, since a thick envelope, opaque to X rays, will form around the collapsed star. Such an envelope will radiate mainly in the visible and ultraviolet regions of the spectrum, and the object will appear to us as a fairly normal star.

The formation of an X-ray source requires, on the one hand, that the rate of mass transfer to the collapsed star not be very large, otherwise the envelope around the star will not remain transparent to X rays. On the other hand, the accretion rate cannot be too small either, otherwise the X-ray source will be too faint to be detectable.

It appears that both of these conditions are fulfilled only during the brief stage between the termination of hydrogen burning in the core of the companion star and the moment when the envelope of the star has expanded to its Roche lobe. During this short interval, lasting roughly between 20,000 and 50,000 years, a star with a mass larger than about 15 or 20 solar masses becomes a blue supergiant. Stars of that type have radii about 20 to 30 times the radius of the sun and are observed to be losing mass from the outer parts of their atmosphere with velocities on the order of 1,000 kilometers per second. This type of mass loss is called a stellar wind. The wind is thought to be driven by the large radiation pressure exerted in the outer layers of these very luminous stars.

The observed rates of mass loss from blue supergiants are on the order of a millionth of a solar mass per year. The compact star in its orbit plows through this outflowing wind and captures only a tiny part of its matter: about .01 or .1 percent. Owing to the high velocity of the wind, only the part of the matter that passes the compact star at very small distances will be captured in the star's gravitational field. The bulk of the stellar-wind flux extends out into space without being much affected by the presence of the compact companion star [*see top part of illustration on page 25*].

The resulting accretion rate onto the compact star, a billionth of a solar mass per year, is enough to give rise to an X-ray source with a strength on the order of 10,000 solar luminosities. At the same time the density of the stellar wind

EVOLUTION OF A CLOSE BINARY SYSTEM leading to the formation of an intense X-ray source is presented in the sequence of drawings on the opposite page. The detailed computations were worked out by de Loore and de Greve on the basis of a model proposed by one of the authors (van den Heuvel) and John Heise of the University of Utrecht. In step *1* the system begins as a normal close binary, the primary and secondary components of which have masses equal to 20 and six solar masses respectively. Hydrogen fusion takes place in the cores of both stars, producing helium. In step *2* the primary star has finished hydrogen fusion and overflows its Roche lobe, triggering the first stage of mass exchange. In step *3* the first stage of mass exchange ends and helium fusion begins in the remaining 5.4-solar-mass core of the primary star, producing carbon. In step *4* the primary star explodes to form a supernova. In step *5*, a short time later, the supernova shell has left the system, leaving behind a collapsed remnant, presumably a neutron star or a black hole (in this case a two-solar-mass black hole). As a result of the explosion the binary period has increased and the entire system has been accelerated to a velocity of some 50 kilometers per second in the direction of the black arrow (*see illustration on next two pages*). In step *6* the secondary star has finished its hydrogen-fusion stage and has become a blue supergiant, which begins to lose mass at a moderate rate in the form of a stellar wind. The capture of stellar-wind material by the collapsed companion turns it into a strong emitter of X rays. In step *7* the second major stage of mass exchange begins, extinguishing the X-ray source.

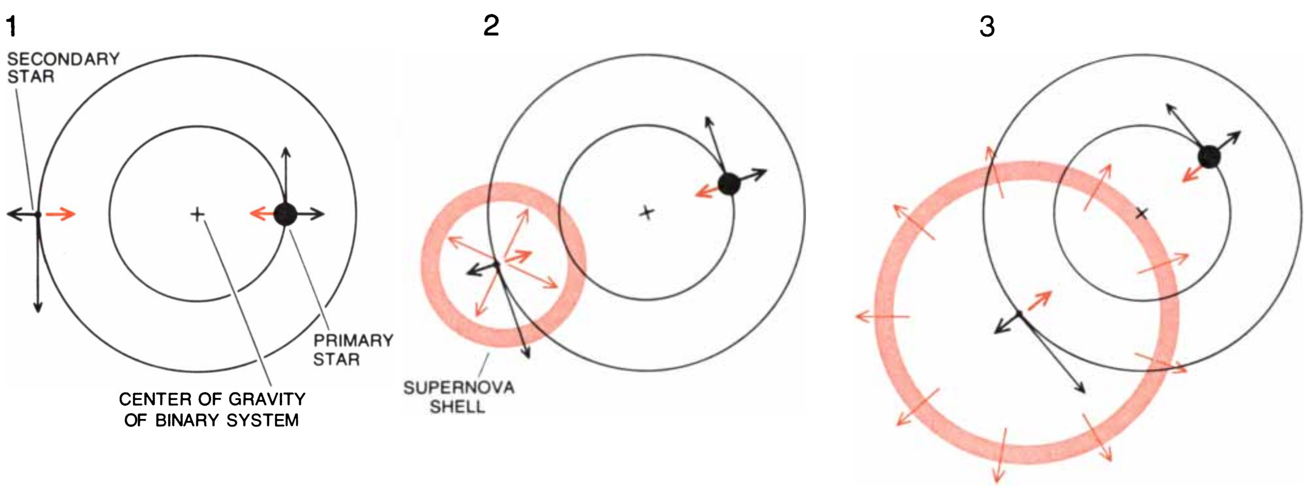
is low enough to ensure that the wind is transparent to X rays.

These facts immediately explain why the massive X-ray binaries almost always include blue supergiant stars. Unevolved massive stars do not have a sufficiently strong stellar wind, whereas stars that have evolved beyond the blue-supergiant stage will have begun to overflow their Roche lobe and therefore will have extinguished any X-ray emission from a compact companion by the ensuing large rates of mass loss. Stars less massive than about 15 or 20 solar masses do not evolve into blue supergiants and therefore will never possess stellar winds that can turn a possible collapsed companion star into an X-ray source.

The evolutionary picture we have outlined here fits many other observed aspects of the massive X-ray binaries quite well. For instance, it explains why the X-ray-binary stage should be a normal one in the evolution of almost all massive close double stars. Furthermore, it predicts that the neutron stars (or black holes) in the massive X-ray binaries are comparatively old, having originated in supernova explosions several million years ago. Thus it explains the very long pulse period of Centaurus X-3 (longer than that of any known radio pulsar), since the pulse periods of all pulsars are known to increase gradually with time. It also explains the complete absence of an expanding supernova shell around any of these systems, since such shells must have disappeared several million years ago. Finally, the relatively short duration of the X-ray-binary stage (on the order of .2 to .5 percent of the total lifetime of a massive close binary) explains the rarity of these objects. It means that on the average only one in every 200 to 500 massive close binaries will be in the X-ray-binary stage (assuming that the formation of massive stars in the galaxy is a continuous process). The total number of close binaries in the galaxy that are more massive than about 15 solar masses (both evolved and unevolved) is estimated to be about 6,000 (with an uncertainty of about a factor of two); hence one would expect the total galactic number of massive X-ray binaries to be roughly a few dozen. This number is, within the boundaries of the uncertainties we have mentioned, in reasonable agreement with the observational estimate of about 50 such systems for the entire galaxy.

These facts appear to add up to the overall plausibility of the outlined evolutionary picture for the massive X-ray binaries. Adopting this picture, therefore, one finds that it has the remarkable consequence that there must be at least several thousand massive quiet collapsar binaries in the galaxy, since the duration of the quiet-collapsar-binary stage (between the supernova explosion of the helium star and the onset of the X-ray-emitting stage) is about 40 percent of the total lifetime of a massive close binary. This figure means that some 40 percent of all massive close binaries can be expected to be in that stage, which implies that there are some 2,400 such systems in the galaxy.

The normal components of such quiet systems are very luminous, hot stars; thus on a statistical basis one would expect some 20 of these systems to be present among the stars visible to the unaided eye. Is there any possibility of recognizing such systems? Only one of the six known massive X-ray binaries is a regular X-ray pulsar, and five of the six are of the same type as Cygnus X-1. Hence one might expect that most of the massive quiet collapsar binaries incorporate a black hole. These systems will appear to us as normal single-line spectroscopic binaries, making it difficult to distinguish them from systems in which the less luminous component is a normal star. The same probably holds for systems in which the collapsed star is a neutron star, since such stars are not expected to emit detectable radio waves. Thus there seems at present to be little hope that a quiet collapsar binary can be distinguished from a normal single-line spectroscopic binary.

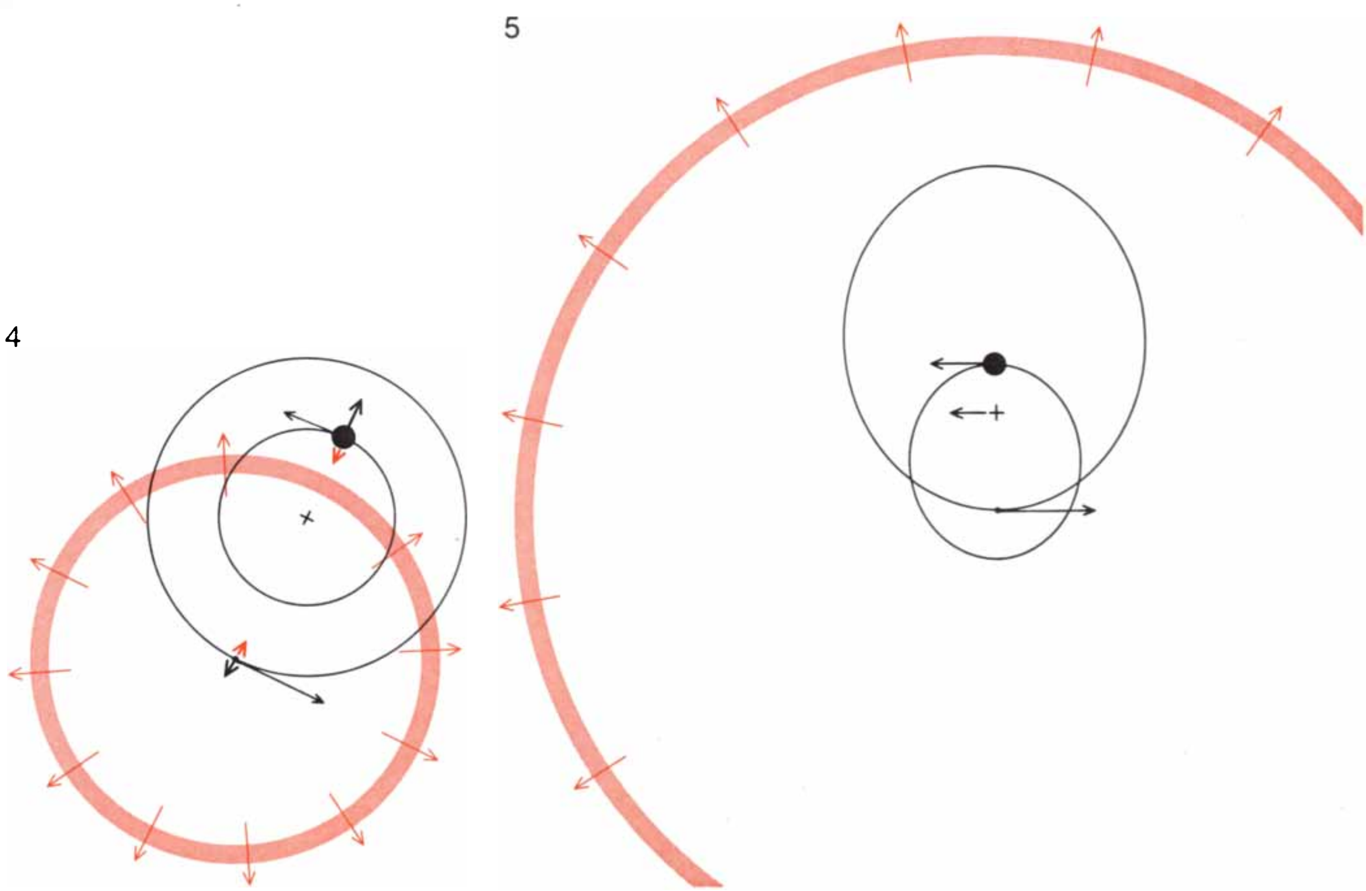


EFFECTS OF SUPERNOVA EXPLOSION on the close binary system depicted on page 32 is explained here from a somewhat different point of view. In this sequence the heavy colored arrows denote the centripetal forces of gravity between the two stars; the heavy black arrows denote the centrifugal forces resulting from the orbital motion of the two stars around a common center of gravity; the light black arrows indicate motion. In the initial stable situation (*1*) both components move in circular orbits and the centrifugal forces are exactly balanced by the gravitational ones. Then one of the components explodes (*2*). As long as the ejected supernova shell has not passed the unexploded component (*3*) that star does not "feel" that its companion has exploded. After the shell passes the unexploded star, however, the gravitational attraction between the two stars is reduced (*4*); moreover, the centrifugal force

How would a massive X-ray binary evolve after the supergiant star has overflowed its Roche lobe and extinguished the X-ray source? Computations carried out at Utrecht and Brussels suggest that at this stage the binary period would decrease rapidly, owing to the fact that the compact star would be unable to accept quickly most of the matter transferred to it and hence would strew the matter about the system, causing the system to lose a great deal of its angular momentum. As a result the system, which in its final stage would consist of a helium star together with the compact star, would have a binary period of only a few hours.

In such a system the helium star would finally explode. Since in most cases the helium star would be more massive than its compact companion, the explosion would probably disrupt the system, leading to the formation of two runaway compact stars, one old and one young. In a few systems, however, the two remaining compact stars would remain bound; such a system would have a highly eccentric orbit and a short binary period. The binary radio pulsar designated PSR 1913+16 may be a system of that type. This unusual system, discovered last fall by Joseph H. Taylor, Jr., and Russell A. Hulse of the University of Massachusetts, seems to consist of two compact stars, one of which is a very young, rapidly pulsating radio pulsar. The system appears to have a highly eccentric orbit, indicating that it must have been nearly disrupted during the supernova explosion that formed the pulsar. The striking resemblance of this system to the anticipated remnant of a massive X-ray binary suggests an evolutionary relation between the two types of system.

In short, the discovery of the X-ray binaries has opened up a rich new field of astronomy. It has produced the first clear evidence for the possible existence of black holes, pieces of matter that may have returned to a condition similar to the primordial state that prevailed before the "big bang" created the universe as we know it. It has shown us that most of the massive close binaries survive the supernova explosion of their more evolved component. Perhaps even more important, it has shown us that there are stars whose enormous energy production results from the simple process of matter falling in an extremely strong gravitational field. That process, in which as much as 40 percent of the rest energy of the matter can be radiated away, is some two orders of magnitude more efficient than the process of nuclear fusion that powers the normal stars. The demonstrated existence of this powerful energy-generating mechanism might be helpful in gaining an understanding of energy production in other abnormally luminous, compact objects, such as quasars and the nuclei of certain galaxies.



acting on the mass-losing component is also lessened. Since the centrifugal force acting on the unexploded component remains the same, this force is no longer balanced by the gravitational attraction of its companion. As a result the unbalanced centrifugal force pulls the unexploded star away. This has two effects: the two stars move somewhat farther apart, and the system as a whole begins to move (5). The first effect causes the orbits to become wider and more elliptical. The second effect causes the system to become a "runaway" star, that is, a star that is moving from its original position (or state of motion) with a velocity of as much as 50 kilometers per second. The actual impact of the supernova shell on the unexploded star may enhance both of these effects, but in cases such as this one, where the exploding star is less massive than its companion, the impact is unlikely to disrupt the binary system.

INTERACTIVE HUMAN COMMUNICATION

Studies of how people communicate in solving problems assist progress toward the development of a conversational computer, with which the user could communicate as if it were a person

by Alphonse Chapanis

Modern computers touch the life of every citizen in varied and often unexpected ways. Not only do computers prepare our utility bills, credit-card bills and bank statements but also they control our traffic, assist us in making travel and theater reservations, keep tabs on the weather for us and help to diagnose our bodily ills. For all that, most of us still have little direct contact with computers. Most computers still require an intermediary between the ultimate user and the computer, someone who is familiar with the way the computer works and with the special language that is needed to address it.

A goal toward which many people have been working is the design and construction of conversational computers: computers that can interact with people in such familiar and humanlike ways that they require little or no special instruction. If such conversational computers are ever to come into existence, however, their designers and programmers will need to know more about how people interact in communicating with each other. With this rationale in mind my colleagues and I at Johns Hopkins University have been working to describe human communication in precise terms and to define its rules.

We have been concerned with three main questions. How do people naturally communicate with each other when they exchange factual information in the solution of problems? How is interactive human communication affected by the devices through which people converse? What other significant variables affect interactive communication?

Let me digress briefly to distinguish between unidirectional and interactive communication. For years psychologists have been concerned with the effectiveness of unidirectional modes of communication such as highway signs, books, lectures and television programs. In unidirectional communication the person to whom the message is addressed is a passive recipient of information. Nothing that he says or does affects the communicator, the communication process or the content of the message.

Interactive communication involves at least two participants. The content of any particular message is determined in part by the content of the prior messages from all participants and so cannot be predicted from the content of the message from any one of them. Conferences, arguments, seminars and telephone conversations are examples of interactive communication. This is the kind of communication that has been the focus of our investigation.

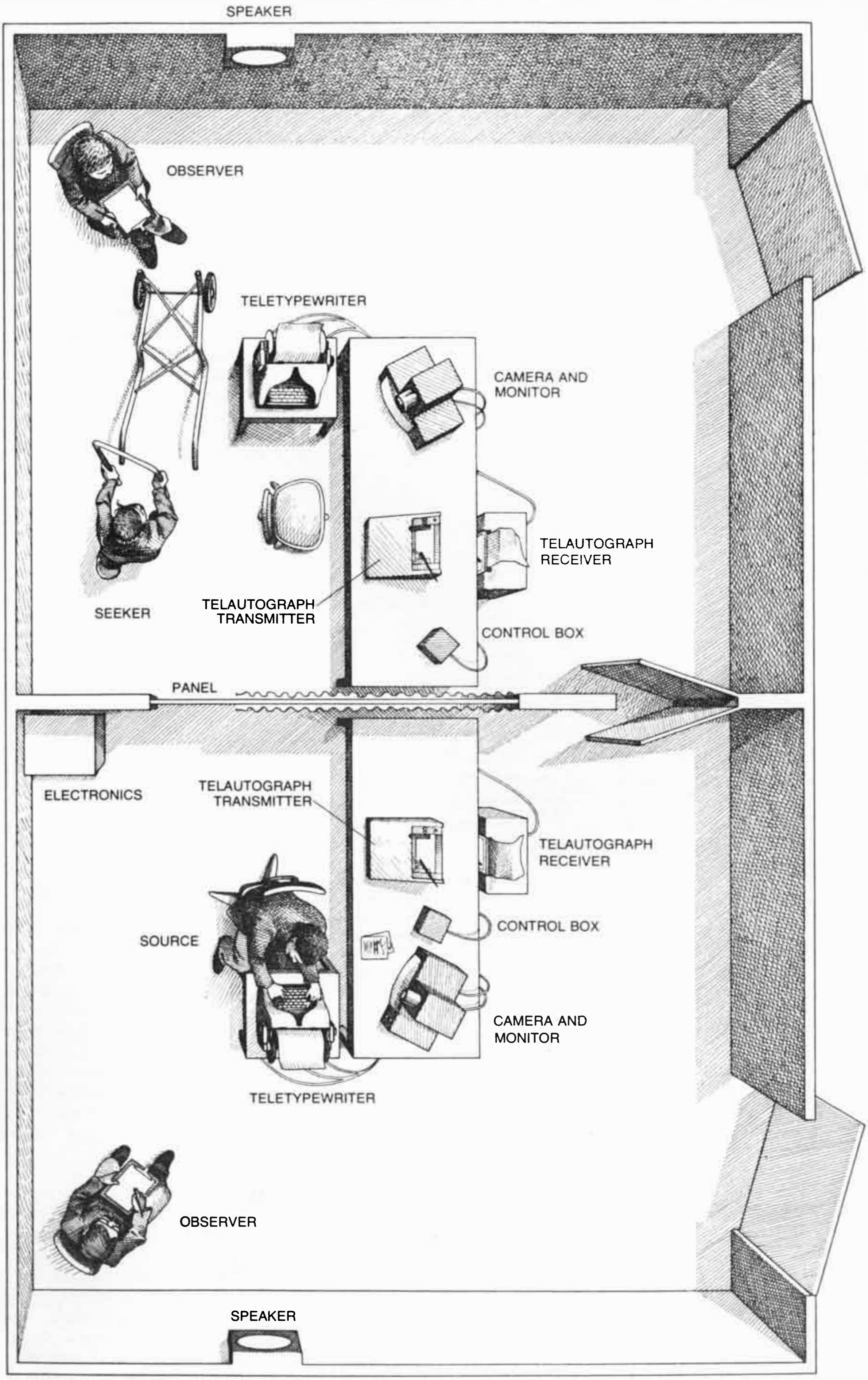
Our experiments are designed to model interactions between man and computer rather than to simulate any existing or planned interactive computer systems. We set up two-person teams and ask them to solve credible problems for which computer assistance has been or could be useful. The exchanges that result represent a limited class of conversations, to be sure, but it is an important class, and we have to start somewhere.

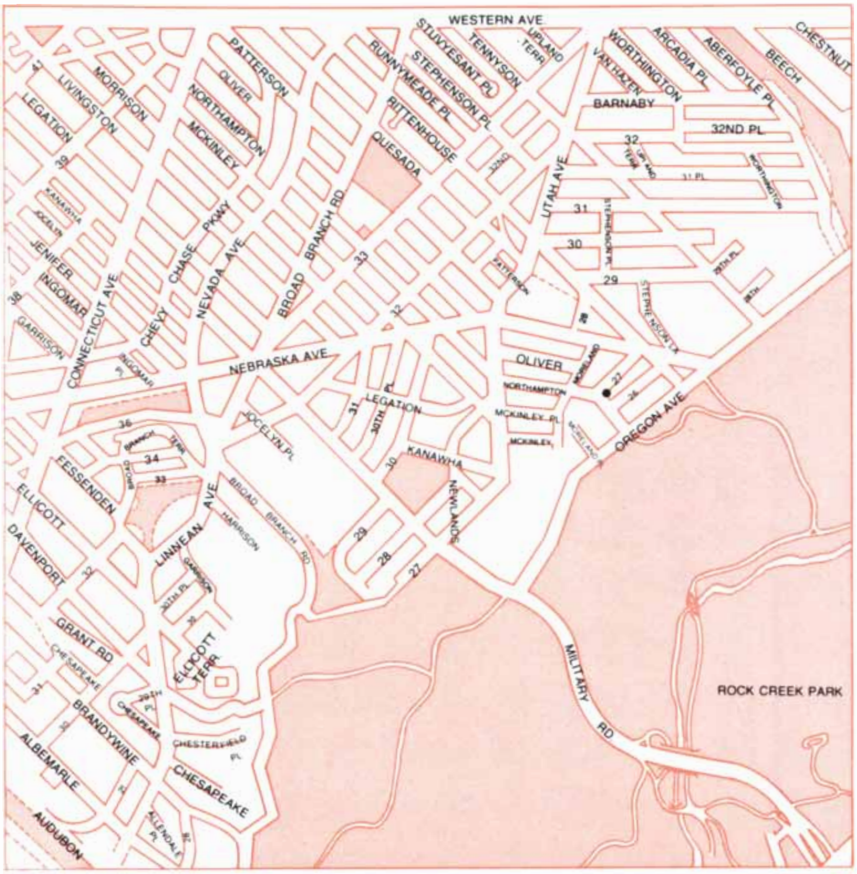
One of our primary interests is the channels and the modes through which people converse. Although the channels of communication that link man and computer are being broadened, most interactions of this kind involve a typewriter or a similar device. Our experiments examine four different channels: voice, handwriting, typewriting and video, the last being the picture part of television without the voice. Three of these four basic channels have been tested singly, and all of them have been tested in various communicative combinations that we call modes. We have tested as many as 10 different modes in a single experiment. As a standard of comparison we typically rely on normal, unrestricted, face-to-face communication, which we call a communication-rich mode.

When we have set up a team, we designate one member as the source (of information) and the other as the seeker. One can think of the source as an ideal computer, that is, a computer communicating in such a human way that a person who did not know he was dealing with a computer might readily believe he was communicating with another person. The seeker can be regarded as the user of the computer. To continue the analogy, our different channels and modes of communication model various input and output channels between the computer and its human user.

The setting for a typical experiment consists of two adjoining rooms connected by a soundproof double door [*see illustration on opposite page*]. The wall between the rooms also has in it a large double-glass panel, which can be covered with an opaque screen so that the source and the seeker cannot see each other. When the panel is not covered, the participants can see each other and can converse freely through a microphone and loudspeaker, but they are still separated physically. Some of our experiments also have test conditions in

LABORATORY SETTING for a typical experiment is depicted on the opposite page. The seeker has been given a trash-can carrier to assemble but has not been told its name or function. The source has the information for assembly. The experiment is designed to elicit communication in the hope of assisting in the design of a computer that would be analogous to the source in communicating much as a person communicates.


SPEAKER
OBSERVER
TELETYPEWRITER
CAMERA AND
MONITOR
TELAUTOGRAPH
RECEIVER
TELAUTOGRAPH
TRANSMITTER
SEEKER
CONTROL BOX
PANEL
ELECTRONICS
TELAUTOGRAPH
TRANSMITTER
TELAUTOGRAPH
RECEIVER
SOURCE
CONTROL BOX
CAMERA AND
MONITOR
TELETYPEWRITER
OBSERVER
SPEAKER



Earich La	I-25	Milmarson Pl	R-11	Tuckerman St	Q-9, R-10, R-12	41st St., S.E.	O-24
Earl Pl., N.E.	R-17	Mills Ave., N.E.	Q-17, R-17	Tulip St.	R-6	42nd Pl	M-6
East Pl	K-13	Millwood Lane	J-8	Tunlaw Road	K-10, 11	42nd Pl., N.E.	R-21
East Capitol St	M-18, Q-22	Milwaukee Pl., S.E.	J-22	U Pl., S.E	M-22	42nd Pl., S.E.	O-24
East Capitol St., N.E.	O-21	Minnesota Ave., N.E	Q-21	U St.	I-9, K-11, M-13	42nd St., N.E.	R-21, Q-22
East Beach Drive	R-7		R-21		N-15	42nd St.	M-7, L-8, J-10
Eastern Ave	S-7, S-10	Minnesota Ave., S.E.	M-22	U St., S.E	L-21, N-23	42nd St., S.E.	O-24
Eastern Ave., N.E	S-21		O-22	U St., S.W	J-19	43rd Pl	L-7
Easy Pl., S.E	P-24	Mintwood Pl	M-12	U St., N.E	O-16	43rd Pl., N.E.	R-21

Allin John O 3800ResrvrRdNw ---------625-7351
If no answer call ----------------223-2200
Alper Melvin G 2141 K Nw ------------223-3300
If no answer call ----------------296-4516
Alpert Hubert J MD 1234 19thNw ------223-5560
If no answer call ----------------223-2246
Res 4761BradlyBlvd ChCh ----------652-4493
Angeles Ramon Jr MD—
Hours By Appointment Only
461 H Nw -------------------- 737-4232
If no answer call Medical Bureau ----223-2200
Angelos Peter G Dr

ORIENTATION PROBLEM imposes on the seeker the task of finding the address of the physician closest to seeker's hypothetical home. Seeker is given a street map of Washington, D.C., with the address marked as shown by black dot. He also receives a street index (*middle*) keyed to the map. The source receives one page from the listing of physicians in the yellow pages of the Washington telephone directory (*bottom*). Subjects occupy separate rooms and must solve the problem by one or more modes of communication.

which the two people can neither see each other nor communicate by voice; instead they use writing machines linked in such a way that anything typed or written in longhand on one machine is reproduced on the other.

Our problem-solving tasks differ significantly from the kind usually found in the problem-solving literature of psychology because they were designed to meet certain special criteria. They sample such psychological functions as verbal skill and psychomotor skill. They are representative of tasks for which interactive computer systems are or could sometimes be employed. Instead of being abstract or artificial puzzles of the kind often devised to measure hypothetical psychological processes, they are of recognizable and practical importance in everyday life. They have definite, recognizable solutions, which can usually be reached within approximately an hour. Finally, their solution requires no special skills or specialized knowledge.

The tasks are formulated in such a way that solving them requires the seeker and the source to work together as a team. The seeker is given a problem for which he has to find the solution. His information folio consists of certain parts of the problem. The source has a folio with the remainder of the information needed to solve the problem. Neither person can solve the problem by himself, but together they have all the information needed for doing so. Remember, however, that our problems are designed to elicit communication between the two members of a team. They do not necessarily represent the way tasks would be assigned to man and computer in any real system.

All together we have constructed 10 problems that meet our needs. The following brief descriptions of three of them will convey their flavor.

In the "equipment-assembly problem" the task of the seeker is to assemble a common household article: a trash-can carrier. His information folio consists of all the disassembled parts of that article exactly as it comes from the mail-order house from which it was bought. He is not told either the name or the function of the device. The source's folio consists of the set of diagrams and instructions for assembly that came with the parts.

In the "information-retrieval problem" the seeker has to find the citation of every newspaper article relevant to an assigned topic that appeared in *The New York Times* during a given year. Usually he is told not to count editorials, reports of public speeches or letters to the editor. The source's information folio consists of *The New York Times Index* for the same year.

In the "geographic-orientation problem" the seeker's task is to find the office or residence address of the physician closest to a hypothetical home address. He is supplied with an index of streets, a gridded street map of Washington, D.C., and a card on which the home address is typed. His hypothetical home address is also marked on the map. The source is supplied with one page of the list of physicians in the yellow pages of the Washington telephone directory.

Our subjects have varied from experiment to experiment. We have relied heavily on that mainstay of psychological experiments, the college student. In one experiment, however, we enlisted high school boys, in another girls from a parochial high school and in a third a mixture of college and high school students.

We began with a series of tests involving four types of communication with less sophisticated equipment and procedures than have characterized our later experiments. In the communication-rich mode the subjects sat side by side at a table with no barrier between them. In the voice mode they were in separate rooms and communicated through a cloth panel that could be heard through but not seen through. In the handwriting mode they wrote messages in a notebook they passed through a slot in the wall between the two rooms. In the typewriter mode we had both experienced and inexperienced typists.

The results show large differences among the several modes of communication [*see illustration on next page*]. The inexperienced typists, for example, took almost two and a half times as long to solve problems as subjects in the communication-rich mode did. Differences of the same order have turned up repeatedly in other experiments.

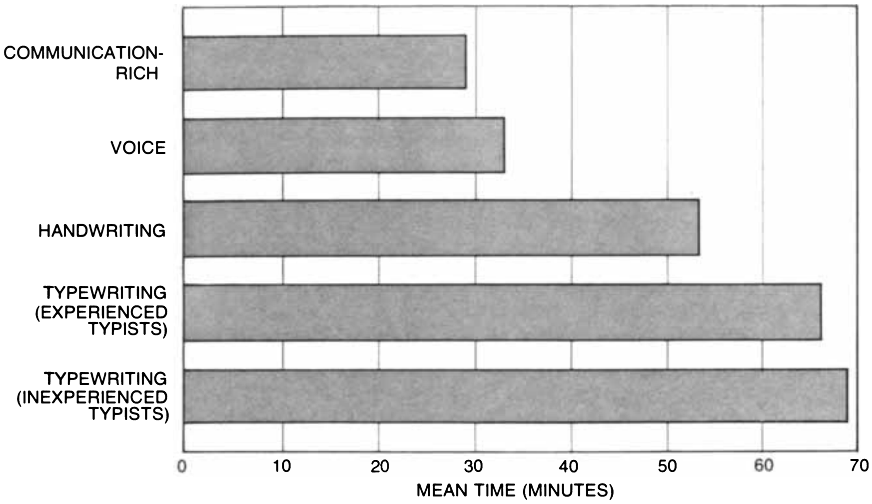
An unexpected finding was the notably small difference in performance between the experienced and the inexperienced typists. This finding seemed so implausible that we later checked it

```
goaheaddoyouknowhowto put this togher
ill tryits a trash  toter   ill  type  you  the  directions  ok
put  axle  thru  38th holes  from  outside
38th holes/ ??yes
put 1 handlebar on back of  each  outer  frame  line  up  bolt  holes
what does outer frame look like ? its  like  a  (W)
put  bottom  frame to outer frames on  front + rear of outer  frames
ok use  1+12  bolts
are  your parts  labled  by  lettrs ???
nookthe  thing  looks  like a cart with  room  for 2 trash cans the part
that looks like this(XX)goes on the bottom +the 2(W)parts go on the sids
put  male ends  ?  into  female  ends
what does that mean?  i dont  no
it  looks  like  3(u)s
what? 2(u)s   go  into  each  other  then  theyare  put on other u +put
on  W  put  top  frame  to  front  of  outer  fr.+to  handlbar  2 1/4
bolts  put  center  support  fr.  inside  topfr.  use 2 1/4  bo.  thru
center  of  top fr.  put 2  1/12  bolts thrub  center  of  side  fr.,
bottomfr. 2 bottom of center  support  fr.
okput  on wheels  3 spoks  on  outside put on  hubcap  with  hammer
put oh handgrips   DO  ALL  THESE  STEPS  FOR  BOTH  SIDES    ok?????
```

TYPED EXCHANGE between a seeker and a source engaged in solving the equipment-assembly problem is reproduced in part, with messages by the seeker underlined. In spite of such confusing messages as "38th holes" in the third line, where the source intended to say 3/8th-inch holes, the two members of the team solved the problem in less than an hour. Both were inexperienced at typing.

```
  SO:  Okay.  And but it's, all it is, is a frame.
  SK:  All it is, is a frame.  And what's supposed to fit inside the frame?
       Do you fit, do you put, you know, a pan like in a wheel barrel or/
  SO:  No, there's no pan.
 {SO:  Okay, now.
 {SK:  Or is it [Message continues below.]
  SK:  like a wagon or what?
  SO:  Now let, now let me read this for a second..It's like a wheel bar-
       rel like I said, there's a handle, there's two wheels..and all it
       is is like it's a frame wheel barrel like. But there is no, y'know,
       water will go through it in other words.
  SK:  Do wheels go in the front?
  SO:  I'm gonna, I'm gonna [Message continues below.]
 {SO:  read this.
 {SK:  Or do the [Message continues below.]
  SK:  wheels go in the back?
  SO:  I'm gonna read the direction now.  I'm gonna, give you, the, you
       know..Let me read it..oh, trash toter. Oh, is that what it is?
```

VOICE EXCHANGES between a seeker and a source working on the equipment-assembly problem are reproduced in a partial transcript in which *SO* is the source and *SK* the seeker. A bracket linking *SO* and *SK* indicates that both were talking simultaneously.



SOLVING TIME of problems is averaged for several modes of communication. In the communication-rich mode, for example, where the two members of a team were in the same room and could communicate freely, the average time for a solution was 29 minutes.

in another experiment with different subjects and more elaborate procedures to help us figure out what was going on. We now think the explanation has at least two different components.

First, by means of detailed measurements of what the subjects were actually doing when they solved problems we found that the average subject spends somewhat less than a third of his time in communication. In interactive problem solving subjects do a great many other things, such as make notes, think about what to say, handle objects and search for information in their respective folios. As a result the advantages one might expect to come from superior typing skill are diluted by the relatively small fraction of time in which the skill can be exercised.

Second, the kind of typing called for in interactive communication is unusual. Typing skill is normally measured by having subjects copy text material. In communication by typewriter, however, the communicator has to decide what to say, compose his thoughts into a message (often fragmented and incomplete) and then type out the message. The transmissions are characterized by hesitations, mistakes, changes of thought and irregular tempos that may at various times be indirect expressions of doubt, amusement or anger. In short, typing skill is usually measured as a strictly mechanical or psychomotor activity, whereas communication by teletypewriter is a much more intellectual process. It is small wonder that the two techniques seem to have so little in common.

From the voluminous literature on kinesics, gestures and "body language" I had been led to predict a large difference between face-to-face communication and communication by voice alone. The voice channel by itself seems impoverished in comparison with the variety and richness of the information-bearing clues available in face-to-face communication. The data did not conform at all to my expectations. The average amount of time taken to solve problems by voice alone was only slightly more than it was in face-to-face communication.

If this were an isolated result, one might well question its validity. It has appeared, however, in other experiments that we have done with different subjects and different problems. One of these experiments, carried out by a former student of mine, Robert B. Ochsman, tested 10 different modes of communication. Arranging them according to richness, one finds that here too the mean problem-solving time is only slightly longer in the voice mode than in the communication-rich mode [*see illustration on opposite page*].

Although solution times tend to increase as the channels of communication become more impoverished, the most striking feature of our data on the 10 modes is that they tend to fall into two fairly distinct groups. The faster five all have a voice link, whereas the slower ones do not. Statistical tests confirm that this one comparison is the only statistically significant one among the 10 communication modes.

However interesting the data on problem-solving time are by themselves, they become even more interesting when they are related to the linguistic output of the communicators. The problems of measuring and quantifying that output have in turn been almost as interesting as the results. Most psycholinguistic research has been done on what I call immaculate prose. Such prose consists of grammatically correct sentences with nouns, verbs and other parts of speech in their proper place. Words are spelled correctly and rules of punctuation are observed. All computer programs based on what is called natural language require immaculate prose because the sentences that are fed into the computer are parsed in one way or another so that the meaning of the ensemble can be inferred from conventional rules of syntax.

The trouble is that people do not naturally speak in sentences. Most of us realize this in an intuitive way, but I suspect that few of us appreciate just how untidy normal human conversations really are. In our experiments recordings are made of conversations in all communication modes having a voice channel, and the recordings are then transcribed into typewritten protocols. Subjects generate their own protocols when they converse in the handwriting and typing modes.

Looking at the transcripts of conversations by voice or writing, one readily sees the untidiness I have mentioned. An example is the record produced by two inexperienced typists who were solving the equipment-assembly problem [*see top illustration on preceding page*]. The first impression is one of complete unruliness. Not one grammatically correct sentence appears in the entire protocol. Words are continually misspelled and run together. Abbreviations, both conventional and unconventional, are common, and violation of the rules of punctuation is commoner than their observance.

The record even includes serious errors of fact. For example, at one point the source gives the seeker an instruction about "38th holes." The seeker queries the source on this point, and the source replies that his original statement was correct. Actually what the assembly instructions stated and the source meant to say was that the seeker should put the axle through the 3/8th-inch holes.

Perhaps the most remarkable thing is that in spite of all this apparent unruliness, the information got through. The two team members who generated the protocol completed their task successfully in less than the average time required by teams for the equipment-assembly problem. Equally remarkable is that the protocols of the experienced typists were also characterized by many of the same kinds of error and ungrammatical feature. Evidently unruliness tends to creep in when the emphasis is on natural communication rather than on precision of typing.

Transcripts of voice conversations have their own special idiosyncrasies that are no less perplexing and difficult to deal with [*see bottom illustration on page 39*]. It is clear that if truly interactive computer systems are ever to be created, they will somehow have to cope with the mispronunciations, errors and violations of format that are the rule rather than the exception in normal human communication. Discovering the rules and characteristics of normal communication is a problem that has been ignored by linguists for too long.

Measuring and counting such characteristics of the protocols as words, sentences and messages had seemed simple in prospect but proved difficult in execution. In the end, however, we were able to formulate sets of rules that enabled us to count the linguistic units. Next came the task of trying to decide what measures of linguistic performance to apply to our linguistic units.

On the basis of hunches, hypotheses and what we could find in the psychological literature we came up with 136 linguistic measures. A number of them turned out to be trivial, and many others were so highly intercorrelated that they were redundant. In the end we were left with only nine meaningful measures of linguistic performance that describe our data. To a certain extent the discarded measures are as interesting as the useful ones because they reveal where it would be fruitless to expend time and effort in the future. To list them here, however, would consume too much space.

The productive measures are: (1 and 2) The number of messages generated by each subject and, closely correlated, the number of sentences. (3 and 4) The number of words per message and, closely correlated, the number of words per sentence. (5) The percentage of sentences that were questions. (6) The total number of words employed by a subject. (7) The total number of different words employed by a subject. (8) The ratio of different words to total words, called the type-token ratio. (9) The communication rate, which is the number of words communicated per minute of time actually spent communicating.

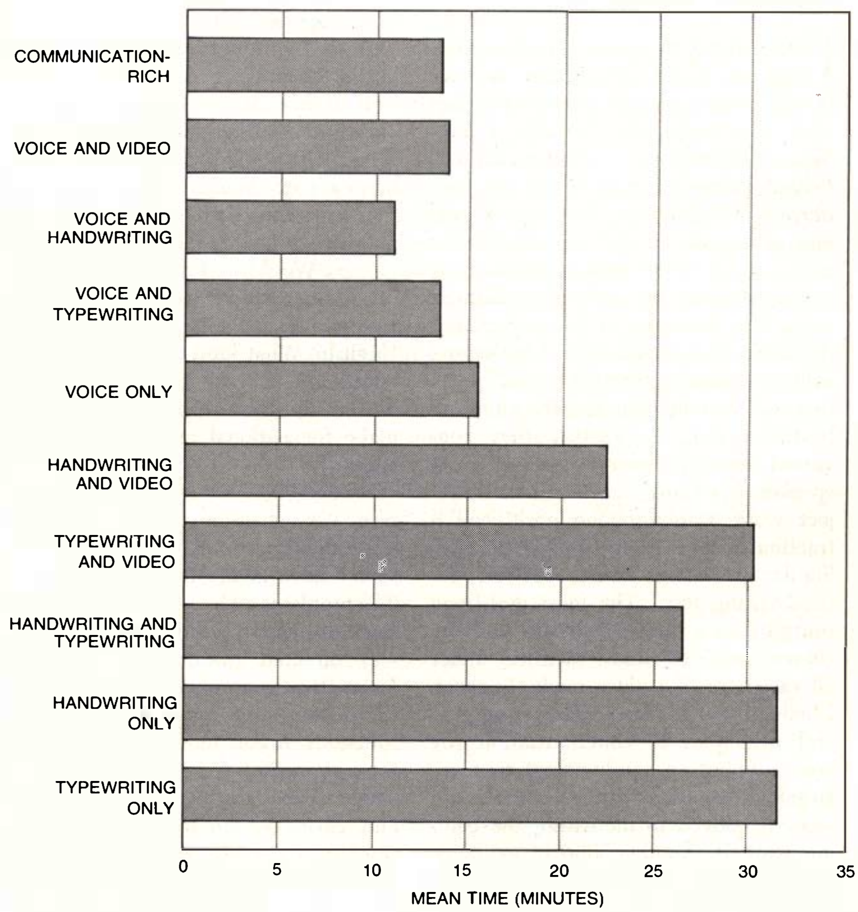
In one sense our findings are disappointing, since there appears to be so little to show for so much effort. In another sense, however, they are gratifying. The linguistic performance of people who communicate naturally can be described by a rather small number of quantitative measures.

When the data are summarized [*see illustration on next page*], the most striking thing about them is that the two fast modes of communication (the two that have a voice channel) are also extremely wordy. Subjects using the two voice modes, as compared with handwriting and typing, delivered about eight times as many messages, eight times as many sentences, five times as many words and twice as many different words; they also communicated words at a rate nearly 10 times as fast. In sum, the voice modes are fast ways of communicating, but they are extremely wordy, no matter how wordiness is measured.

The higher type-token ratio for the handwriting and typing modes confirms that they tend to be more parsimonious and less redundant than the voice modes. (Measuring redundancy in these experiments is extremely difficult because the data do not lend themselves to the conventional measures of redundancy relied on by information theorists. By making certain plausible assumptions, however, we concluded that in the voice modes subjects employ about 13 times as many words and four times as many unique words as are really required to solve problems.)

Interruptions in normal conversations are so common and apparently so important that we have tested their effects in a separate experiment. In brief, we found that if subjects do not have freedom to interrupt, they use fewer messages and more words per message. They also maintain a relatively constant number of messages in both speaking and typing.

Allowing subjects freedom to interrupt does not shorten the time needed to solve problems, nor does it result in any reduction in the number of words needed to reach solutions. What does happen is that when subjects have the freedom to interrupt, they package their words differently. They use more messages and fewer words per message, and they maintain a relatively constant message length whether they speak or communicate by typewriter. A final point of interest is that a subject is much more



MODE OF COMMUNICATION **influenced the time required to solve problems. Here the average time taken by teams to solve problems is charted for 10 different modes of communication. The data fall into two fairly distinct groups. The faster five all involve the use of the voice in communication, whereas the voice is excluded in the five slower modes.**

	COMMUNICATION-RICH	VOICE	HANDWRITING	TYPEWRITING	
				EXPERIENCED TYPISTS	INEXPERIENCED TYPISTS
SOLUTION TIME IN MINUTES	29	33	53.3	66.2	69
NUMBER OF MESSAGES	230.4	163.8	15.9	27.2	31.5
NUMBER OF SENTENCES	372.6	275.9	24.9	45.8	44.1
TOTAL NUMBER OF WORDS	1,563.8	1,374.8	224.8	322.9	257.4
TOTAL NUMBER OF DIFFERENT WORDS	397.5	305.9	118.5	150.5	133.4
TYPE-TOKEN RATIO	.3	.3	.6	.5	.6
NUMBER OF WORDS PER MINUTE	190.3	171.2	17.3	18.1	10.2

EXPERIMENTAL RESULTS are enumerated for the solution of problems by various modes of communication. "Type-token ratio" is ratio of different words to total words. Problem solving by voice takes the least time but is wordier than the other modes are.

likely to take control of a voice channel than of a channel or any combination of channels lacking the voice.

In considering the generality of our findings one might ask whether we have found certain general principles of human communication or have merely found out about the ways in which particular pieces of equipment are employed. We think we have found general principles. In our experiments we tested some of the various channels of communication in distinctly different ways. For example, in one experiment the voice channel was tested by having subjects converse through a cloth panel that was visually opaque but acoustically transparent. In another they conversed through a microphone and loudspeaker. In a third experiment each subject wore a microphone positioned a fraction of an inch in front of his lips. Similar variations were devised for handwriting tests. The most gratifying thing to me is that the results and the comparisons are almost identical under all variations of a given mode. In short, I believe we have been discovering general principles of communication (by voice, typing or handwriting) that are largely independent of the particular devices employed in mediating the communications. Another kind of evidence bearing on the generality of our findings is that the results obtained with our different modes of communication hold for all the different problems we have tested and for both job roles assigned to the communicators: source and seeker.

As so often happens in research, the questions our findings have raised are more numerous than the answers they have supplied. Every reader will have his own list of questions. Four in particular intrigue me.

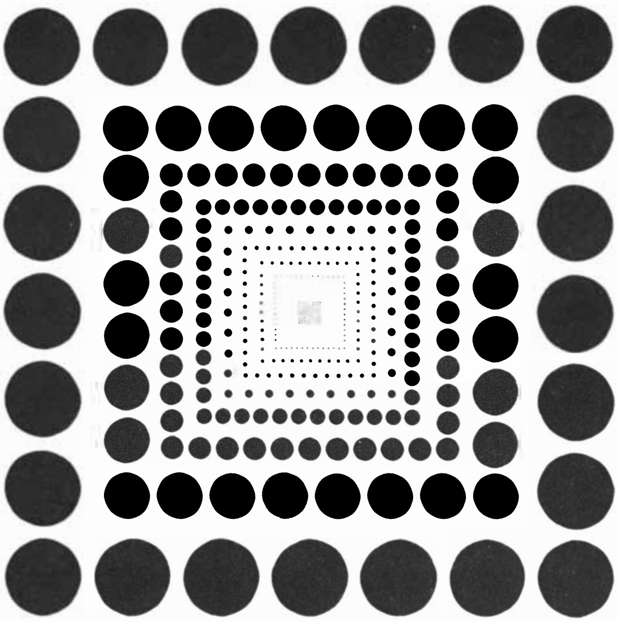
First, how do communication patterns vary among different nationalities? Anthropologists and sociologists tell us (and our own experience seems to confirm) that communication patterns differ markedly among different cultural groups. Would we have obtained results of the same kind if we had studied people who speak a language other than English? What kind of concessions will telecommunication systems in general and computers in particular have to make for national and cultural differences?

Second, how do communication patterns vary according to the purpose of the communication? All our experiments so far have involved factual problems. The problems all had single answers, and the information needed to solve them was directed toward that one goal. Interactive communication, even with computers, may serve many different functions. A communicator may browse through masses of data for items that he needs or that merely excite his interest and curiosity. He may want to have briefings and status reports that will help him to anticipate weather conditions for the next few hours or days, review the state of the economy or update his information about the condition of patients in a hospital. Communications can also provide information necessary for reaching decisions among conflicting alternatives and can serve in argumentation, bargaining and persuasion. Virtually nothing is known about how communication patterns vary according to these diverse purposes.

Third, what happens to communication patterns as the number of communicators increases? Our experiments were all done with teams of two people, but communications often involve a group of people and perhaps a computer. The full implications of this kind of communication are not known.

Finally, what are the rules that govern normal human communication? Perhaps the most interesting question of all concerns the grammatical, syntactical and semantic rules that apply to such communication. In spite of the apparent unruliness of natural communication, it obviously follows some rules, because problems do get solved, often with surprising speed. How are meanings conveyed in natural conversations? How can we even go about investigating this problem?

This brief introduction to our program of research has conveyed only a few of the many interesting findings we have made. Still, it may be enough to excite the interest of others to try to understand what happens when people communicate. If enough people work on these problems, who knows? Perhaps at some future time you and I shall be able to find out about the latest developments in science not by reading articles such as the ones in this magazine but by conversing in ordinary English with a computer.



DOTS

or

Complexity is no longer objectionable

It began to happen quite recently. Engineers, realizing that the mythical mind of the well-programmed and capacious digital computer does not boggle, put it to work translating mind-boggling complexities of logic into physical form. The ancient philosophers who invented logic as a game that is all in the mind had perhaps never actually encountered minds that do not boggle. To put logical constructs into a physical form that is readily multiplied, one turns to a craft generally credited to Gutenberg. The particular physical form, whether a metal-oxide-semiconductor (MOS) chip, a bipolar integrated circuit, or a charge-coupled device (CCD), rests also on subsequent contributions by the artists who invented photography, by the inorganic chemists who vied to discover new elements in rocks, by the physicists who invented quanta and electrons to explain their experiments, by the organic polymer chemists, even by us in one small way or another.

The physical form comes down to a bit of jewel patterned in lines and dots much finer than those by which a printing press prints pictures. The scanning electron microscope discloses thousands of electronically active and passive components tied together by logic. Economies of duplication and size for costly materials of the rarest purity open mass markets for all manner of artifacts that interact with the customer's central nervous system, homunculi who live in tiny boxes and are looking for new work beyond telling time and performing difficult calculations. When complexity is affordable, one can do almost anything.

The technology that generates the dots for the printing press turned out to be the way to carve and deposit the patterns the computers generate. From serving the printing and graphic arts industry for a long time, we know that technology well and currently serve the photolithographic departments of today's electronics industry not only with high-resolution plates and light-sensitive resists but by encouraging communication within their craft.

Here, for example, are the titles of papers presented at the most recent of the Kodak Microelectronics Seminars:

The "Oxygen Effect" in Photoresists and Its Manifestation in Processing Parameters

A Laser Reticle Generator

Performance of an EBPG (Electron Beam Pattern Generator) as a Link in a Semiconductor Chip Manufacturing Chain

Mechanistic Aspects of Photocrosslinking in Negative Photoresists

Modelling Positive Photoresist

An Electrical Method for Detecting Pinholes

Effective Water Removal from High-Resolution Plates

Advances in Contact and Proximity Printing

Holographically Prepared Gratings for Integrated Optics

Chrome Is for Car Bumpers and LSI Mask Technology

The Characterization of Positive Photoresists

Processing Techniques

The Chemistry and Control of Processing High-Resolution Plates

Lenses and Optical System Used in Microelectronics

Monosubstituted Vinyl Polymers as Highly Sensitive Negative Electron Beam Resists

If any of them specifically interest you drop us a note at Dept. 55W, Kodak, Rochester, N.Y. requesting a copy.

Message to a few very special people who all too well know the reality of it all

There is more to the art than photolithography, of course. Now is the time we start building reliability into clever Kodak apparatus as much as a decade away from the marketplace. S. C. Wright, Business and Technical Personnel, Kodak, Rochester, N.Y. 14650 seeks the acquaintance of a few individuals highly qualified by direct practical experience to begin a new career in a research laboratory environment where the scientific underpinnings of MOS technology are to be extended. Wright also would like to know of others who are very good at designing IC's, or at directing fabrication of prototypes of new designs to prove out in working products. This must long precede production orders placed with suppliers of IC's. We are an equal-opportunity employer f/m.

Kodak

EPA's highway test average.

mph, 38 mpg,
Rabbit.

Happy days are here again.

†See your dealer for more details. ©Volkswagen of America, Inc.

The Effects of Ice on Scotch

How fast a drink of Scotch whisky over rocks loses its flavor depends on the proof of the Scotch and the richness of its blend. These two factors are optimized for "on the rocks" Scotch drinkers in 90-Proof Famous Grouse, a venerable old brand from Scotland only recently introduced to America.

by Allen Mac Kenzie

In countries where Scotch has been consumed for centuries, ice and whisky rarely mingle. But on this side of the Atlantic, the picture is quite different. While a small percentage of American Scotch drinkers take it neat, better than 35% drink it "on the rocks." The rest of us add varying amounts of water, club soda, etcetera. And ice. Always plenty of ice—the great American drink requisite.

It would seem then that the American Scotch devotee, particularly our on-the-rocks fancier, has a right to raise a serious question: *Is the Scotch I drink ideally suited to enjoyment over ice?*

Pursuing a Perfect Proof

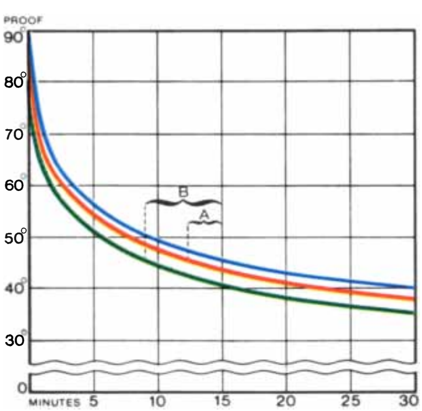
Let's turn our attention first to the proof at which Scotch whisky is bottled. Consider the hypothesis that there is indeed a better proof for on-the-rocks Scotch drinking than that of the brand you currently favor.

Practically every Scotch sold in this country is bottled at 80, 86, or 86.8 Proof. So at the instant you pour Scotch over ice, it contains between 40% and 43.4% alcohol by volume. (Proof is double the percentage of alcohol.) The chilling effect of the ice is accompanied by dilution. And when your drink has been properly cooled—in 30 seconds to a minute—you achieve what one Scotch connoisseur refers to as "the ideal sip." From then on, the Scotch drinker's enjoyment typically runs downhill, as the drink loses its freshness.

While there is no way to preserve that fresh Scotch flavor indefinitely, we submit that you can sustain the freshness substantially longer with 90-Proof *Famous Grouse*. If you have never heard of this brand, we are not surprised. It is a well established name in Scotland, but only recently introduced to America. So far as we know, *Famous Grouse* is the only Scotch now available in this country at 90 Proof.

A Revealing Experiment

To demonstrate the merits of a slightly higher proof, we performed a simple experiment: 50 millilitres of Scotch (about 1.7 ounces) was chilled with 100 cc of ice. The ensuing dilutions at 80, 86.8 and 90 Proof are charted in the graph at left.



DILUTION BY ICE OF SCOTCH AT THREE DIFFERENT PROOFS (72°F).
80 Proof 86.8 Proof 90 Proof

You'll notice that after 15 minutes on the rocks, the proof of *Famous Grouse* is diluted to a level which occurs after 12½ minutes when the Scotch is 86.8 Proof, and after 9 minutes when it is 80 Proof. In essence, the *Famous Grouse* brand has remained about 2½ minutes fresher than 86.8-Proof Scotch (Interval A on graph), 6 minutes fresher than 80-Proof Scotch (Interval B). If you "nurse" a drink beyond 15 minutes, the advantages of 90 Proof Scotch are even more pronounced.

Proof, of course, is not the only influence on the flavor of a blended Scotch. The proportion of malt to grain whiskies, origins of the malts, aging methods—these are also important factors determining the relative richness of Scotch flavor.

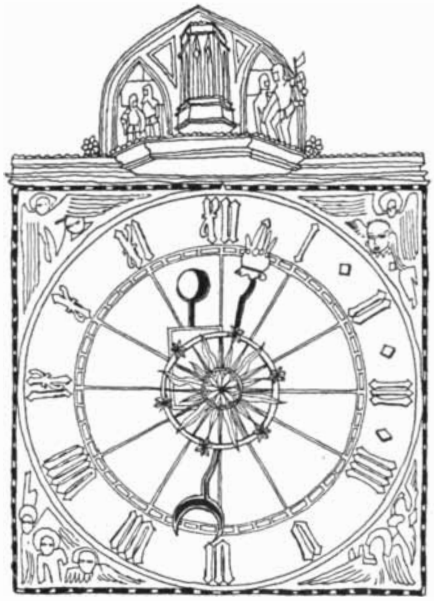
The makers of *Famous Grouse*—Matthew Gloag & Son of Perth, Scotland—have been producing Scotch in the same family for six generations. And they have performed their most noble feat in the rich blend they created for *Famous Grouse* Scotch. Its flavor—so remarkable at the outset—holds firmly to its character during prolonged contact with ice.

Knowledge of Scotch, however, cannot be indefinitely pursued in the abstract. Your learning process must ultimately include a leisurely sip of *Famous Grouse* on the rocks. For Scotch drinking is one of those pleasures enjoyed most, not in the pursuit, but in the conquest. Scotland's greatest bard, Robert Burns, said it best:

"Gie me a spark o' Nature's fire,
That's a' the learning I desire."

90 PROOF, BLENDED SCOTCH WHISKY BOTTLED IN SCOTLAND
IMPORTED BY AUSTIN, NICHOLS & CO., LAWRENCEBURG, KY

SCIENCE AND THE CITIZEN



The Arms Race Institutionalized

Ever since the announcement from Vladivostok last November that a tentative strategic-arms agreement had been worked out between the U.S. and the U.S.S.R. it has become increasingly evident that the main short-term effect of this latest step in bilateral arms-control negotiations would be to authorize increased military expenditures on both sides. The Vladivostok understanding, which is now serving as the basis for the third round of the Strategic Arms Limitation Talks (SALT III) in Geneva, sets an upper limit of 2,400 on the total number of long-range delivery vehicles (bombers, land-based missiles and submarine-based missiles) permitted on each side for the next 10 years. The agreement also stipulates that neither side may equip more than 1,320 of its missiles with multiple independently targetable reentry vehicles (MIRV's).

As was pointed out at the time, these ceilings are substantially higher than the present strategic arsenals of both superpowers. Moreover, the total-delivery-vehicle figure is, in the case of the U.S., even higher than the ceiling of 2,160 envisioned for 1977 by the SALT I Interim Agreement on Offensive Weapons. (The stated purpose of the Vladivostok discussions was to extend beyond 1977 the terms of the SALT I interim agreement, which was described as merely a "first step," albeit a "historic" one, along the complex negotiating path.)

At present the U.S. has installed MIRV's on some 800 of its land-based and submarine-based missiles; this figure is scheduled to rise to approximately 1,050 by the time the current missile-conversion program is completed in 1977. The U.S.S.R. has tested MIRV's and is just beginning to install them on some of its land-based missiles. Clearly the new ceiling of 1,320 MIRVed missiles is far in excess of what either side now has, or what at least one side (the U.S.) apparently ever planned to have.

At a press conference in December, President Ford declared that the U.S. has an "obligation" to build its strategic forces up to the permitted ceilings. He also observed that the Russian land-based missiles tend to be larger and to have a greater "throw weight" than the American ones, and that the Vladivostok agreement "does not preclude the U.S. from increasing its throw-weight capability." A short time later Secretary of Defense James R. Schlesinger was reported as saying that the "restructuring" of the U.S. strategic arsenal necessitated by the new agreement would entail "some upward adjustment" in the U.S. strategic-arms budget.

The extremely high ceiling of 1,320 MIRVed missiles, it has been argued by some critics of the Vladivostok accords, must be viewed by the U.S.S.R. as a virtual invitation to embark on a large-scale MIRVing program in an effort to narrow the wide U.S. lead in this area. On the other hand, Secretary Schlesinger has warned that any "massive" deployment of MIRV's by the Russians would be viewed by the U.S. as "a potential source of strategic instability" and might therefore force the U.S. "to take countermeasures to maintain the strategic balance."

Meanwhile the continued omission of any meaningful qualitative controls in the Vladivostok understanding, a defect that threatens to carry over into the forthcoming SALT III treaty, leaves the way open for the two superpowers to engage in a costly strategic-arms "quality race," a particularly risk-filled course that many advocates of arms control warned against in their original criticism of the SALT I agreements. As William Epstein, a former United Nations disarmament official and a long-time observer of the armaments race, remarked recently: "It seems that the agreements already concluded, and indeed those being negotiated, are designed not to halt or reverse the arms race but rather to institutionalize it and regulate it."

The Geneva Protocol Accepted

Biological weapons have now been negotiated out of the arsenals of most of the world's major military powers, and poison gas may be on the way out. In January the U.S. acceded to the Geneva Protocol of 1925, banning any first use of gas and bacteriological weapons, and to the Biological Weapons Convention of 1972, banning the development, production or possession of bacteriological and toxin weapons. This month the Conference of the Committee on Disarmament is scheduled to meet in Geneva to take up proposed treaties that would move the world toward actual chemical disarmament. That will involve troubled issues of verification and inspection, however, and also a major difference in definition: the U.S., unlike the rest of the world, has held that riot-control gases and herbicides are not agents of chemical warfare.

In ratifying the Geneva Protocol the U.S. became the 106th party, half a century late, to a treaty this country initiated soon after World War I. The first time around, in 1926, the Senate failed to approve the protocol after strong lobbying by the chemical industry and other opponents. The Nixon Administration revived it as part of a new effort on chemical and biological weapons that included the unilateral renunciation of all biological weapons in 1969, the destruction of bacteriological- and toxin-weapons stockpiles in 1971 and the negotiation of the biological-weapons convention in 1972. When the Administration resubmitted the Geneva Protocol in 1970, however, it added an explicit interpretation to the effect that the treaty did not cover riot gas or herbicides, both of which the U.S. was then deploying in Vietnam. The Senate Foreign Relations Committee did not agree, and it deferred action on the biological-weapons treaty as well as on the Geneva Protocol.

Last fall the Ford Administration negotiated a compromise with the Senate committee. It agreed not to write its interpretation into the treaty. Instead Fred C. Iklé, director of the Arms Control and Disarmament Agency, announced that the President was prepared, "in reaffirming the current U.S. understanding of the scope of the Protocol [in other words, reaffirming that the treaty does not cover riot gas and herbicides] to renounce as a matter of national policy" the first use of such agents—except in limited, specified circumstances. The committee approved both treaties, the Senate agreed and President Ford signed the instruments of ratification in January.

The 1925 protocol bars the first use (not the retaliatory use) of gas and bacteriological weapons. The 1972 convention bars the development, production or other acquisition of any biological weapons. As of now there is no prohibition on the production or stockpiling of chemical weapons, including nerve gases. The U.S. and presumably the other major powers have large stocks of extremely potent nerve gases. Public concern about nerve-gas tests and the possibility of accidental discharge has prompted the U.S. Army to conceive the binary gases: two nonlethal constituents of a nerve gas would be stored in separate containers that could be assembled in the field to provide lethal artillery shells or missiles. Congress refused to put up the money for such weapons last year, but the Army has advertised for bids nonetheless.

The U.S. and the other signatories of the 1972 biological-weapons convention committed themselves then to eventually eliminating all chemical weapons. The Geneva conference this spring will have before it draft treaties, presented by the U.S.S.R. and by Japan, aimed at full chemical-warfare disarmament. The Japanese proposal would simplify the difficult problem of verification by barring only the production and transfer of such weapons as a first step. It remains to be seen what effect the U.S. moves with regard to binary weapons will have on the talks and also what the U.S., having somewhat obscured its position on riot gas and herbicides in acceding to the protocol, will say about those weapons during the new negotiations.

Big Tokamaks

The decision of the Ford Administration to include in the Federal budget for fiscal year 1976 a request for some $7.5 million to start work on a major new fusion-power test facility at Princeton University places the future of the U.S. controlled-thermonuclear-research program in the hands of Congress. If the requested funds are approved, detailed design of the proposed installation could begin almost immediately, with component fabrication and site construction scheduled to get under way late next year. Assuming that everything goes according to plan, the experimental fusion reactor—the first U.S. system of its kind that is expected to reach the "break even" threshold for net power output—would be finished and ready for operation by 1980 at a projected total cost of approximately $215 million.

The Princeton machine would be the latest, largest and by far the most costly in the series of toroidal (doughnut-shaped) plasma-confinement systems known generically as tokamaks, after the original Russian devices of a decade or so ago. According to a three-volume conceptual-design study prepared jointly by workers at Princeton's Plasma Physics Laboratory and at the Westinghouse Electric Corporation, the new tokamak would be of a type described as a two-component torus, or TCT, machine. The name refers to the proposed scheme for obtaining fusion reactions within the magnetically confined plasma, or gas of electrically charged particles. In the TCT approach this is accomplished in part by directing a beam of accelerated neutral atoms at the lower-energy bulk plasma. The neutral atoms, which are able to cross the magnetic-field lines unhindered, enter the plasma and are promptly ionized (that is, stripped of their electrons); the resulting energetic ions are then trapped by the magnetic field, thereby adding both energy and particles to the plasma.

It is the addition of this second, injected component to the tokamak's primary "ohmic" mode of heating (so named because it results from the resistance offered by the plasma to an induced electric current flowing through the core of the plasma column) that is expected to make it somewhat easier to raise the plasma to the break-even threshold. (By definition the break-even criterion does not take into account such extrinsic factors as the considerable power still needed to maintain the magnetic field. In an operational fusion reactor superconducting magnets would presumably be employed to minimize the latter power requirement.)

In many respects the proposed TCT machine would be simply an enlarged version of another tokamak, called the Princeton Large Torus (PLT), currently under construction at the same laboratory. When the PLT is completed later this year (at an estimated cost of $13 million), it will be for a time the largest tokamak in the world. Experiments with both machines are expected to begin with a pure hydrogen plasma. At a later stage of the TCT's operation, however, fusion reactions involving the heavy hydrogen isotopes deuterium and tritium would be studied, since these are the fuels most likely to be burned in a full-fledged fusion-power reactor.

Elsewhere two other large tokamaks, both on roughly the same scale as the TCT, have been proposed: one in Europe and one in Japan. Meanwhile at the I. V. Kurchatov Institute of Atomic Energy in Moscow (the birthplace of the tokamak concept) plans are reportedly being made to proceed with the construction of an even larger machine, to be designated the T-20, which would be the lineal descendant of the original tokamaks.

Doctor's Dilemma

In the past 10 years physicians and hospitals have been sued for malpractice with increasing frequency. Settlements for large sums of money have become commonplace. In California, for example, there were three malpractice settlements of $300,000 or more in 1969; in 1974 there were more than 30. In the nation as a whole there have been between 30 and 35 malpractice awards of $1 million or more; all but two or three have been made since 1970.

The trend reflected in these statistics has been interpreted in several ways, but most of the explanations have at least one element in common: the rising expectations of the public. Patients are apparently more inclined to blame the physician and to take legal action when they are dissatisfied with the medical care they have received. In addition juries are evidently more willing to believe that the physician has erred, and that he should be held accountable for his mistake.

One factor that contributed to the proliferation of malpractice suits is pointed out by David S. Rubsamen, a physician and attorney who serves as a legal consultant to California physicians and hospitals. The factor is a new interpretation of a legal doctrine that once required expert witnesses to be physicians practicing in the same community as the defendant. Expert testimony is crucial to the success of most malpractice cases; before about 1960 it was rarely available because physicians were unwilling to testify against colleagues in their own community. Twenty states have now relaxed or eliminated the geographic restrictions.

When it became possible to obtain expert testimony, successful malpractice suits became more common, and the first large settlements were widely publicized. As a result, Rubsamen says, "the public learned that physicians can be negligent and that they can be sued."

In the past three or four years the character of malpractice law has been transformed again. Formerly charges of negligence usually involved a fairly obvious error or mishap, such as leaving an instrument in the patient's body during surgery. Recently physicians have been held liable for errors of judgment, such as the failure to diagnose breast cancer in its earliest stages or the failure to manage a complex chronic illness.

Among the consequences of escalating malpractice settlements has been a rapid increase in the cost of malpractice insurance. The company that insures most New York physicians nearly doubled its rates in July, 1974: physicians in high-risk specialties were charged an annual premium of $14,500. In December a further increase was announced that would have almost tripled that premium; the increase was rescinded, but after June the company will no longer offer malpractice insurance in New York.

Higher insurance premiums charged both to physicians in private practice and to hospitals will obviously lead directly to higher fees charged to the patient. Malpractice suits also raise the cost of medical care in a more subtle way: through the adoption of various procedures characterized by physicians as "defensive medicine." Included are laboratory tests and consultations when they are ordered not from medical necessity and not primarily for the benefit of the patient but in order to provide a defense in the event that negligence is later charged. Because these precautions increase the capital expenditures of the hospital their cost is shared by all patients.

Numerous solutions to the malpractice problem have been proposed, but so far no state has adopted any of them. Rubsamen has recommended that malpractice cases be tried before a judge instead of a jury; a judge specializing in malpractice law, he contends, would be better qualified to evaluate the merits of a case and could set awards more consistently and more equitably. Rubsamen would also eliminate the practice of compensating a patient for future medical expenses in a single cash payment and would modify or eliminate the contingency-fee system, by which attorneys receive a percentage of the settlement.

Another possible solution has been proposed by Gerald J. Lustig, a thoracic surgeon and president of the New York State Federation of Physicians and Dentists. His plan would abolish malpractice as a legal concept and remove it from the courts. A patient who had been injured would be compensated regardless of the culpability of the physician, perhaps through a program analogous to workmen's compensation insurance. A related "no fault" plan was recently endorsed by the American Medical Association. "When a patient is aggrieved, he should be paid appropriate compensation, and he should not have to take his chances in court to get it," Lustig says. "What we are interested in is improving the quality of medical care; the present system of malpractice law is not doing that."

The Disappearing Open Hearth

Steel is made by burning other elements out of iron. For many years the combustion was carried out with air in the Bessemer converter or the open-hearth furnace, but in 1952 a new "basic oxygen process" employing gaseous oxygen of very high purity was introduced on a commercial scale in Austria. By the end of 1974, according to Joseph K. Stone of Kaiser Engineers, Inc., the process had caught on so well that it accounted for almost 65 percent of steel production in the U.S. and almost 60 percent of world production.

The basic oxygen process, which is also known as the L-D process (the origin of the initials is usually given as Linz and Donawitz, the two places in Austria where the first commercial plants were built), did not challenge the open-hearth method in the U.S. until 1960. Stone, in a review titled "History of the L-D Process," attributes the "phenomenal growth" of the process to its "low operating costs and low capital cost," with the high quality of the steel and the relative simplicity of the plant as "added factors." He predicts that by 1980 the process will account for 75 percent of the steel made in the U.S.

Variable Star

The sun is a variable star. Sunspots appear on the solar disk and change as they move across it, and all solar activity from flares to magnetic storms follows an 11-year cycle. Even the speed at which the sun rotates on its axis is not constant from one year to the next. The sun's variability is still more apparent when it is examined over longer periods of time. In a recent discussion E. N. Parker of the University of Chicago gives several examples of exactly how dramatic solar variations have been over the past few centuries.

On September 25, 1909, a blast of ionized particles from the sun gave rise to a violent magnetic storm on the earth. Auroral displays are normally created by the interaction of solar particles, the earth's magnetic field and the atoms and molecules of the earth's upper atmosphere at high latitudes; on this occasion an aurora was seen in Singapore, one degree north of the Equator. A similar storm on May 14 and 15, 1921, produced an aurora that was visible from Samoa, 13 degrees south of the Equator. Over the rest of the earth radio, telegraph and telephone communications were profoundly disturbed.

On the other hand, in the 70-year-period between 1645 and 1715 solar activity was so low that sunspots, which had been monitored since the time of Galileo, were unprecedentedly few and were present only on one hemisphere of the sun. Auroras all but vanished, with perhaps only one notable display per 11-year cycle even at high latitudes. The solar corona, the silvery surround that normally appears during a total eclipse of the sun, was absent in the eclipses of the period.

The quiet period was discovered 75 years ago by E. W. Maunder, a British solar physicist, when he was searching through old records of solar observations. He published two reports on the discovery, the first at the turn of the century and the second in 1922. The second paper came to the attention of A. E. Douglass of the University of Arizona, in the course of his pioneering studies of tree rings and tree-ring dating. He had previously noted that trees from arid parts of the world, which normally exhibit pronounced annual rings as a result of growth during a brief wet season, showed erratic growth patterns during the last half of the 17th century. After reading Maunder's paper Douglass realized that the absence of cyclical growth patterns coincided with the period of the curious lack of solar activity.

The activity of the sun is stronger at higher solar latitudes. Since the earth travels around the sun in a plane that is tilted only seven degrees with respect to the sun's equator, it never moves through regions of interplanetary space where the sun's activity (and variations in it) would be most apparent. Moreover, the spacecraft from which studies of the sun have been made have also been confined to the plane of the earth's orbit. Parker suggests that a fuller picture of the sun's variability could be obtained by sending a spacecraft out of the earth's orbital plane to view the sun from above its poles.

THE EARTH'S MANTLE

This great body of hot rock accounts for 83 percent of the volume of the earth and 67 percent of its mass. Although the mantle is inaccessible, much has been learned about it by indirect evidence

by Peter J. Wyllie

The mantle of the earth is a thick shell of red-hot rock separating the earth's metallic and partly melted core from the cooler rock of the thin crust. Starting at an average depth of from 35 to 45 kilometers (22 to 28 miles) below the surface and continuing to a depth of some 2,900 kilometers, it accounts for nearly half of the earth's radius, 83 percent of its volume and 67 percent of its mass. Its influence on the crust is profound; indeed, the crust and its thin film of ocean and atmosphere are distillates of the mantle, and the driving forces that move the continents slowly about on the earth's surface arise within the mantle. Knowledge of the mantle is therefore crucial to an understanding of the structure and dynamic behavior of the earth. Notwithstanding the inaccessibility of the mantle, a considerable amount of information about it has been assembled by more or less indirect means.

The role of the mantle in influencing conditions at the surface is manifold. For example, during the 4.6 billion years since the origin of the earth in the solar nebula, melting of the more fusible constituents of the mantle has produced lavas that rise to the surface and solidify, adding new rocks to the crust and giving off the water vapor and other gases that go to the atmosphere and the oceans. On another scale gaseous carbon compounds that came from the mantle began the story of life on the earth when they provided the raw material for organic molecules.

Similarly the mantle as a driving force has multiple effects. The surface of the earth is shaped by the action of the mantle, moving very slowly below the crust. Mountains rise and persist because of this movement; without it erosion would wear them down to sea level within 100 million years or so. Movements of the mantle also cause volcanic eruptions, earthquakes and continental drift.

A cross section through the earth shows the concentric layers of the core, the mantle and the crust [*see top illustration on opposite page*]. They differ from one another in composition or physical state or both. The mantle is composed of silicate minerals rich in magnesium and iron, with an average composition corresponding to that of the rock peridotite. (The name comes from the fact that the most abundant mineral in peridotite is olivine, which is more familiar as the transparent green gemstone called peridot.) The mantle is solid but within the relatively thin zone ranging from about 100 kilometers below the surface to about 250 kilometers the rock may be partially melted, with thin films of liquid distributed between the mineral grains. This zone is called the low-velocity layer for reasons that will become apparent below.

The density of the mantle increases with depth from about 3.5 grams per cubic centimeter near the surface to about 5.5 grams near the core. It does not increase smoothly; the curve of density displays distinct steps [*see illustration on page 52*]. They indicate significant changes in the mantle rocks at depths near 400 and 650 kilometers. The distribution of density provides the basis for the calculation that the mantle makes up 67 percent of the mass of the earth.

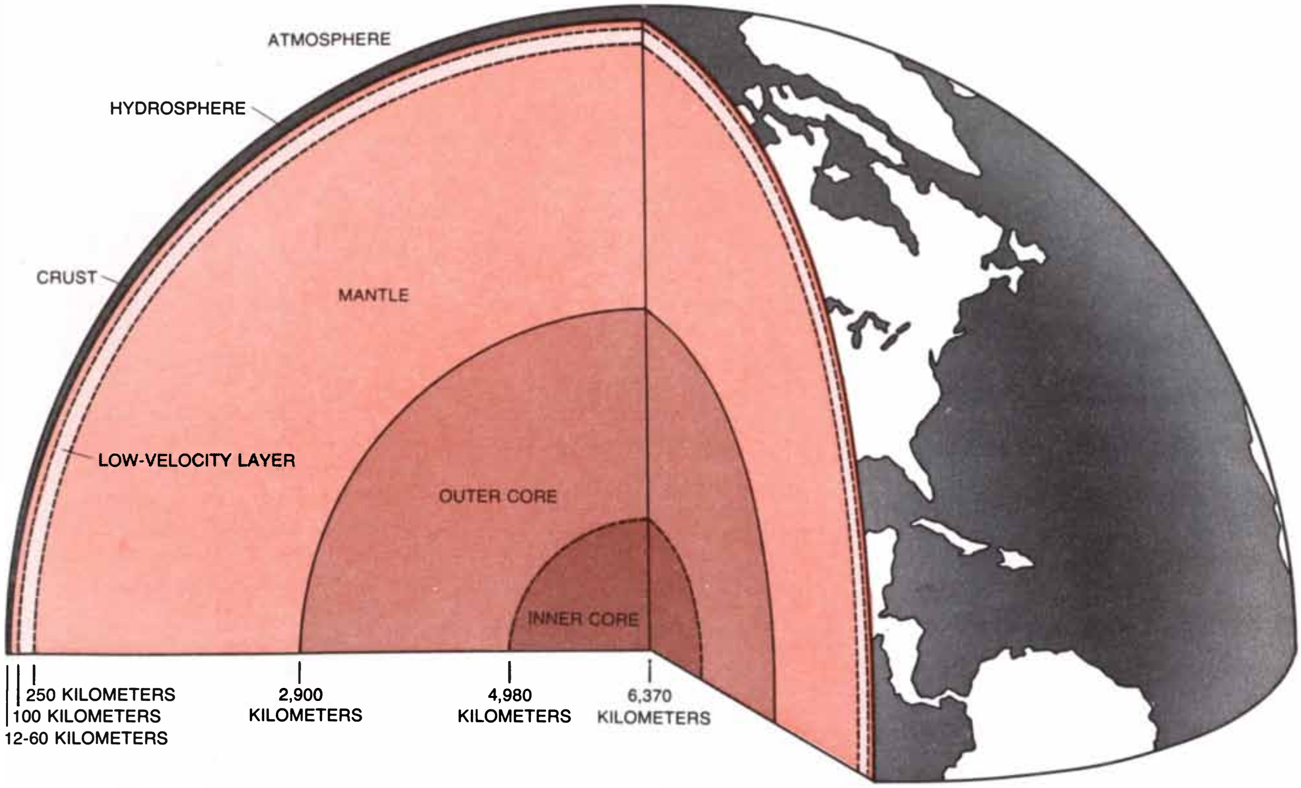
Plate Tectonics

This static picture of a concentric, layered earth is modified by the theory of plate tectonics, which deals with the movement of lithospheric plates. The lithosphere includes the crust and part of the upper mantle and is distinguished from the asthenosphere below it by the fact that it is cooler and therefore rigid. The theory of plate tectonics provides a dynamic picture of a mobile mantle, with plates of lithosphere about 100 kilometers thick moving laterally over the asthenosphere. The surface of the earth is covered by a few large lithospheric plates and several smaller ones. These shell-like plates move with respect to one another, and the geological activity represented by earthquakes and volcanoes is concentrated along the plate boundaries.

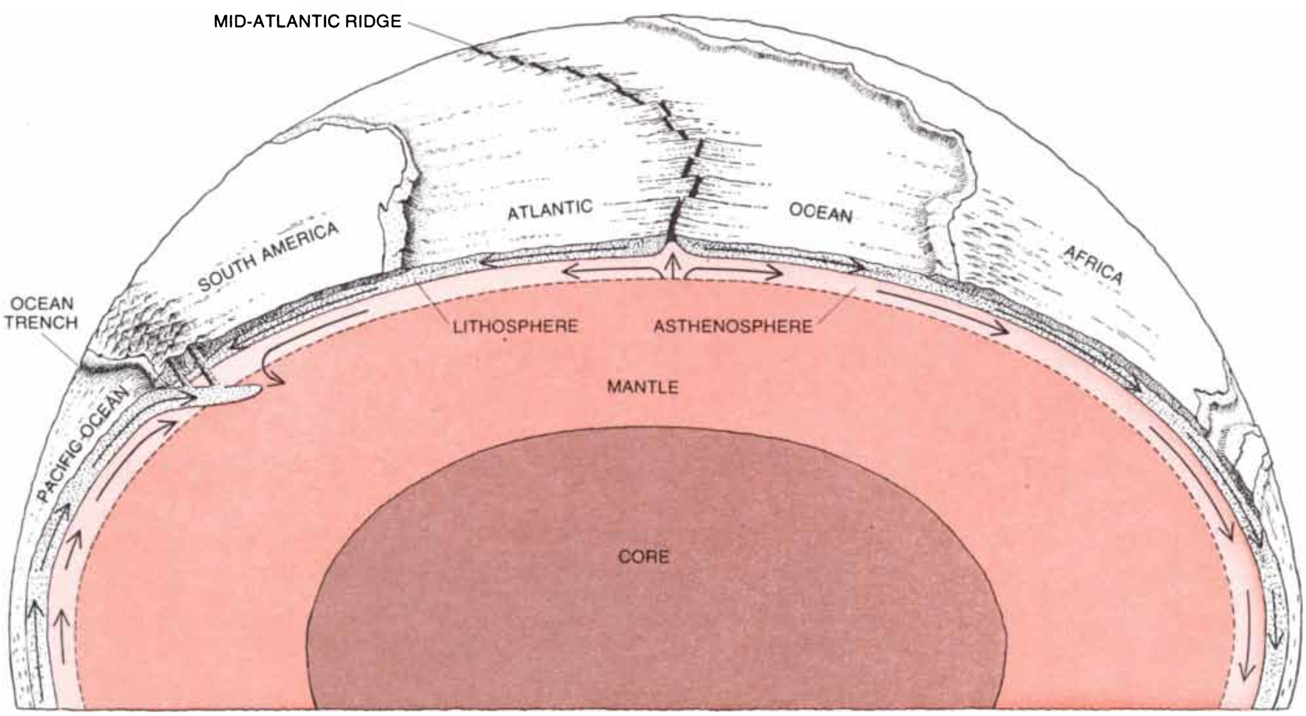
Plate boundaries include divergent and convergent types. Below the crest of an ocean ridge material from the asthenosphere rises, melting as it moves and thus producing lava that is erupted in the central rift valley of the ridge to produce new crust. Convective movements in the mantle cause the plates to diverge from one another as new lithosphere is generated.

At convergent boundaries plates may collide, pushing up crust to form mountains, or one plate may move under another, carrying lithospheric material back into the mantle. The growth of new lithosphere is therefore balanced by the destruction of lithosphere elsewhere. Such boundaries are associated with ocean trenches and with lines of volcanoes, including arcs of volcanic islands and the volcanoes in active mountain ranges.

Many of the physical properties of the earth as a whole have been determined. The planet's size, shape and mass have been measured with precision. Its known volume and mass give a mean density of 5.5 grams per cubic centimeter, which is much higher than the density of the rocks making up the accessible crust. Therefore a significant part of the interior must be composed of material



CONFIGURATION OF THE EARTH is portrayed by means of layers alone, without regard to the active processes that go on in the interior. The rocks of the thin crust are cool and rigid. Mantle rock, which is hot, is capable of slow movement. Evidence from earthquake waves indicates that the outer core consists of molten metal. Hydrosphere consists of surface and atmospheric waters.



DYNAMIC EARTH is depicted in cross section as it is envisioned in the theory of plate tectonics. Plates of the lithosphere, which includes the crust and part of the upper mantle, migrate laterally over the asthenosphere, which is a hot and perhaps partly molten layer of the mantle. Material from the asthenosphere rises below the crest of an ocean ridge, melting to produce lava that is erupted to form new crust in the ocean floor. The lithospheric plates diverge as new lithosphere is generated from the rising material. The growth of new lithosphere is balanced by the destruction of an equivalent amount of lithosphere at convergent plate boundaries, where the lithospheric layer moves down into the mantle. These boundaries are associated with ocean trenches and lines of volcanoes.

with a density higher than 5.5 grams per cubic centimeter. Studies of the earth's gravitational field and the physical properties of the rotating sphere indicate that the mass is concentrated toward the center.

Evidence on the Mantle

The main body of information about the physics of the earth comes from studies of earthquake waves, which provide the equivalent of X-ray pictures of the interior. The earth rings like a bell when it is shaken by a major earthquake. The vibrations of a bell depend on its shape and physical properties; similarly the vibrations of the earth as recorded by sensitive instruments can be interpreted in terms of the properties of the earth.

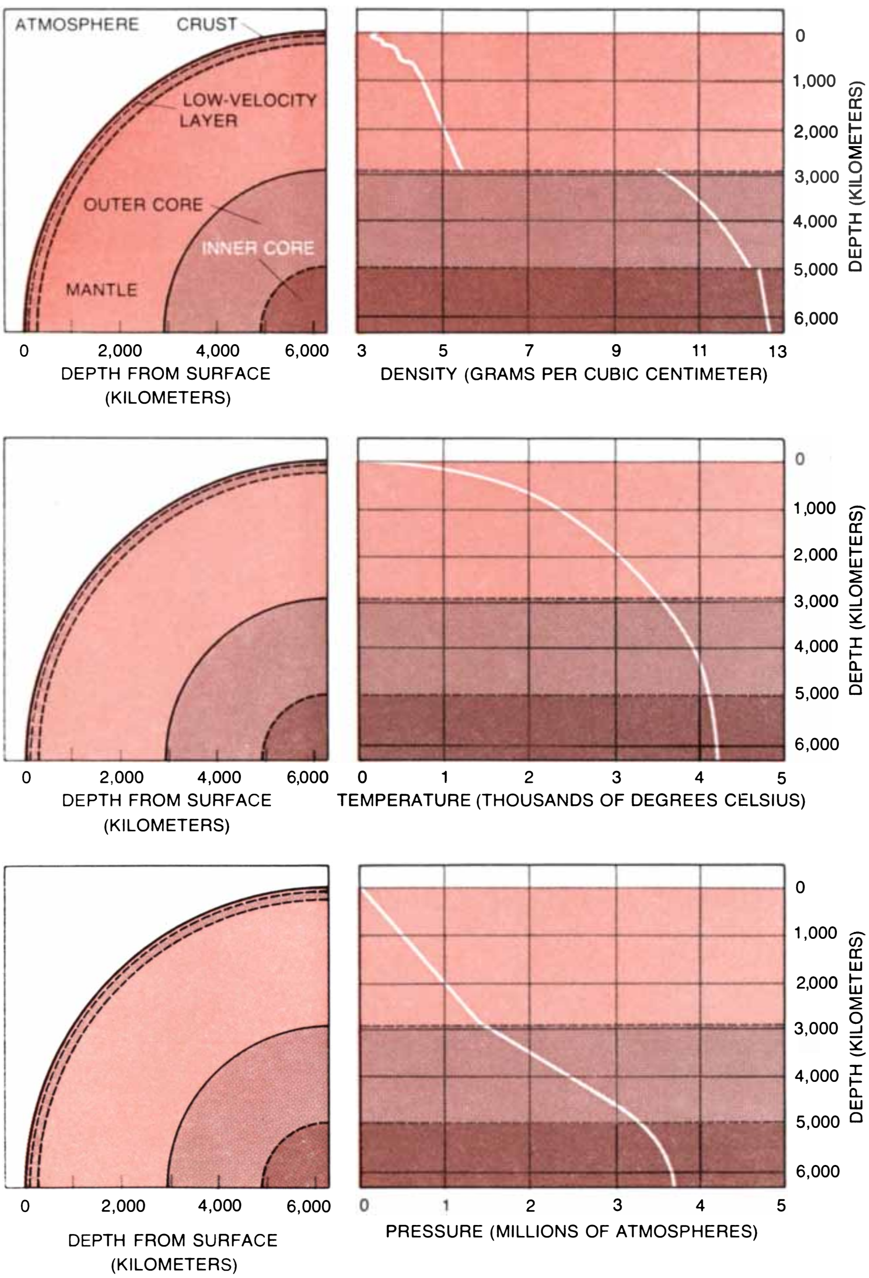
The release of energy at the focus of an earthquake produces several types of wave. The primary, or *P*, waves and the slower secondary, or S, waves pass through the interior of the earth. The abbreviations also serve as reminders that the energy is transmitted along the path of a ray by different phenomena: the *P* waves are compressional, or push-pull, waves and the *S* waves are shear, or "shake," waves [*see illustration on opposite page*]. *P* waves can be transmitted through both solids and liquids; *S* waves can be transmitted only through materials that can support shear stresses, that is, materials that can be deformed or bent. An *S* wave cannot be transmitted through a liquid, because liquids cannot sustain shear; they flow too easily.

If the earth were composed of material with uniform properties throughout, the waves from the focus of an earthquake would follow straight lines, and each type of wave would travel at a constant velocity. The times taken for *P* and *S* waves to reach a particular recording station at a known distance from the focus would give the velocity of each type of wave. In actuality, however, the results obtained from many earthquakes show that the waves travel faster through the earth than would be predicted from their known velocities in surface rocks. The results show in addition that the waves that travel the greatest distance have also traveled faster. These findings mean that the velocity of earthquake waves is greater at depth than it is near the surface and also that it increases progressively with depth.

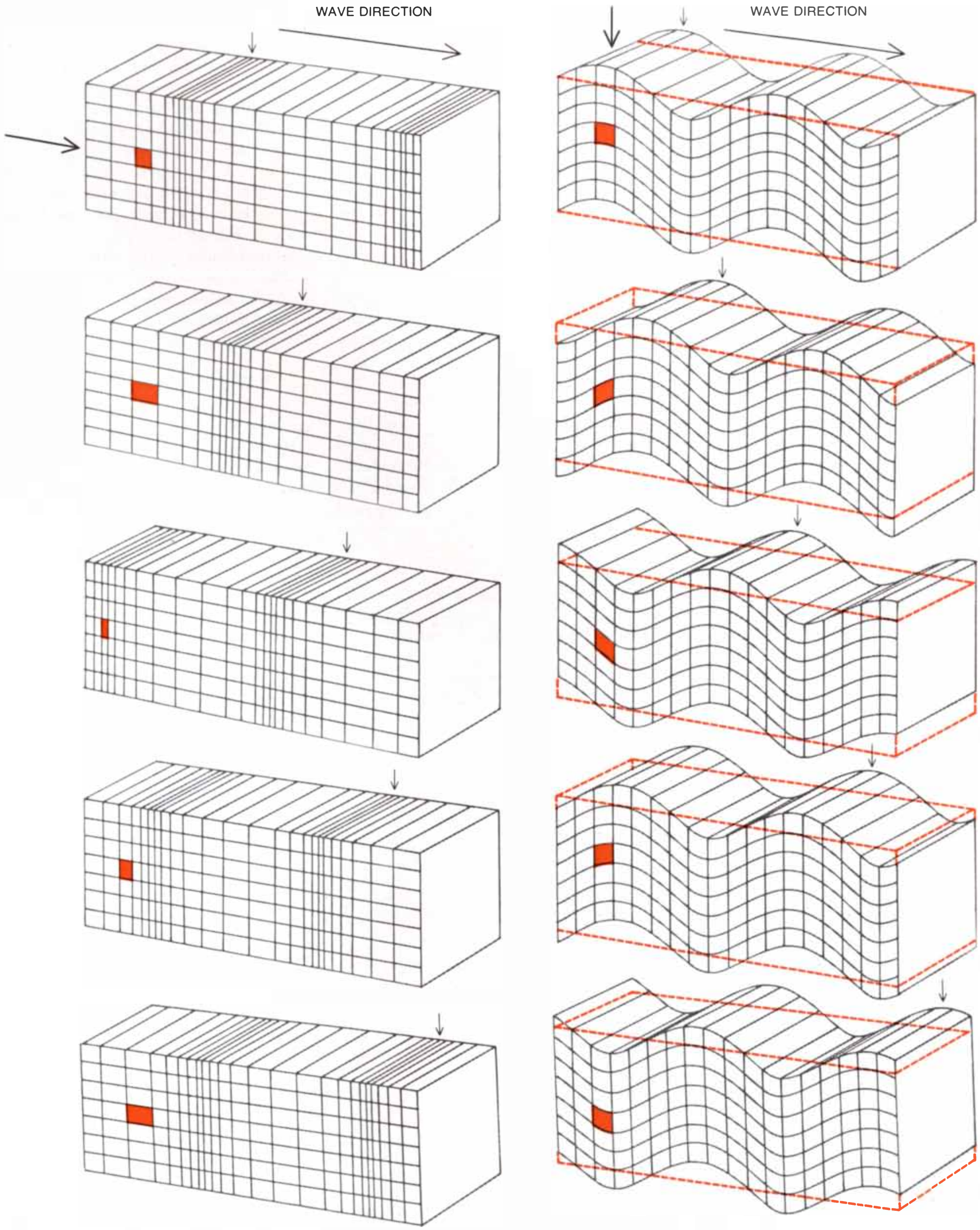
From these observations and others it is known that the waves from earthquakes are refracted and reflected within the earth. Refraction causes them to follow paths that are concave upward. Rays are reflected at levels where there is a distinct change in physical properties between layers. Reflection therefore occurs at boundaries: between the crust and the mantle, the mantle and the core and the inner and the outer core. This wave pattern demonstrates the concentric structure of the earth. *S* waves follow paths that reach as deep as the core-mantle boundary, but they do not pass through the core. This finding is evidence that at least the outer part of the core is liquid.

Measurement of the times of travel of the waves over various paths in the mantle provides a means of calculating the velocity at which the material at each depth within the mantle transmits waves. The wave-velocity profiles show a progressive increase with depth, but the increases occur in a series of steps down to about 1,000 kilometers. These observations indicate that the upper mantle has a layered structure.

The velocity of both *P* and *S* waves decreases in the upper mantle within a layer lying from approximately 100 to 250 kilometers below the surface. This



PHYSICAL PROPERTIES of the earth vary with depth from the surface. The depths on the graphs of density, temperature and pressure correspond to the depths on the cross-section diagrams at left. One atmosphere is the air pressure at sea level, 14.7 pounds per square inch.



EARTHQUAKE WAVES of two types pass through the interior of the earth from the focus of an earthquake, where energy is released. They are portrayed here passing through a block of rock. At left is a compressional, or *P*, wave, which is started by a sudden push or pull in the direction of the wave path. The action compresses the rock, and the nearby particles move forward. They then rebound to their former positions and beyond, continuing to vibrate in this way for some time. Each small volume of matter (*color*) contracts and expands as the crest of compression moves through the rock. A shear, or *S*, wave (*right*) is started by pressure at a right angle to the wave path. Particles of rock vibrate up and down. A small piece of matter (*color*) undergoes shear deformation. The wave crest transmits energy. The illustration is adapted from ***The Heart of the Earth***, by O. M. Phillips (Freeman, Cooper & Company).

is the low-velocity layer that I have mentioned. It is considered to be equivalent to the asthenosphere of the plate-tectonic model. The layer's physical properties, deduced from the wave-velocity profile and other geophysical evidence, are consistent with the presence of an interstitial liquid, which is generally thought to be a molten fraction of the rock in the layer.

Plate Boundaries

Earthquakes provide not only this static picture of the concentrically layered earth but also some aspects of the dynamic picture of plate tectonics. Since earthquakes occur only in rocks that are cool and rigid enough to fracture, the distribution of earthquake foci delineates the boundaries of the stable plates. The depths of the foci along the boundaries associated with ocean ridges are less than 100 kilometers, indicating that the mantle below the lithosphere is hot.

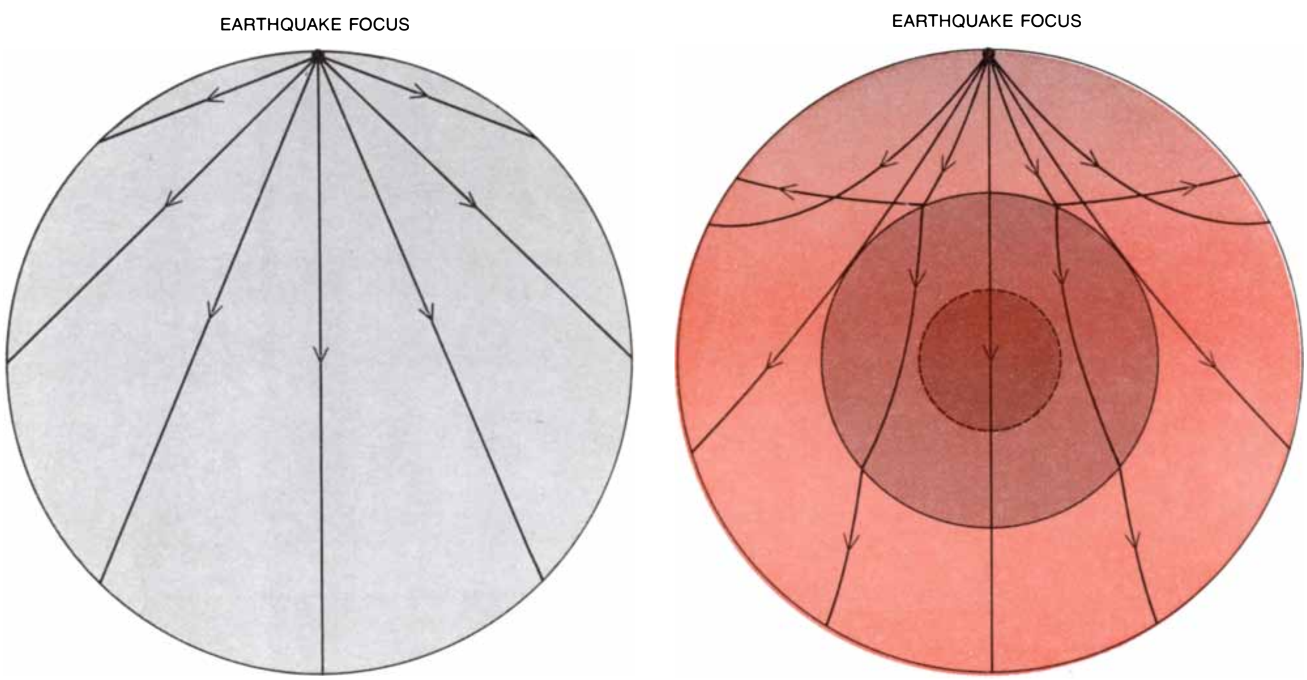
At the convergent boundaries where one plate sinks below another and moves into the mantle the foci of earthquakes are at depths as great as 700 kilometers. The distribution of these deep foci provides a means of mapping the layers of sinking lithosphere. Detailed studies of wave velocities in these regions are consistent with the existence of slabs of cooler lithosphere extending to considerable depths within the mantle.

The information provided by earthquake waves about the structure and physical properties of the mantle is fairly direct. In contrast, the effort to obtain information about the chemical properties of the deep interior requires resorting to mostly indirect evidence. The only chemical samples available are a few rocks in the crust that have been carried up from the topmost layers of the mantle. To form an estimate of the composition of the earth as a whole and of the core and mantle one must rely on the chemistry of extraterrestrial bodies, including stars, the sun and meteorites. The presumption is that such bodies have a composition essentially similar to the earth's. The approach involves the formulation of physical and chemical models for the origin of the solar system and the earth.

It is generally held that the solar system formed about 4.6 billion years ago through the gravitational collapse of matter that previously had been dispersed in interstellar space. A large precursor of the sun, a protosun, was surrounded by a thin, disk-shaped nebula of dust particles and gas. Local aggregation of particles and condensation of gas within the rotating nebula formed small objects that combined to produce planetary bodies. The material of meteorites was also formed during this period. Since the sun accounts for more than 99.6 percent of the mass of the solar system, its composition is effectively the same as that of the system as a whole. The abundances of the elements in the sun and other stars have been measured by spectroscopic methods, whereby each element is identified by its characteristic electromagnetic radiation.

What Meteorites Reveal

Meteorites now travel through the solar system in elliptical orbits that occasionally intersect the earth. There is evidence that they come from the asteroid belt: the swarm of small planetlike bodies orbiting at distances of 2.2 to 3.2 astronomical units from the sun. The meteorites vary widely in chemistry, mineralogy and structure, but for the purposes of this discussion it is sufficient to note the distinctions between two main groups: the iron meteorites and the stony meteorites. The iron meteorites consist essentially of iron-nickel



PATHS OF EARTHQUAKE WAVES through the interior of the earth provide information about its structure and the physical properties of its concentric layers. Energy from the focus of an earthquake is transmitted in all directions. If the earth had uniform properties throughout, the wave paths would follow straight lines (*left*) and the wave velocities would be constant. Measurements of the time of travel show, however, that the wave velocities, and therefore the physical properties, change abruptly at certain levels (*right*), thus revealing the concentric layers. *P* waves can be transmitted through both solids and liquids. *S* waves, which cannot pass through liquids, do not pass through the core of the earth, showing that at least the outer core is in the liquid state.

alloy, with the nickel content ranging from 4 to 20 percent; they also contain a small amount of iron sulfide.

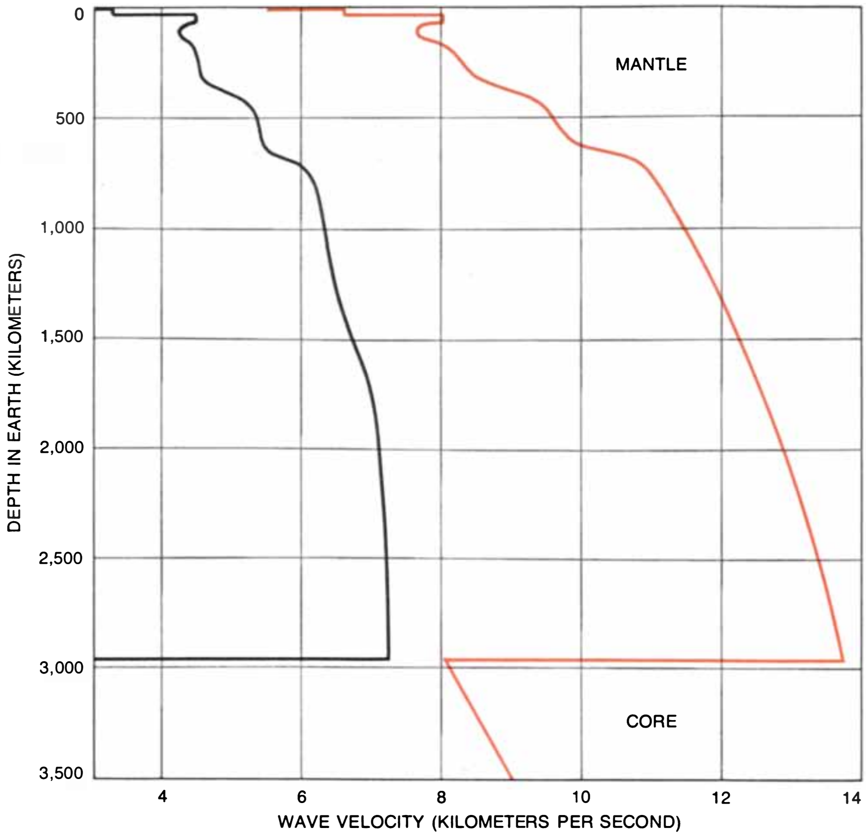
Stony meteorites are composed mainly of silicate minerals, together with various proportions of metal alloy and iron sulfide. The relative abundances of nonvolatile elements such as magnesium, silicon, aluminum, calcium and iron are about the same in many types of stony meteorite as they are in the sun and other stars. It is therefore argued that these abundances provide a good basis for estimating the overall abundances of elements in the earth and other planets.

In complex models of the origin and evolution of the solar system the composition of the earth is derived by starting with a volatile-rich stony meteorite and formulating a series of processes and chemical changes that could account for the other types of meteorite and for the present structure of the earth. The calculations yield estimates of the composition of the core and the mantle.

A simpler approach, which gives a similar result, is to select a specific group of stony meteorites and to assume that the average composition of the silicate portion is equivalent to the composition of the earth's mantle. The composition of the earth's core is estimated from the iron sulfide and an appropriate portion of the iron-nickel alloy in the same meteorites. The iron meteorites serve as a check on this calculation.

Whatever procedure is followed, the estimates of the composition of the mantle agree in the following respects: (1) More than 90 percent by weight of the mantle is represented by oxides of silicon, magnesium and iron (SiO_2, MgO and FeO), and no other oxide exceeds 4 percent. (2) The oxides of aluminum (Al_2O_3), calcium (CaO) and sodium (Na_2O) total between 5 and 8 percent. (3) More than 98 percent of the mantle is represented by these six oxides, and no other oxide reaches a concentration of as much as .6 percent. The concentrations of other elements, which are present in trace amounts, are not defined. The oxides are combined in various minerals within the mantle rock.

Of all the rocks found in the crust, only peridotites correspond to the estimates of the composition of mantle rocks made by studying extraterrestrial bodies. It is therefore to such peridotites that geologists interested in the mantle turn for details, including the concentration of trace elements. The problem, of course, is to make sure that the specimen of peridotite one is examining originated in the mantle rather than in the crust. For this reason particular interest attaches to the rounded boulders or nodules that are found in kimberlite "pipes" (cylindrical intrusions, with a diameter of a few hundred meters, that puncture the crust from the mantle in certain regions).



VELOCITY PROFILES of earthquake waves passing through the mantle are plotted for *S* waves (*black*) and *P* waves (*color*). The velocities are affected by the increase in pressure and temperature with depth, but the steps appearing at intervals down to a depth of about 1,000 kilometers correspond to changes in physical properties of the mantle. The decreased velocity between 100 and 250 kilometers is probably due to the presence of partly melted rock, which is an appropriate property for mobile asthenosphere of plate-tectonic model.

Kimberlites and Lavas

Kimberlites are famous as the rocks that bring diamonds to the surface of the earth. Diamond is a form of carbon that is stable only at very high pressures. One can therefore be confident that kimberlites originate in the depth interval from 150 to 300 kilometers below the surface—well within the upper mantle.

Evidence indicates that a kimberlite pipe originally rose rapidly through the crust as a fluidized system of solids, molten rock and gases, breaking through to the surface with a tremendous blast in a brief volcanic explosion. Fragments of rock ripped from the walls of the pipe and carried upward include nodules of peridotite and subordinate nodules of eclogite, another mantle rock. They are rounded and polished by repeated impact with other gas-driven fragments in the explosive pipe.

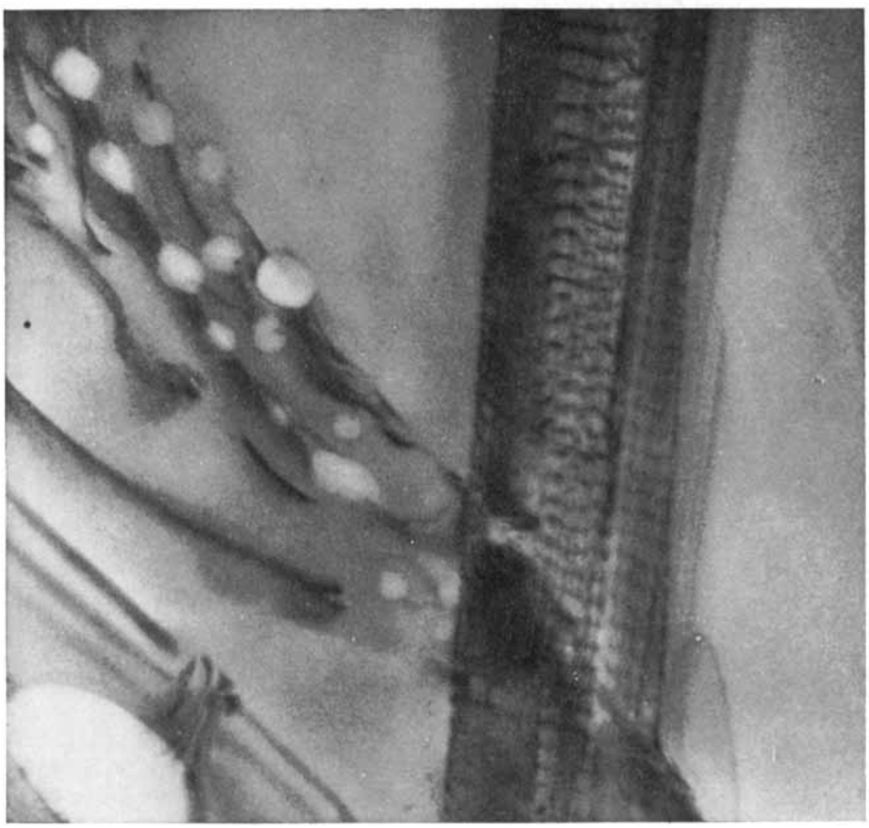
Suites of nodules similar to those in kimberlites are also found in certain volcanic lavas. In general they are derived from shallower levels of the mantle than the nodules in kimberlites. Many estimates of the composition of the mantle have been based on the mineralogy and chemistry of nodule suites in kimberlites and lavas. Peridotite rocks found in other geological environments have also served for this purpose, although with them it is particularly important to take into account the nature of the geological association with other rocks and the effect of geological processes in order to be sure that the judgments made on the basis of the evidence in the rocks truly relate to the mantle rather than to events in the crust.

A third approach, in addition to the study of extraterrestrial bodies and mantle rocks, is to calculate a hypothetical peridotite having the chemistry appropriate to yield the volcanic lavas that are derived by partial melting of the mantle. The melting gives rise to the basaltic magma erupted abundantly onto the surface from volcanoes and leaves behind in the mantle a residual peridotite, which can be presumed to lack the more fusible elements that melted and were carried off in the magma. Assigning an appropriate chemistry to the residual peridotite, one arrives at the hypothetical composition of the upper mantle. Pyrolite (pyroxene-olivine rock) is the name given to one of these hypothetical peridotites.

The Heterogeneous Mantle

It was once assumed on the basis of physical models that the upper mantle was homogeneous in composition. As we have seen, however, the detailed geophysical studies of recent years indicate a layered structure. Indeed, the nodules that have been studied confirm the notion that the upper mantle is quite heterogeneous, both chemically and mineralogically. The specimens sample the mantle from the mantle-crust boundary to the sources of kimberlite pipes, a depth interval that could reach 250 kilometers. One could hope to find among the nodules specimens of original mantle peridotite that has never been melted; specimens of depleted residual peridotite from which a molten fraction of the more fusible components has been removed; specimens of the molten fraction that failed to escape to the surface in magma, instead crystallizing at high pressure in the mantle to form eclogite; specimens intermediate between these types, and specimens involving the same mineral assemblage but formed by other processes too complex for consideration here.

Estimates of whole-mantle composition based on the examination of extraterrestrial bodies yield values similar to those obtained from the study of rocks originating in the upper mantle. Together the estimates support the idea that the chemistry of the mantle does not change much from top to bottom. In comparing estimates, however, one finds that the potassium content of the upper mantle according to the hypothetical pyrolite is considerably higher than the amount of potassium found in peridotite rocks derived from the mantle, and both are much lower in potassium than the estimate derived from studying extraterrestrial bodies.



CARBON DIOXIDE BUBBLES appear in this electron micrograph of a specimen of peridotite. The enlargement is 20,000 diameters. The bubbles, which appear in discontinuities in the crystal structure of the rock, indicate the presence of carbon dioxide in the mantle. Micrograph was provided by Harry W. Green II of the University of California at Davis.

Large uncertainties also remain about the concentration and distribution of other trace elements and of volatile components such as water and carbon dioxide. These components are of fundamental significance for such factors as the generation of heat by radioactive decay (of such elements as uranium and thorium and the radioactive isotope of potassium, potassium 40), melting temperature (in the presence of small amounts of water at high pressure rocks begin to melt at temperatures lower than the melting point of dry rock) and the physical strength of the mantle (where the strength of rock would be significantly reduced by the presence of small amounts of molten material between mineral grains and by interstitial bubbles of gas).

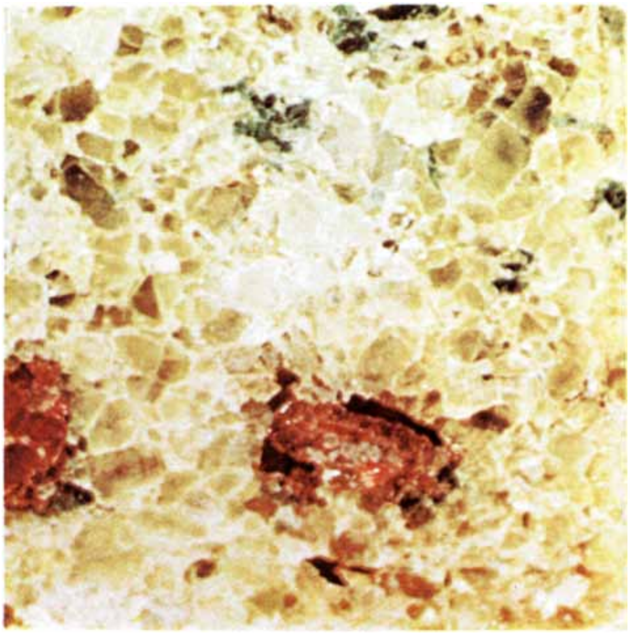
Water and Carbon Dioxide

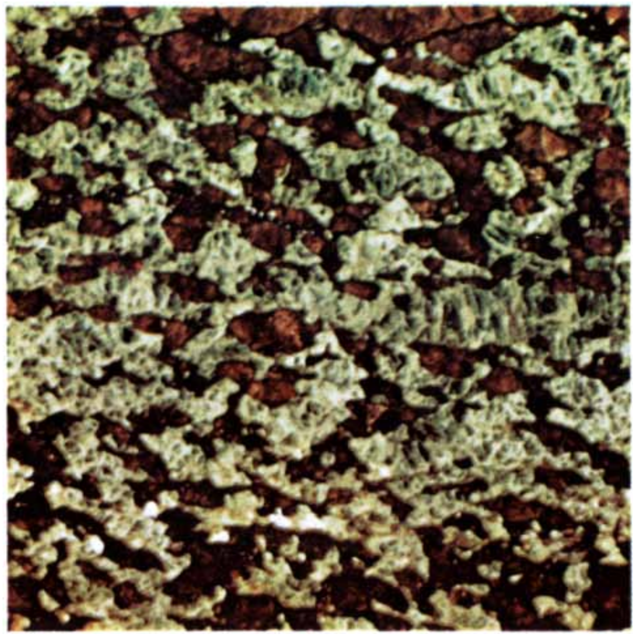
In certain nodules of peridotite one finds the minerals phlogopite and amphibole, which are hydrous, that is, incorporate water. This finding is accepted as evidence for the existence of water in at least some parts of the upper mantle. It is doubtful that the amount could exceed .1 percent by weight, and it is probable that the distribution of the water is not uniform.

Examination under the microscope has shown that the crystals of olivine and pyroxene in some peridotite nodules from kimberlites and lavas are crowded with tiny cavities up to five micrometers in diameter. Many of them are filled with dense, liquid carbon dioxide trapped at high pressure. This finding indicates the presence of carbon dioxide in some form in the upper mantle.

High-voltage electron microscopy has recently yielded remarkably detailed pictures of crystal defects: the discontinuities in crystal structure that occur within individual grains of minerals. The defects cannot be seen with light microscopes. In the electron micrographs minute bubbles of carbon dioxide are abundant along the discontinuities in the olivine and pyroxene of some peridotite nodules. This evidence suggests that the carbon dioxide was originally dissolved in the solid minerals and was exsolved and precipitated as gas bubbles because of elastic strain near the crystal defects.

The independent determinations of the physical properties and chemistry of





ROCKS FROM MANTLE are a mica-garnet peridotite (*left*) and an eclogite (*right*). The peridotite consists mainly of olivine (*yellowish green*) but also contains garnet (*red*) and orthopyroxene and clinopyroxene (*bright green*). A sample including mica, a mineral with water bound in its structure, is rare. This sample is from Tanzania. The eclogite, which is from a mine in South Africa, is composed of garnet (*red*) and a clinopyroxene (*green*). The minerals are arranged in layers. Mantle rocks are brought to the surface in kimberlite "pipes" and in certain volcanic lava. The photographs were provided by J. B. Dawson of the University of St. Andrews.

the mantle must be consistent with one another. In order to test their consistency one needs to know the physical properties of the estimated composition of the mantle through the range of pressure and temperature existing in the mantle. If one can determine the way the mineralogy varies as a function of pressure and temperature, one has a means of estimating the depth from which a rock sample came and the temperature at that depth when the rock was formed or reached equilibrium with its surroundings.

Pressure and Temperature

The structure of silicate minerals is dominated by the packing of oxygen atoms. The other atoms, which are much smaller, occupy spaces between the oxygens. At low pressure each silicon atom is surrounded by four oxygens whose centers form a tetrahedron; the silicon is said to be in a fourfold coordination. At much higher pressure the oxygen atoms are squeezed closer together. They adjust into a densely packed arrangement with silicon atoms in a sixfold coordination. This readjustment of mineral structure is a phase transition. The steplike changes in value of the physical properties in the mantle are caused by successive phase transitions.

Starting with a material of fixed composition, such as the peridotite thought to exist in the mantle, the phase transitions depend on pressure and temperature. A diagram of the transitions for peridotite has been determined by experiments in laboratory apparatus that carried the pressure up to 200 kilobars (200,000 atmospheres, equivalent to the pressure 600 kilometers below the surface of the earth). The diagram has been extended to higher pressures by indirect methods.

It is known from the nodules in kimberlites and lavas that peridotite in the mantle can crystallize in at least three mineralogical assemblages: plagioclase peridotite, spinel peridotite and garnet peridotite. The high-pressure experiments demonstrate that the assemblages are related through phase transitions. With increasing pressure plagioclase peridotite is transformed first into spinel peridotite and then into garnet peridotite [*see illustration on next two pages*].

Experimental studies show that at still higher pressure garnet peridotite undergoes a phase transition involving an increase in density of almost 10 percent; the dominant olivine of the upper mantle is transformed into a spinel-like material, and the aluminous pyroxene is transformed into a garnet structure that combines in solid solution with the garnet already present. At pressures near 200 kilobars the minerals are further compressed into structures with all the silicon atoms in the sixfold coordination, giving rise to minerals that are unknown at the surface of the earth. This compression results in another increase in density of about 10 percent. The actual pressure at which a phase transition occurs increases with higher temperatures.

At any fixed depth an increase in temperature eventually brings a rock to a point where it begins to melt. This temperature increases with pressure, as shown by the boundary labeled solidus in the phase diagram [*next two pages*]. A rock composed of several minerals melts progressively through an interval of temperature in which solid crystals coexist with liquid. Complete melting is marked by the boundary labeled liquidus.

The effect of temperature can be studied by means of a geotherm, which is a line giving the temperature at each depth in the earth. If the line is drawn on the phase diagram for peridotite, each point on the line occupies one of the phase fields and thus also defines the mineral assemblage for the peridotite at each depth. A cross section through a hypothetical mantle composed of peridotite is constructed by following a geotherm through the phase diagram. Each layer consists of a particular mineral assemblage.

The boundaries between layers of the mantle are presumably at depths where

Some of the metals we mine are more precious than gold.

An ounce of cold steel can do wonders for a warm smile.

But it must be a very special kind of steel. Strong, yet ductile. Hard, yet smooth.

It must not rust or stain. And it must remain unchanged through ice-cold sodas and red-hot pizzas.

At Union Carbide we mine or process the alloying metals that make possible this special steel and hundreds of others.

We produce over 60 different alloys and alloying metals.

Manganese, tungsten, silicon, chromium, vanadium.

Not exactly household words. But—combined with iron, aluminum and other metals—they have transformed the world we live in.

If it weren't for alloys there would be no high-strength steels for buildings and bridges.

No jet engines or aircraft bodies.

No sophisticated tools.

No electric motors for shavers, typewriters or vacuum cleaners.

No computers, or lightbulbs, or television sets.

When you think of them this way, our alloying metals become very precious indeed.

Whether they're as far away as a rocket on its way to the moon. Or as close to your heart as a brace shaping a beautiful smile.

UNION CARBIDE

Today, something we do will touch your life.

the geotherm crosses phase boundaries. It turns out that these boundaries correspond closely to the depths where the velocity of seismic waves changes. This finding is regarded as being good evidence that the upper mantle does have a composition close to that of the hypothetical peridotite and that the layered structure of the upper mantle is caused by transitions of phase rather than by changes in composition.

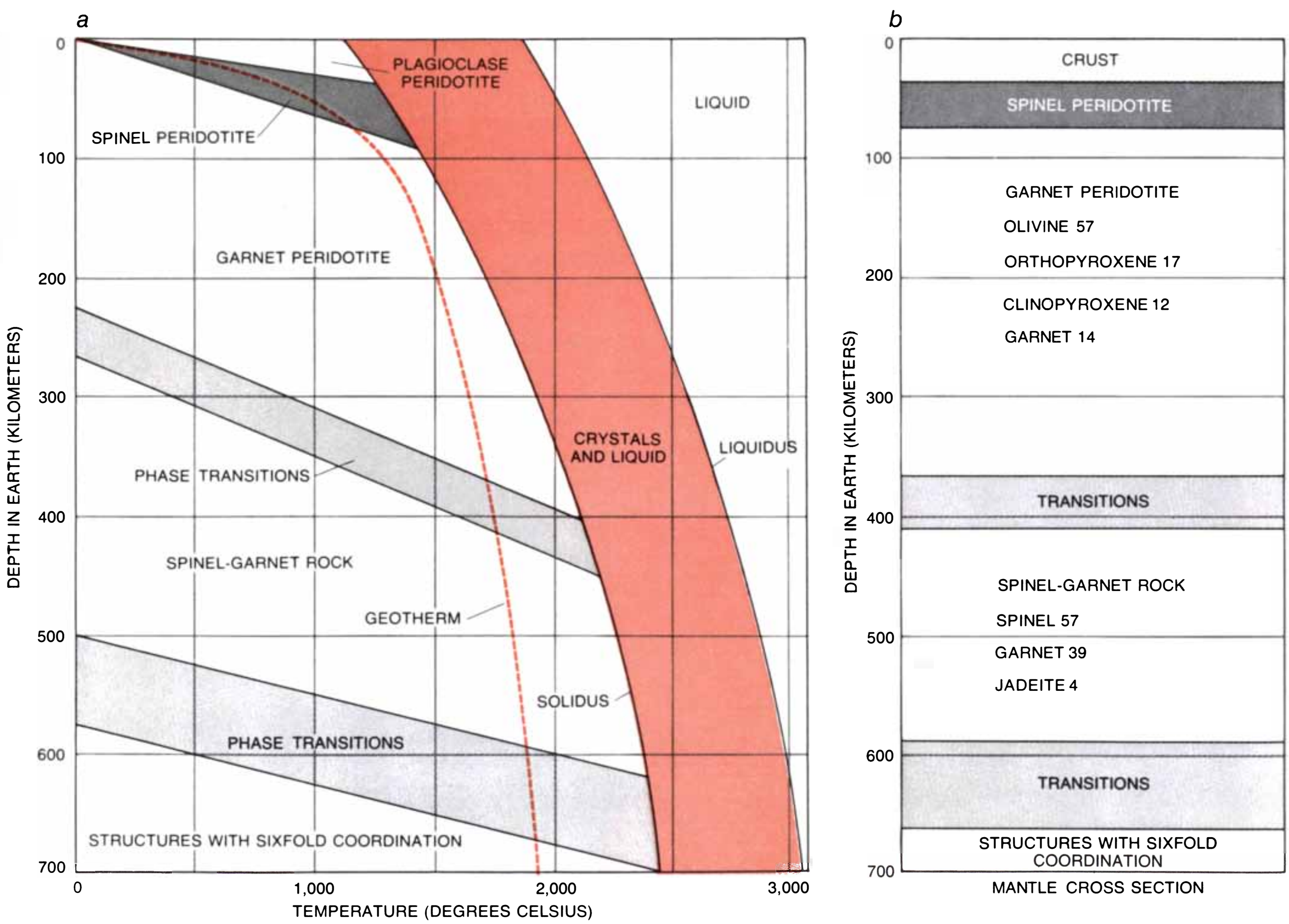
The decrease in the velocity of earthquake waves in the low-velocity zone can be explained by the presence of water or carbon dioxide in the upper mantle. Either one could cause a trace of melting in the peridotite of the upper mantle. The result is a change in the physical properties of the rock. If no water were present, a similar effect could perhaps be produced by intergranular carbon dioxide.

During the first half of this century it was widely believed among geophysicists that convection, the upward flow of hotter material and the downward flow of cooler material, could not occur in the solid, rigid mantle. That was one of the reasons the theory of continental drift failed for so long to gain many adherents. Recently, however, a number of models have proposed convection in the mantle as the driving mechanism for the migration of lithospheric plates. Details of the movements in the mantle and the scale of convection remain uncertain, but there is little doubt that the rates are exceedingly slow—so slow that the mantle is effectively motionless within the normal human framework of time.

The lithospheric plates and the continents that ride on them drift at the rate of a few centimeters per year. Suppose the driving material in the mantle moves at a rate of five centimeters per year, which is equivalent to about .005 millimeter per hour. The tip of the hour hand on a typical household clock moves five centimeters per hour, and the movement is not directly apparent to the human eye. Yet the speed is 10,000 times faster than the proposed movement in the mantle. Even so, a movement of five centimeters per year adds up significantly over geologic time; a parcel of rock could move from the bottom of the mantle to the top in 58 million years, which is only a small fraction of the earth's age of 4.6 billion years.

Movements of the Mantle

How is it possible that solid rock can flow, even so slowly? When a blacksmith picks up a bar of cold steel, he is unable to bend it, but when he heats it to a red glow, he can bend it easily, even though it is still a solid bar. Similarly the rocks of the mantle, which are at

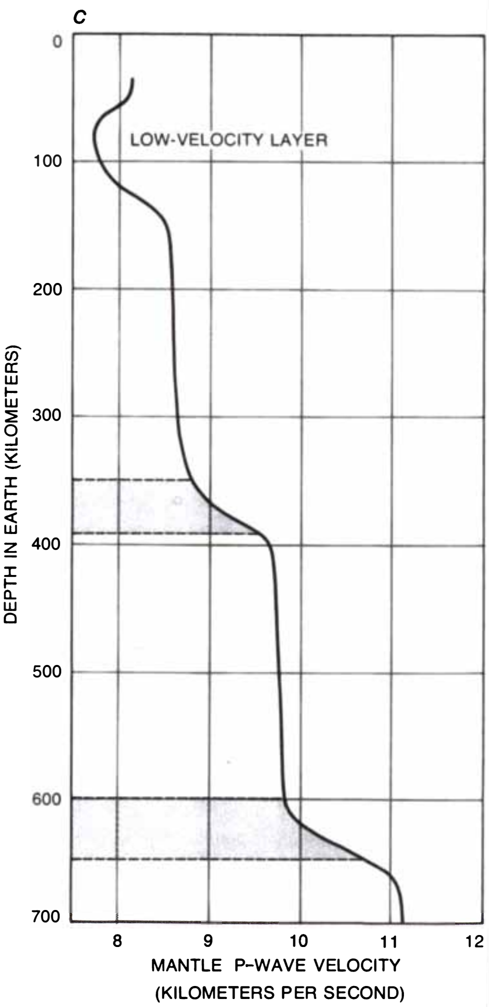


PHASE DIAGRAM for peridotite is stated in terms of pressure, or depth, and temperature. At a fixed depth an increase in temperature eventually brings a rock to a point where it begins to melt. The beginning and end of melting are bounded by the solidus and liquidus curves. Reading downward in the two parts of the diagram at left, following the geotherm (*a*), one sees that the upper mantle has a layered structure owing to successive phase transitions from spinel peridotite in the uppermost region to a rock composed of minerals with elements in a sixfold coordination, that is, having closely packed groups of oxygen atoms enclosing atoms of magnesium, iron and silicon, at a depth of about 600 kilometers. The numerals in the cross-section diagram (*b*) represent the percentages

high temperatures, can be deformed even though they are still in the solid state.

Olivine, pyroxene and peridotite have been subjected to strains at high pressures and temperatures in laboratory experiments. They deform. The deformed products have recently been studied in the high-voltage electron microscope in attempts to establish the mechanisms of plastic flow. A suggestion emerging from this work is that deformation occurs first in individual crystals, which then recrystallize to form a mosaic of new grains. The mechanism of flow varies as a function of pressure and temperature.

Schemes have been proposed for large convection cells extending through the entire thickness of the mantle [*see illustration on page 63*]. An alternative model, based on the argument that the absorption of heat in phase transitions would reduce the driving force, confines



of minerals in the rocks. The profile of *P*-wave velocity (*c*) shows that the depths of actual changes in the mantle indicated by earthquake waves coincide closely with the depths of phase transitions as inferred from the phase diagram and the geotherm.

convection to the upper mantle, above the olivine-spinel transition. A third model limits convection to the asthenosphere, which as a layer of the mantle lies at depths of from 100 to 300 kilometers. Subduction zones, however, where lithospheric material returns to the mantle and thus is involved in convective forces, appear to run as deep as 700 kilometers. Another model for the plate-driving mechanism involves thermal "plumes" in the mantle. According to this argument, all upward movement of mantle material is confined to about 20 plumes, each plume a few hundred kilometers in diameter, rising from the core-mantle boundary. The return flow is accomplished by a slow downward movement of the rest of the mantle.

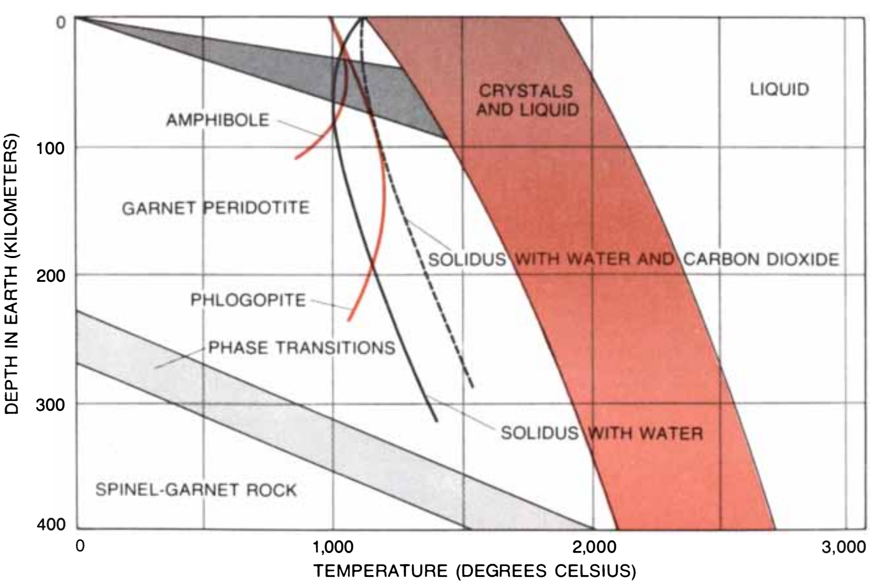
Where a plume reaches the lithosphere the flow becomes horizontal, spreading radially in all directions. A plume creates a hot spot with volcanic activity at the surface, and it may cause an upward doming of the lithosphere. In such ways the plumes would cause the movement of the lithospheric plates.

The hot-spot hypothesis is currently a hot idea. Many geologists are exploring its implications for various phenomena, including the features of volcanic island chains such as Hawaii. The hypothesis is also under challenge, however, by earth scientists who express doubt that a plume could retain its coherence as it rises through 2,800 kilometers of mantle.

The vertical movement of mantle material in any kind of convective system causes changes in the temperature distribution; at a given depth the temperature is increased where hotter material is rising and decreased where cooler material is sinking. The movement changes the shape of the geotherm curve from time to time and from place to place as the convection proceeds. The compositions of the minerals in mantle peridotite vary as a function of pressure and temperature, as laboratory experiments dealing with the phase diagrams of peridotite and its constituent minerals have established.

A parcel of mantle peridotite is subjected to changes in temperature and pressure as a result of convective motions in the mantle, and its mineralogy will adjust by recrystallization as it strives to attain equilibrium in the changing environment. The movements are slow enough so that an equilibrium mineralogy is normally attained. If the rock is suddenly transported to the surface, however, as in a kimberlite eruption, the time for readjustment is not enough and the sample reaches the surface with the mineralogical signature corresponding to its position and temperature where it last reached equilibrium in the mantle.

The compositions of coexisting minerals in peridotite at any particular pressure and temperature have been measured directly in the laboratory experiments. Using these data for the purpose of calibration, it is now possible to take samples of mantle peridotite, such as the nodules from kimberlites, to measure the composition of the minerals in each sample and thus to estimate the pressure (or depth) and temperature at their source in the mantle.

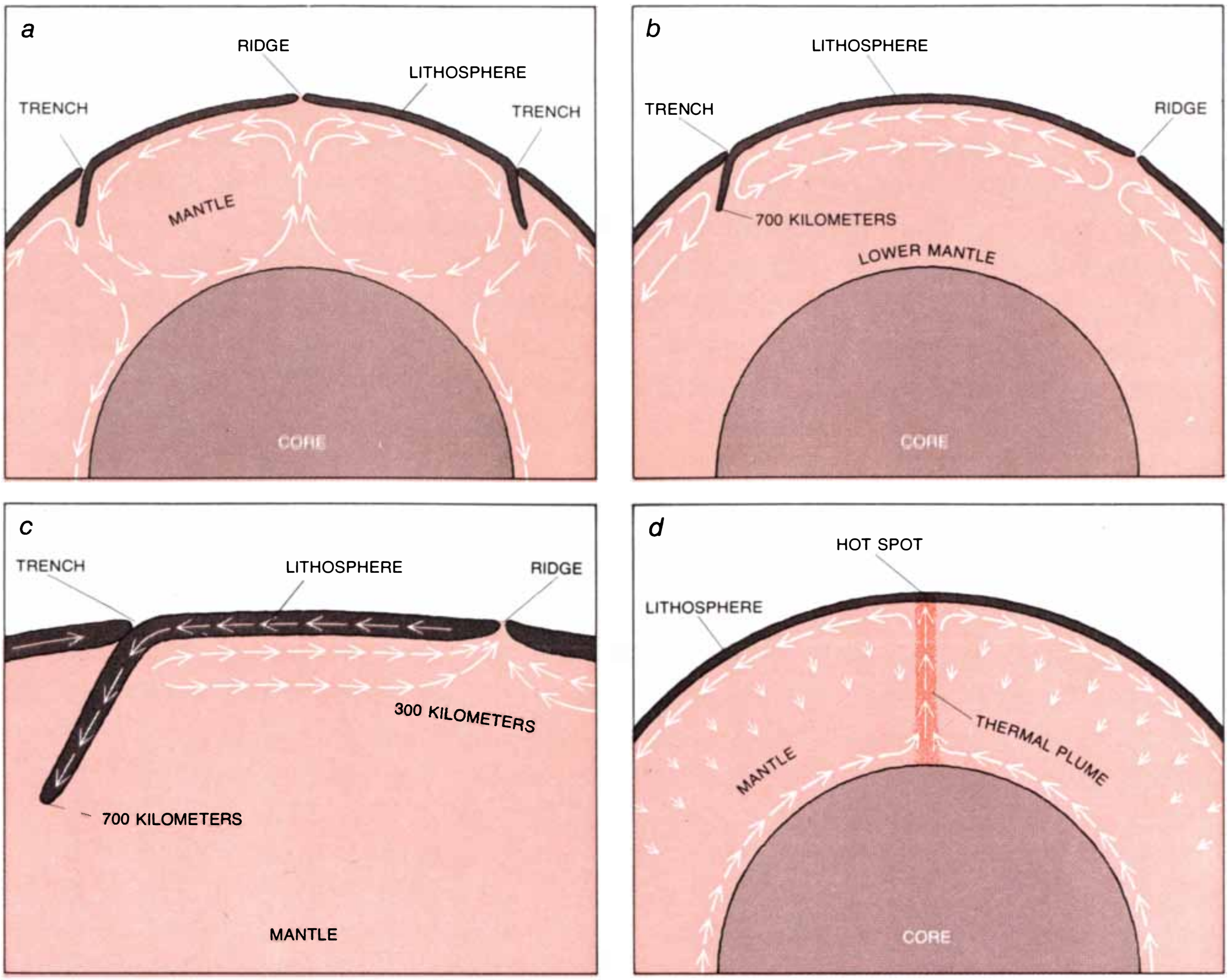


LOW-VELOCITY LAYER in the uppermost mantle appears to be explained by the presence of water and carbon dioxide, which lower the solidus curve, so that rock melts at lower temperature and pressure. The hydrous, or water-bearing, minerals amphibole and phlogopite are stable through a limited temperature range (*color*). Comparing this diagram with the one for dry mantle rock, one sees that the geotherm curve passes from a phase of solid peridotite to a phase of partly melted peridotite at a depth of about 100 kilometers, which corresponds closely with the top of the low-velocity zone observed in the uppermost mantle.

Ancient Geotherm

The application of these methods to nodules from kimberlites has produced intriguing results during the past year or two. A suite of nodules collected from a single kimberlite pipe gives a series of points, each specified for a given nodule by the estimated pressure and temperature of equilibrium before eruption. The locus of these points on a diagram of depth (pressure) and temperature corresponds to the geotherm that existed at the time of eruption of the kimberlite. In other words, each kimberlite nodule contains stored within its mineralogy the record of its equilibrium pressure and temperature in the mantle before it was abruptly carried upward. The experimental mineralogists engaged in this work have found that the results from kimberlite pipes in South Africa give fossil geotherms with normal gradients down to depths of 150 kilometers or so but with steeper gradients at deeper levels; apparently the temperature was higher than normal at those depths for some interval of time before the kimberlite pipes erupted nearly 100 million years ago.

These results are of great interest to geophysicists trying to establish the thermal history of the earth, the dynamics of the mantle and the driving forces of plate tectonics. One interpretation of the data is that the inflection point in a fossil geotherm corresponds to the top of the asthenosphere at a time about 120 million years ago when the African lithospheric plate began to move rapidly as the Atlantic Ocean opened up. According to this interpretation, frictional heating in the asthenosphere as a result of movement shifted the geotherm to higher temperatures, producing an inflection point at the lithosphere-asthenosphere boundary, whereas the conduction rate of heat through rocks is so slow that the original geotherm in the lithosphere remained effectively unchanged. Another interpretation is proposed by geophysicists who conclude that friction in the asthenosphere is not capable of producing such a large thermal effect. They argue that the steepened portion of the fossil geotherm below the inflection point could be caused by the upward convection of a local



MODELS OF CONVECTION have been proposed to explain how activity in the mantle drives the lithospheric plates. In convection warmer material moves upward and colder material moves downward. One model (*a*) holds that convection cells extend through the entire mantle. According to a second model (*b*), they are confined to depths above the phase transition from spinel to olivine. A third model (*c*) confines movements of the mantle to the asthenosphere. In the thermal-plume model (*d*) all upward movement is confined to a few thermal plumes, and the downward flow is accomplished by slow movements of the remainder of the mantle.

thermal plume, which initiated the eruption of the kimberlite.

This example illustrates the way plate tectonics, with its focus on the mantle, has brought together into the same symposium rooms at scientific meetings research workers in areas that once were considered quite distinct from one another. Field geologists, mineralogists, geophysicists and experimental chemists and physicists have together discovered in the minerals of kimberlite nodules information that is grist for the mills of the theoreticians who are trying to work out how the mantle moves now and has moved through the 4.6 billion years of the earth's existence.

A view of the entire earth from a spacecraft shows a large, spinning globe as smooth as a billiard ball. The deepest hole drilled in the surface reaches a depth of only nine kilometers; it is a mere pinprick, penetrating less than .15 percent of the earth's radius. It is therefore quite remarkable that so much is known about the inaccessible mantle.

Uncertainty of Models

Nonetheless, the amount of information now in hand is insufficient for a full understanding of the dynamic behavior of the mantle, which is the key to many geophysical and geological phenomena. The extent of the uncertainty about what is really happening in the mantle can be illustrated by comparing two hypotheses about the Hawaiian volcanic island chain. Each island formed as a result of eruption above a melting region fixed in the asthenosphere. Then the moving lithospheric plate carried the island away, creating over a long period of time the island chain.

According to one interpretation of these events, the melting is localized in a hot spot above a thermal plume. According to another interpretation, the melting is localized by friction in rocks flowing from the asthenosphere into a column moving downward through the mantle. The column is said to form a gravitational anchor, maintaining the region of downward flow in a position more or less directly above it.

These diametrically opposed hypotheses are put forward by respected earth scientists. Is Hawaii to be explained by a rising thermal plume or by a sinking gravitational anchor? I am confident that before too many years have passed the accumulation of additional evidence and the refinement of hypotheses will have placed much closer constraints on the picture that can be drawn of the structure and dynamics of the mantle.

VISUAL PIGMENTS AND COLOR BLINDNESS

Color vision depends on three types of cone cell, each containing one of three visual pigments. Those who are color-blind either lack one of the pigments or have an anomalous pigment in one type of cone

by W. A. H. Rushton

"See that little boy; he's blind. Let's put some obstacle in his way and watch him stumble over it." It is almost unthinkable that anyone would play such a trick on a blind person, yet much the same is often done to people who are color-blind. "This little boy is color-blind. Let's ask him the color of these objects and enjoy his confusion when he names the green one red." I mention this because I have been told by my friends, the color-anomalous subjects of my experiments, how they were teased at school and how glad they are to cooperate in a scientific study in which their abnormality is not a joke but a precious pathway to a better understanding not only of abnormal color vision but also of normal color vision.

People with defective color vision can nearly always see a range of colors, particularly along the yellow-blue axis. Thus they tend to resent the implication that they are somehow blind. Their deficiency almost always shows up as abnormal red-green discrimination. The condition is a sex-linked recessive genetic character; therefore the defect is much rarer in women, who have two gene-bearing sex chromosomes (the *X* chromosomes), than it is in men, who have no second chance if they lack the necessary gene on their single *X* chromosome. Among men 8 percent have defective red-green vision; among women only .4 percent show the defect.

Two types of red-green defective are recognized, both of which are further subdivided. One type, Type *A*, consists of dichromats who can match any spectral color by a suitable mixture of lights of two colors, for example red plus blue (without the need for green, which normal eyes require). Such dichromats are monochromatic in the red-green band of the spectrum; hence they can match any wavelength there (that is, in the band between 540 and 700 nanometers) with any other by adjusting only the relative brightness.

Dichromats are divided into two classes: A_1, called protanopes, and A_2, called deuteranopes. Both will say that any wavelength from the red part of the spectrum will match a given green, but protanopes require that the red light be about 10 times brighter than the red selected by the deuteranopes to make the match [*see top and middle illustrations on opposite page*].

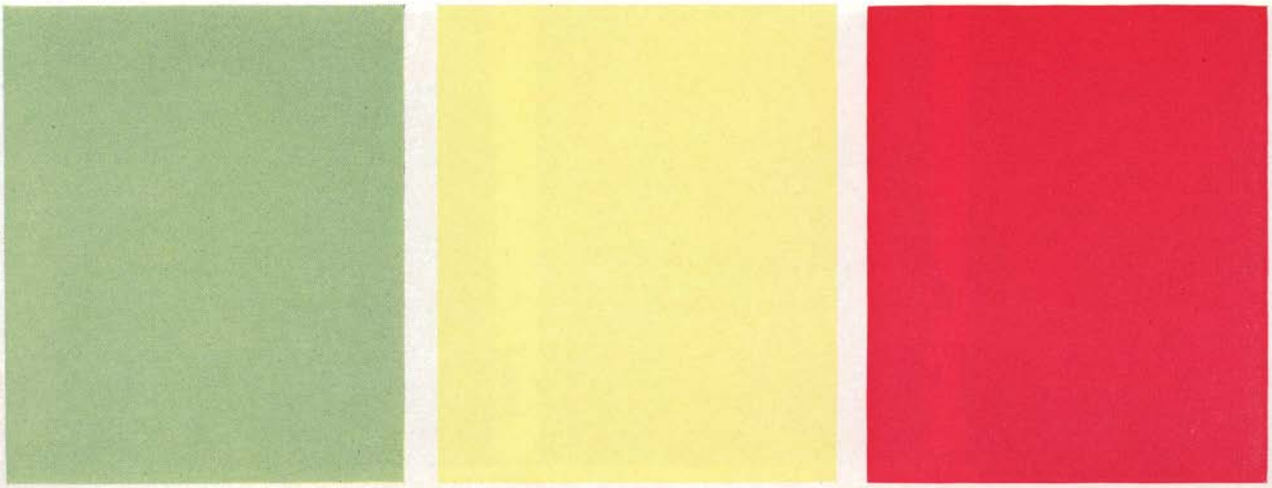
The other type of red-green defective, Type *B*, is the anomalous trichromat. Defectives of this type resemble normal people in needing three colors (for example red, green and blue) to match all spectral colors, but they insist on abnormal proportions of red and green in their mixtures. They fall into two classes according to the proportions of red and green in the mixture that they match with yellow. Protanomalous trichromats (Type B_1) introduce too much red by normal standards; deuteranomalous trichromats (Type B_2) introduce too much green [*see bottom illustration on opposite page*].

The anomalous trichromats are not just different from normals; they are indeed somewhat defective. They cannot distinguish nearly as many hues in the spectrum as normals can, and they are confused by autumn tints: the browns and olive greens. Our experiments have shown that in all these defective conditions what is abnormal is one of the photosensitive pigments in the cone cells of the retina. By means of special techniques, which I shall describe, we have been able to measure the spectral sensitivity in the red-green range of each cone pigment in normal eyes and in the eyes of subjects with various color defects.

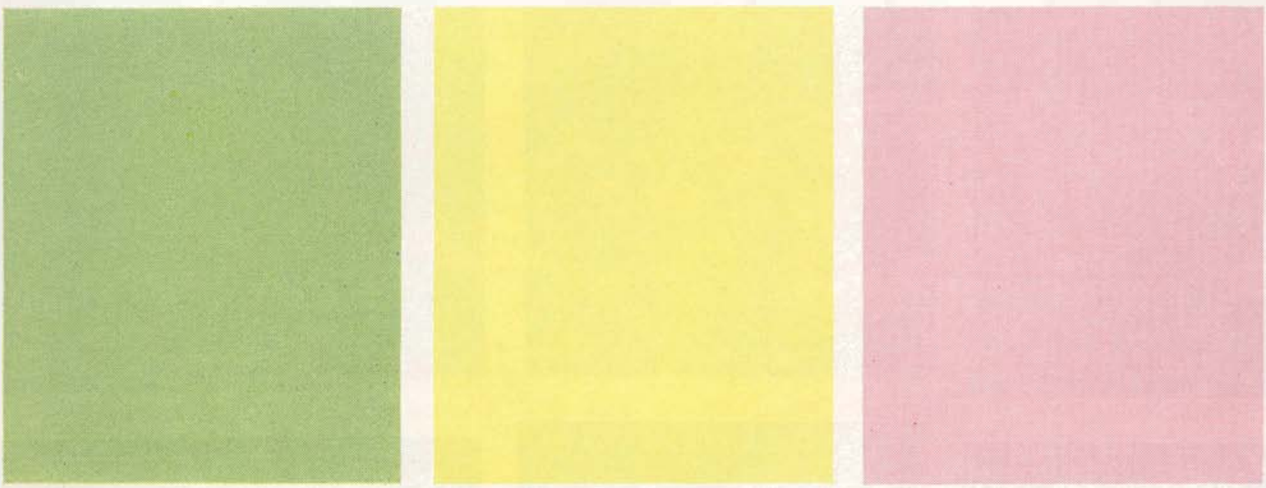
Since antiquity it has been known that paints can be mixed to obtain new colors. By the 18th century it was generally accepted that there are three primary colors and that all others can be created by mixing them in various ways. In 1802 the English physician Thomas Young located the three primary "colors" not in the physics of light or in the nature of paints but in the color-sensitive mechanisms of the eye. Since Young feared that his patients would lose confidence in him if his addiction to science were known, he wrote very little about his trichromatic theory. Those who assume that what a man has not explicitly stated he has not perceived suppose Young did not see very far.

This ungenerous attitude was not adopted by Hermann von Helmholtz or James Clerk Maxwell, who independently assumed that Young had seen the full implications of his theory. Both made a point of naming him the originator of the three-vector analysis of color vision, which they proceeded to work out. Young had postulated the existence of three resonators in the eye capable of responding maximally to red, yellow (green in a later statement) or blue light. He had concluded that the sensation of color depends on the relative amplitudes of the vibrations induced in the three resonators.

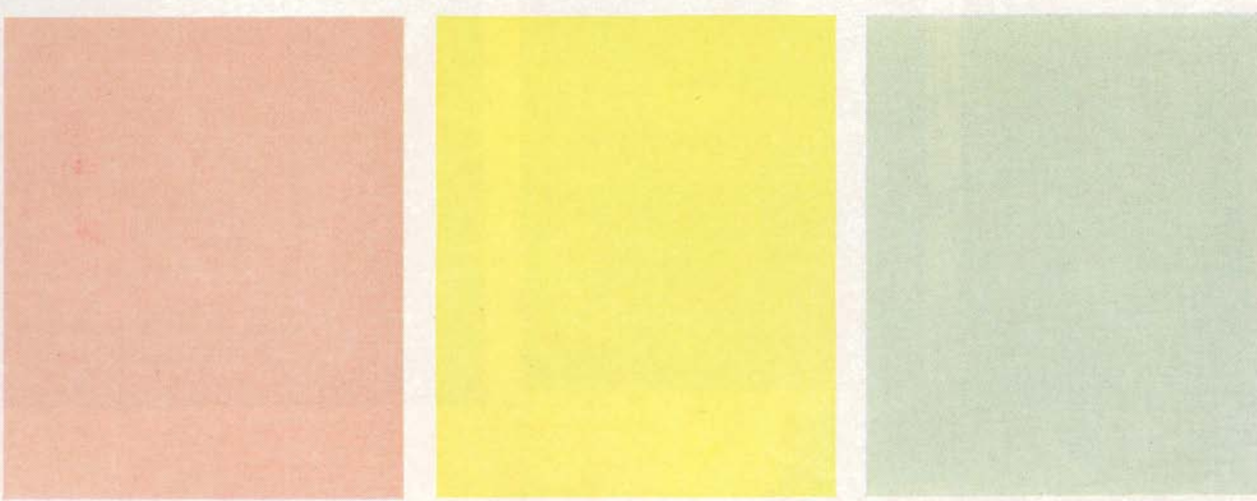
Light, we now know, acts on the photoreceptors of the retina, which are the rods and the cones. The rods are responsible for twilight vision, which is colorless; the cones are responsible for the colored vision of daylight. Toward the end of the 19th century Willy Kühne of Heidelberg discovered the photochemical basis of rod vision. He dissected out frog retinas in the faint yellow light from the flame of a wick steeped with salt and found that the pink retina thus obtained was bleached to a pale yellow on exposure to daylight. This pink pho-



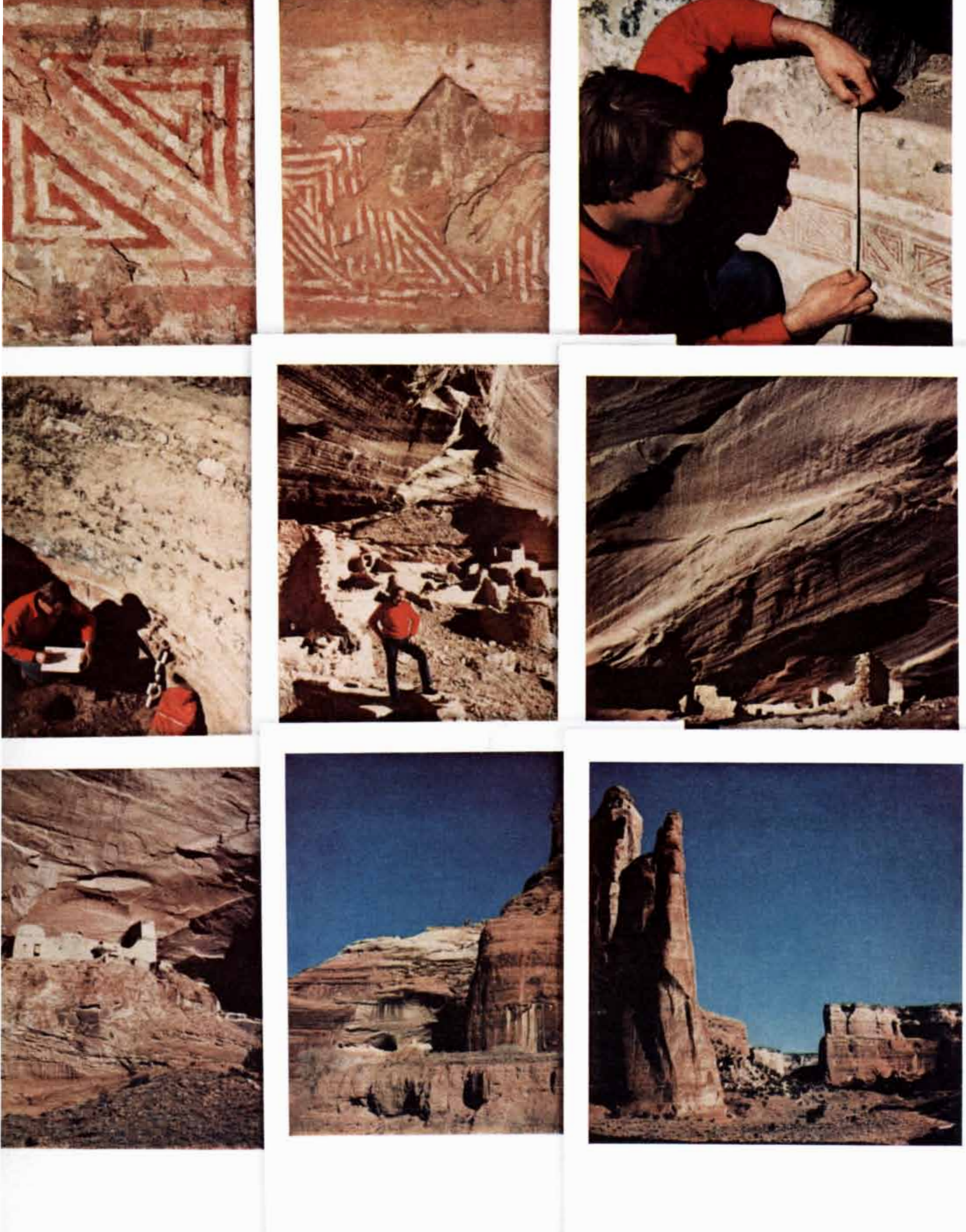
PROTANOPE, one type of dichromat, has defective color vision because he lacks the red-sensitive cone pigment erythrolabe and as a result cannot distinguish between green and red lights or any spectral color in between. These three color blocks simulate the appearance to the normal eye of lights the protanope judges identical except for small differences in brightness. When a protanope is asked to match red with a given green, he selects a red about 10 times hrighter than the red selected by deuteranopes (*below*).



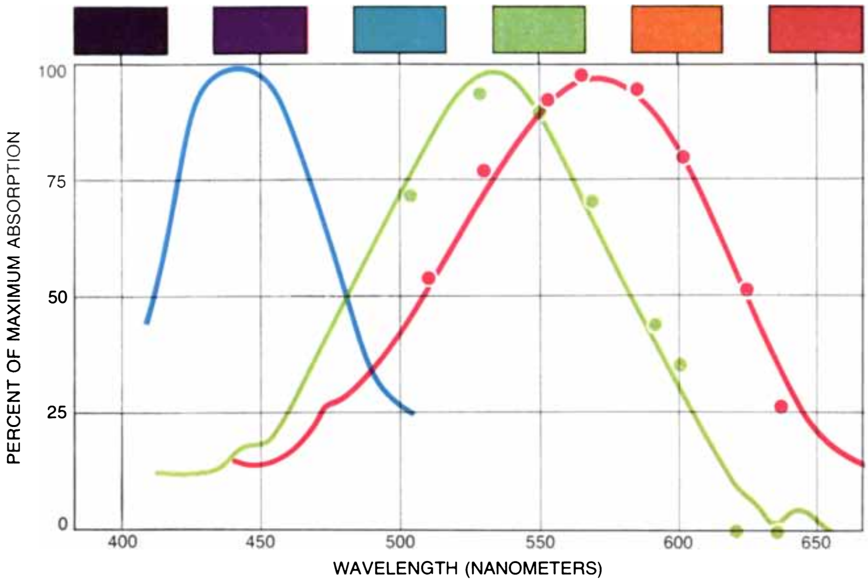
DEUTERANOPE is a dichromat who lacks the green-sensitive pigment chlorolabe. Like the protanope he cannot distinguish between red, green and colors in between. These three blocks show the appearance to normal eyes of lights deuteranope regards as identical.



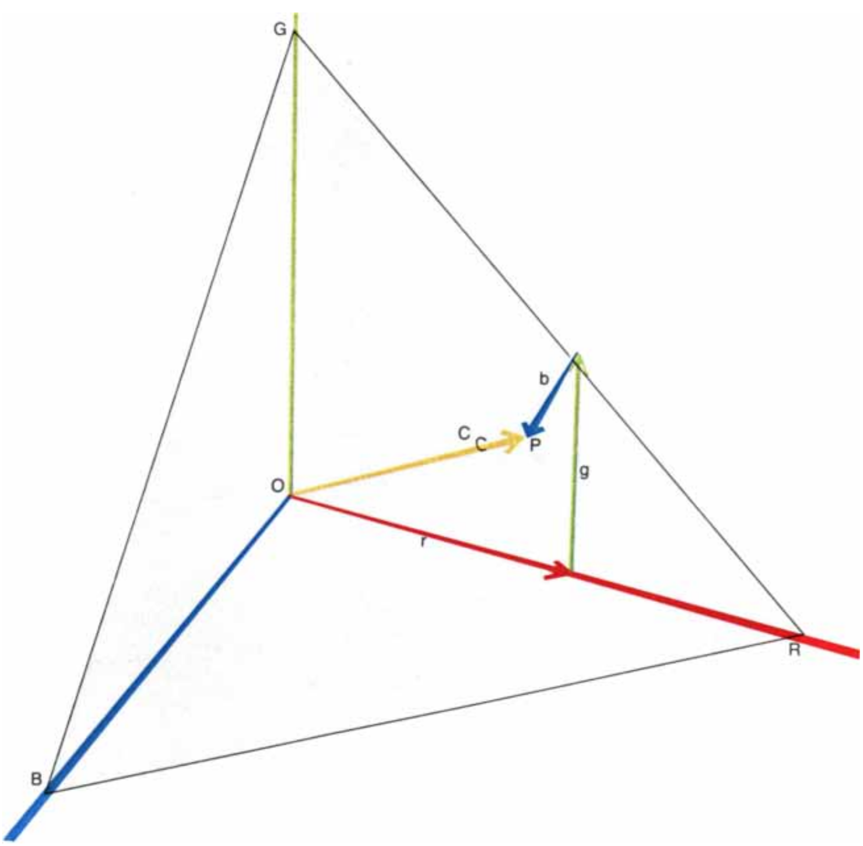
ANOMALOUS TRICHROMATS resemble normal people in that they need a mixture of red, green and blue to match all colors. The mixture, however, requires abnormal proportions. When someone in the protanomalous class of anomalous trichromats is asked to match a yellow light (*middle*), he selects a pinkish light (*left*). Someone in deuteranomalous class selects a greenish light (*right*).



Sequence of photographs showing a 12th or 13th Century ceremonial room mural in Anasazi (Pueblo) Indian cliff ruin in Canyon del Muerto, Arizona. Photographs ranging from 10.4 inches (upper left) to 5 miles (lower right) display the find in relation to its surroundings. All photographs are reproduced at their actual size.



THREE CONE PIGMENTS of normal color vision absorb lights of different wavelengths as plotted here. The curves are the average spectral absorbance from single cones in excised eyes of humans or monkeys scaled to the same maximums. The measurements were made by Edward F. MacNichol, Jr., and his colleagues at Johns Hopkins University. The colored dots represent the pigments in the green-sensitive and red-sensitive cones as they were measured in the living human eye by H. D. Baker and the author at the University of Cambridge. The coincidence of the two sets of measurements demonstrates that single cones contain single pigments. The color patches that appear above the curves approximate the colors of the spectrum at the wavelengths indicated at the bottom of illustration.



CONCEPT OF "COLOR TRIANGLE" is based on theory of three-pigment color vision. Any given spectral color *P* (*in this example a pale yellow-orange*) can be matched by a mixture of lights consisting of *r* units of red, *g* units of green and *b* units of blue. The vector sum of *r*, *g* and *b* is the three-dimensional vector *OP*, the length of which represents brightness. The location *C*, where arrow *OP* pierces the color triangle *RGB*, represents hue.

tosensitive pigment he named rhodopsin. Observing the retina under the microscope, he saw that the pigment was contained in the outer segments of the rods.

Since rhodopsin is bleached by light, it is natural to suppose the change is the basis of rod vision. If that is true, a critical relation must follow: if lights of different wavelengths are adjusted in intensity to look identically bright by the colorless rod vision, they should all bleach rhodopsin at the same rate. This was shown to be the case by German investigators in the 19th century and has been confirmed by more accurate methods since.

By analogy it seemed likely that cones must also contain a photosensitive pigment, different from rhodopsin because the spectral sensitivity of the eye in daylight is not the same as it is in twilight. And presumably in man there are three kinds of cone, each with its own pigment to account for the trichromacy of normal color vision. How are these three cone pigments related to Young's resonators? Young knew the approximate velocity of light from the measurements of Olaus Roemer in 1675, which were confirmed by James Bradley in 1728. Young also inferred the wavelength of light in his own studies of interference fringes, from which he concluded that the resonators vibrate at nearly 10^{15} times per second. No gross object can vibrate as fast as that. What is thrown into oscillation must be something within the "atoms" that John Dalton had just postulated to account for chemical change. We now know that the resonators are electrons, in particular the pi orbitals of molecules, and that their resonance to light at different frequencies is expressed by the pigment's absorbance spectrum.

Young's idea that color results from the independent excitation of three kinds of cone lends itself to a three-vector graphical representation. Three axes mutually at right angles represent the magnitude of excitation of Young's three resonators, or the pigments in the three kinds of cone [*see bottom illustration at left*]. The color represented by a particular point (*P*) is the vector sum of an excitation (*r*) of the red-sensitive cones, together with the excitation (*g*) of green-sensitive cones and the excitation (*b*) of blue-sensitive cones. If all three components are increased in equal proportions (that is, if the intensity of light *P* is increased without changing its spectral composition), the result is an increase in the brightness of the mixture without change of color. Therefore the length of the radius vector *OP* signifies brightness, whereas the direction of *OP* signifies color, or hue.

If we imagine a diagonal plane intercepting the three axes at points *R*, *G* and *B*, we obtain a triangle (the "color triangle"), which is pierced by vectors *OP* representing every possible color; each vector will pierce the triangle at a different point. Many different color triangles have been proposed, depending on the meaning chosen for the term "cone excitation." They are all linear transformations (projections) of one another. Maxwell, who made the first color triangle, chose the most straightforward physically: the vectors of any color *P* were the intensities of three primary lights, a red, a green and a blue whose mixture matched *P*. Intensity units were chosen so that mixing one of each produced white. Thus white was at the center of the triangle.

A more physiological set of vectors is the relative quantum catch (the relative number of photons actually caught) in each of the three types of cone, since this is the immediate precursor of nerve excitation. Such a cone-pigment triangle is shown at the right. The measurement of the quantum catch in each cone pigment, however, is too difficult and imprecise to commend itself to such a meticulous body as the International Commission on Illumination (CIE). The CIE color triangle (like the square root of -1 in alternating-current theory) gives neat and exact answers to calculations by introducing a nearly incomprehensible concept of what is really going on.

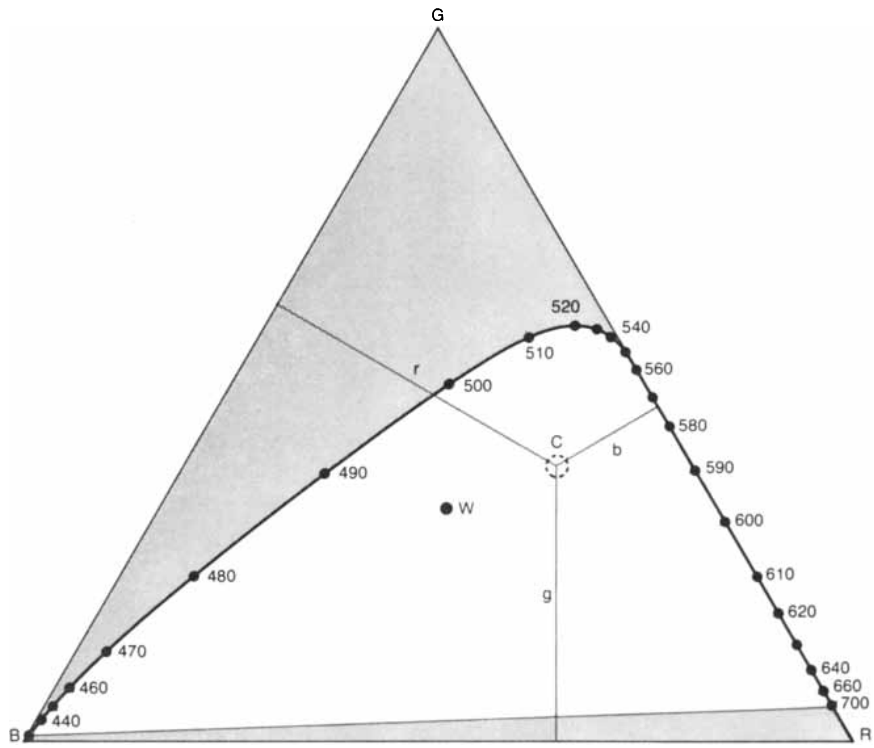
If we knew the spectral-absorbance curve for each cone pigment (that is, the spectral sensitivity of each cone), we should be able to calculate the relative quantum catch for any known light. We could then say, for instance, what quantities of red, green and blue Maxwell needed in his mixture to match any spectral light *P*. In other words, we could establish the condition where the quantum catch in each cone is the same from *P* as it is from the mixtures of *R* plus *G* plus *B*.

One might have hoped that the converse would hold, so that from knowing the quantities in the mixture that matches each spectral light we could deduce the spectral sensitivity of each cone. Unfortunately this is not the case. It would be if the red primary light excited only the red-sensitive cones, the green primary excited only the green-sensitive cones and the blue primary excited only the blue-sensitive cones. This, however, is not true. Until we know how much each primary light excites each of the three kinds of cones, we can say only that the color-mixture functions provide not the three spectral-sensitivity curves but weighted means of those curves, the weighting coefficients being unknown. The most insightful way out of this difficulty was a suggestion made by Arthur Koenig of Berlin at the end of the 19th century.

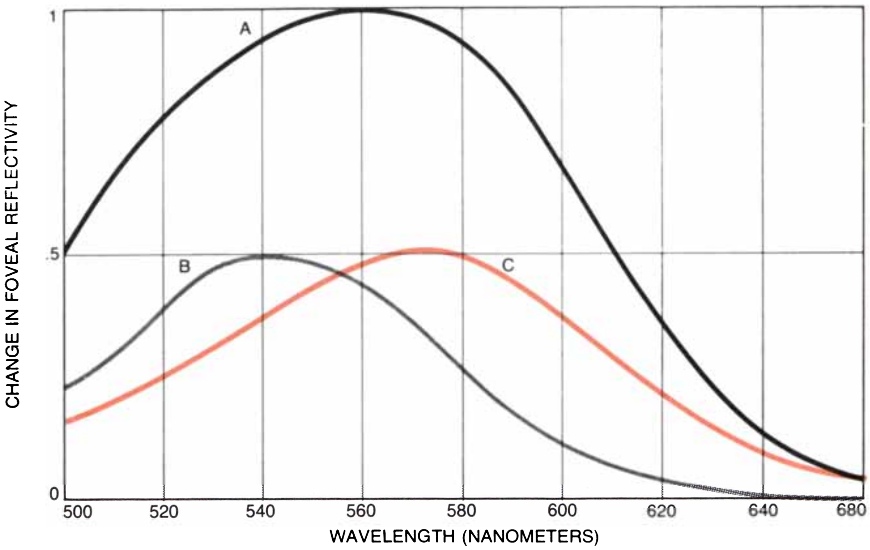
Koenig recognized that the principal difference between normal vision and the color-defective vision of the dichromat is that to match any wavelength in the red-green spectral range (between 540 and 700 nanometers) the normal person needs two knobs on the color-mixing box, one to regulate the red intensity in the mixture and one to regulate the green, whereas the dichromat, who cannot distinguish red from green, needs only one knob. It makes no difference whether the knob controls a red light or a green one (as perceived by normal people); the dichromat can match any mixture of red and green simply by turning the one knob to achieve matching brightness. Koenig saw that this is exactly what would follow if the dichromats had only one kind of cone instead of the normal two kinds active in the red-green range: two cones, two dimensions of color, two knobs; one cone, one dimension, one knob.

Two expectations follow and both are true. The first is that there should be two types of red-green confusers, those lacking the red-sensitive cones and those lacking the green. As we have seen, there are indeed two types of dichromats: protanopes, to whom red lights are dim, and deuteranopes, to whom the same lights seem bright. The second expectation is that both types should accept all color matches made by normal eyes. When a person with normal vision matches two fields of color, the red-sensitive pigment catches quanta as fast from the one field as it does from the other, and the green-sensitive pigment does the same. Thus a subject who has only one of these normal pigments will also catch light quanta equally from the two fields and say that they match. This too is found to be true.

In spite of its seductive simplicity, Koenig's theory is not the only one that



COLOR TRIANGLE, one of many possible, is the triangle *RGB* in the bottom illustration on the opposite page. Perpendiculars *r*, *g*, *b* from *C* to the opposite sides *R*, *G*, *B* are proportional to *r*, *g*, *b* in the preceding illustration. Their lengths correspond to the "quantum catch" (the actual number of photons caught) in *R*, *G* and *B* cones respectively, each expressed as the fraction of the catch from white light, point *W* in the center of the triangle. The curve shows the position of various monochromatic lights. Note that wavelengths greater than 540 nanometers are not absorbed by blue-sensitive cones. Lights between 510 and 480 nanometers, on the other hand, have appreciable perpendiculars onto all sides of the triangle. Thus these lights are absorbed by all cones, and white appears as a diluent in the resulting sensation. This effect may be reduced by previous adaptation to strong red light.



DIFFERENCE SPECTRA of green- and red-sensitive pigments in the normal eye can be determined by shining lights into the eye and measuring the fraction reflected from the foveal region in the center of the retina. Curve *A* is the change in reflectivity at various wavelengths when a normal dark-adapted fovea is bleached by a strong white light. Curve *C* is the change from dark adaptation when the bleaching light is a red of a strength sufficient to make the peak of *C* half the height of the peak of *A*. Curve *B* is the change when bleaching light is a bluish green of a strength sufficient to make the peak of *B* also half that of *A*.

could account for these facts. An alternative explanation proposed by Adolf Fick of Kassel has been widely entertained, particularly to explain deuteranopia. Fick suggested that deuteranopes have the normal red and green cones but their messages are mixed, so that the brain cannot tell whether red or green light provoked the resulting sensation. The distinction between the Koenig and Fick hypotheses can be paraphrased as follows. An artist with one pot of red paint and one of green who intends to paint a colored picture will be reduced to monochromacy if some prankster either removes one pot (Koenig) or thoroughly mixes the two paints together and leaves nothing but a mustard-colored mixture (Fick). For the Fick hypothesis to be true it would not be necessary for the messages from red and green cones to be mixed on the way to the brain; the equivalent result would obtain if the red and green pigments were mixed in a single class of cones.

Koenig's theory rests on two assumptions, and it will be true if they can be shown to be true: (1) In protanopes and deuteranopes one pigment is absent. (2) The remaining pigments are normal. We have obtained some good evidence to establish the truth of both of these assumptions, as I shall now relate. In order to learn if a subject has one cone pigment or more in the red-green range one would like to be able to measure the pigments objectively. My colleagues and I have been able to do this by reflection densitometry, which involves shining lights into the living human eye and measuring the amount reflected back [see "Visual Pigments in Man," by W. A. H. Rushton; SCIENTIFIC AMERICAN, November, 1962].

When a cat's eye is caught in the headlights of a car, it is seen to shine back green. Clearly the light has gone into the eye and undergone some selective absorption. By comparing the composition of ingoing and returning light we can measure the spectral absorption. Much of the absorption is due to factors other than visual pigments, but visual pigments are distinguished from everything else by the fact that they alone are bleached away by exposure to light. Thus if we analyze not simply the returning light but its change as a result of bleaching, we have a measure of the change in amount and composition of the visual pigments involved.

Since 1952 my colleagues and I have built and used six reflection densitometers. We have found, contrary to the general expectation, that the pigments in the cones are dense enough for satisfactory measurement. This has encouraged many people to engage in excellent densitometry studies both in intact eyes and in excised retinas. The surprising success of reflection densitometry in measuring cone pigments in the living human eye is mainly due to two anatomical features evolved by nature to perfect our cone vision and now used backward, as it were, to analyze it.

The first feature is that the fovea centralis, the tiny central retinal region of sharpest vision, is devoid of rods, the containers of the pigment rhodopsin, which elsewhere outnumber cones by 100 to one. By confining our densitometry to the fovea we can entirely exclude the overwhelming rhodopsin contribution to the pigments measured. The second helpful feature is that light falling on the retina enters the base of the cones and is focused onto the fine tip where the pigment is, in the same way that a burning glass focuses sunlight. This makes, of course, for an efficient use of the incident light. If one looks through a burning glass at the little area charred, however, one sees the tiny area greatly magnified. In similar fashion the speck of pigment at the tip of each cone is magnified to fill the base, so that the entire foveal floor is virtually filled with cone pigment.

We need not consider here the construction and use of our densitometers. Suffice it to say that the spectral reflectivity of the eye is measured by sending in lights of various wavelengths but of an intensity always too weak to bleach the pigments appreciably. The reflected light is measured by a photomultiplier cell. Bleaching is produced by lights of adjustable wavelength and strength, applied in pulses that alternate with the pulses of measuring light. In this way the change in pigment level can be followed at any wavelength throughout the course of any sequence of bleaching and regeneration.

Now let us see how the densitometer can be used to determine what it is that protanopes and deuteranopes lack. Is it a cone pigment or something in nerve organization? Densitometry measures pigments, not nerves. If nerves are the only thing wrong with dichromats, the densitometer will not be able to detect any abnormality. The fact is that in the normal fovea the densitometer detects two cone pigments (in the red-green range), but in the fovea of a dichromat it detects only one. Hence we can say with confidence that these defectives lack one kind of cone pigment, as Koenig maintained. (This was established some years ago in our work at the University of Cambridge.)

The way to detect two pigments in the normal eye is first to measure the foveal reflectivity with lights of different wavelengths in the unbleached eye. One then exposes the eye to a white light so strong that all pigments are bleached away.

The change of pigment density (not simply the pigment density itself) that is measured at various wavelengths after full bleaching yields a bell-shaped curve with a maximum at about 550 nanometers called the difference spectrum [*curve A in illustration on opposite page*]. The subject is then kept in total darkness for seven minutes to allow his cone pigments to regenerate fully. We now bleach the pigments with a red light of a strength such that the change in pigment density (the difference spectrum) corresponds to curve *C* in the illustration, whose peak is half as high as the first curve, *A*, and turns out to be situated at about 580 nanometers. After another pause for cone regeneration we repeat the bleaching, this time with a bluish green light of a strength sufficient to give the difference spectrum *B*, also half the height of *A*. The maximum of curve *B* occurs somewhat below 550 nanometers.

These results in the normal eye are what would be expected if there are two cone pigments, one absorbing more in the red (and thus being more sensitive to red light) and one absorbing more in the green. The red bleaching light will remove mainly the red-absorbing pigment and hence give a difference spectrum with a maximum toward the red. The bluish green light similarly will give a difference spectrum with a maximum toward the green.

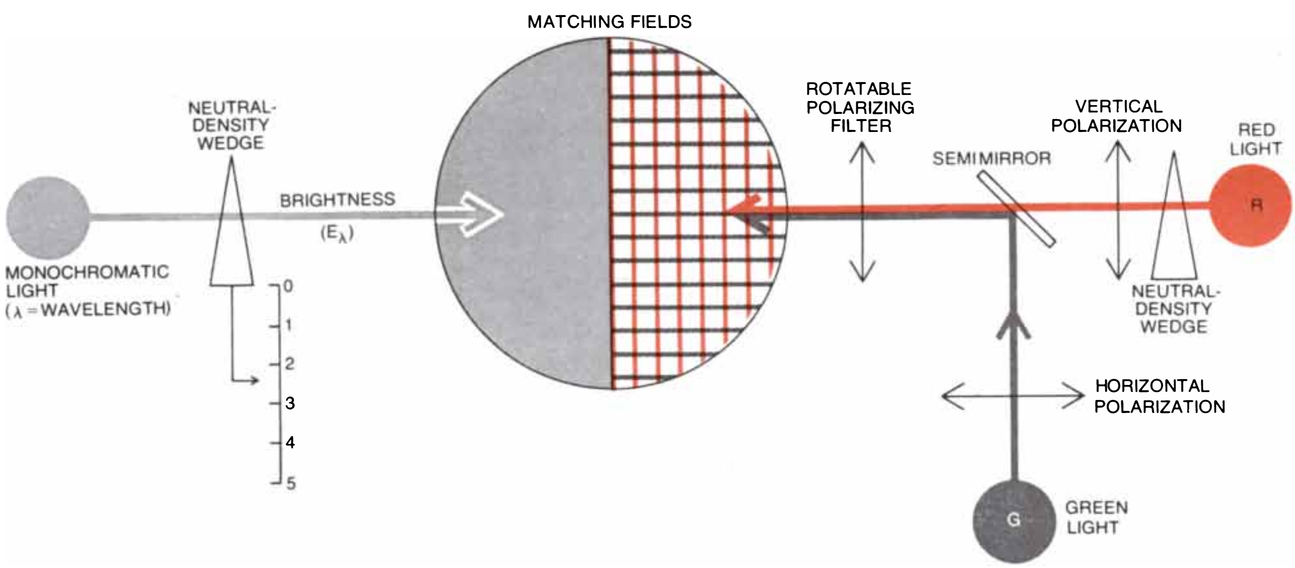
These two curves, *C* and *B*, which are expected in normal eyes and are found in them, should not be found in the eyes of dichromats, according to Koenig's theory. If these color-defective subjects have only one foveal pigment bleachable in the red-green range, then any light that bleaches it 50 percent must do the same thing and have the same result. That is precisely what is found. Half-bleaching by either red or bluish green light results in an identical difference spectrum. This must happen if dichromats have only pigment that is measurable in the red-green range; it could not happen if they had two or more pigments that were measurably distinct. We may thus conclude with some confidence that dichromats have only one dimension of color vision in the red-green range instead of two because they have only one cone pigment instead of the normal two.

The evidence is strong, therefore, that these dichromats possess only one cone pigment that is sensitive in the red-green range, since they consistently produce only one kind of difference spectrum regardless of whether the half-bleaching light is red or green. It is also clear that the pigment in the deuteranope is more red-sensitive than the pigment in the protanope. Hence Koenig's first assumption is justified.

It is much easier to compare various measurements in a single eye than to compare the eye of one person with that of another, because reflection and scattering usually differ between different eyes. Hence to test Koenig's second assumption—that the cone pigments dichromats do possess are normal pigments—we have not simply measured different spectra in various eyes but also used a more reliable approach.

I have mentioned how the rod pigment rhodopsin was proved to be the pigment of twilight vision: when lights of various wavelengths are adjusted in intensity to bleach rhodopsin at the same rate, all appear equally bright by twilight vision. At Florida State University, D. E. Mitchell and I applied the same test to see whether or not the cone pigments chlorolabe or erythrolabe in the dichromat are the pigments of daylight vision in the protanope or the deuteranope. If lights of different wavelengths are adjusted in strength to bleach the pigment at a fixed rate, all should appear equally bright to the dichromat. This experiment can be done with a high degree of accuracy, as I shall now undertake to explain.

It is easy to adjust the strength of a steady light of any wavelength so that it keeps the pigment 50 percent bleached in a steady state. The pigment bleached to this level is continuously regenerating at a fixed rate while it is continuously being bleached by the steady light. Thus when the light intensity is adjusted for equilibrium, bleaching equals regeneration. Now we can change the wavelength of the steady bleaching light and simultaneously change its intensity to the value that is found to keep the pigment still at the 50 percent bleached level. In this way the appropriate energy can be found for each wavelength



METHOD FOR MAKING RAYLEIGH MATCHES uses an instrument in which a yellow light (*represented by light gray*) is presented in the left half of a field and a mixture of green (*G*) and red (*R*) lights is presented in the right half (*green is represented by dark gray*). The *G* and *R* beams are polarized at right angles to each other and are mixed to produce a beam that passes through a polarizing filter. By rotating the filter the subject can produce any mixture of green and red needed in order to match the light in the left half of the field. He matches the brightness of the two fields by moving the neutral-density wedge in or out of the left-hand beam.

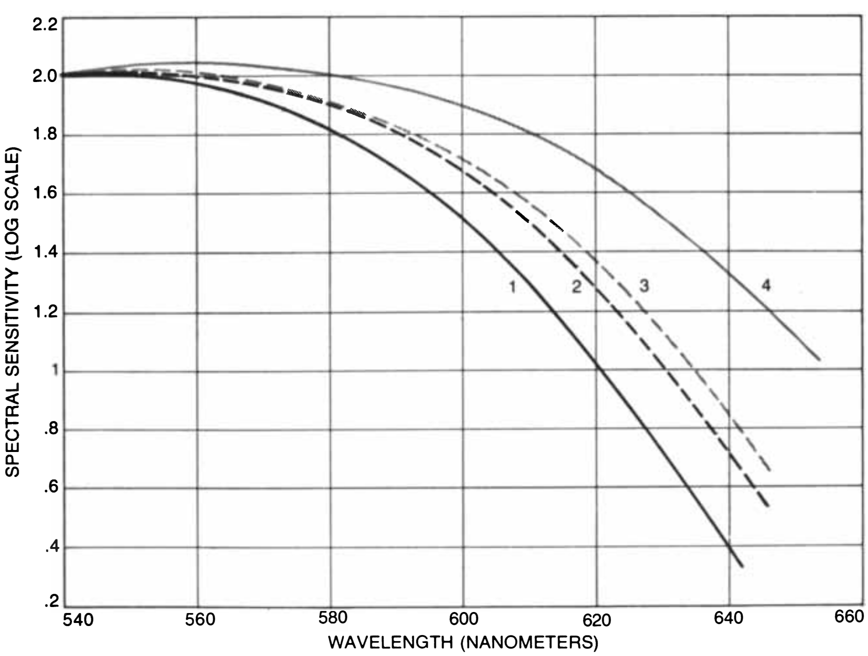
that bleaches at this constant rate. How about the appearance of brightness?

The dichromat is presented with a field divided into two halves; the left half is filled by yellow light of adjustable intensity; the right half is filled sequentially by bleaching lights of various wavelengths, from 540 nanometers in the green part of the spectrum to 650 nanometers in the red, all adjusted to maintain the pigment steadily at the 50 percent level. The dichromat cannot distinguish in color the yellow from the various bleaching lights; hence he can make a perfect match in each case by simply adjusting the intensity of the yellow. The result he gives is this: those lights that had been found to bleach equally fast (since they each kept the pigment in equilibrium at the 50 percent level) all need the same intensity of yellow for a perfect match. Thus lights that look equally bright to our dichromat bleach equally fast the pigment chlorolabe or erythrolabe that we measure in his eye.

This proves that the chlorolabe in the protanope and the erythrolabe in the deuteranope is the pigment with which that subject sees in the red-green range. The experiment also gives us the spectral sensitivity of these two pigments, since at wavelengths where the sensitivity is doubled the light energy for bleaching can be reduced by half, so that in general the sensitivity is the reciprocal of the energy needed to produce a constant bleaching rate [*see curves 1 and 4 in illustration below*].

The idea that protanopes are "red-blind" has long been widely held, since they find red lights dim. The fact that they lack erythrolabe accords well with this view. Many people have nonetheless found difficulty in accepting the symmetrical idea that deuteranopes lack chlorolabe, since deuteranopes experience no corresponding dimness in the green part of the spectrum. The asymmetry lies, however, in the shapes of the absorbance spectra of the two pigments, as is shown in the illustration on page 70. Nearly all the light at 650 nanometers or longer is absorbed by erythrolabe (peaking at about 575 nanometers). Hence the protanope, who lacks this pigment, will catch little light in the red and will be "red-blind." But the absence of chlorolabe (peaking at about 540 nanometers) still leaves erythrolabe absorbing well in the middle of the spectrum, so that the deuteranope, even though he lacks chlorolabe, is not "green-blind." He can still see this light as being bright, although he cannot distinguish its color from red.

Let me summarize the facts about dichromacy. Instead of having two kinds of cone pigment in the red-green spectral range protanopes and deuteranopes have only one. They accordingly have only one dimension of color in this range.



ANOMALOUS PIGMENTS in cones of anomalous trichromats have absorption curves between those of chlorolabe (*1*), the normal green-sensitive pigment, and erythrolabe (*4*), the normal red-sensitive one. Protanomalous trichromats have cones containing anomalous pigment protanolabe (*2*) in addition to cones containing chlorolabe. Deuteranomalous trichromats have cones containing deutanolabe (*3*) in addition to cones containing erythrolabe.

The spectral sensitivity of a protanope is the same as that of his cone pigment chlorolabe; the spectral sensitivity of a deuteranope is the same as that of his cone pigment erythrolabe. This spectral sensitivity is best measured by matching lights of equal brightness, and it is now well established.

The majority of color defectives are not dichromats, however. They need a mixture of three spectral lights to match all the colors of the spectrum, but they do not mix them in normal proportions. The first important step in classifying these anomalous trichromats was taken by Lord Rayleigh, who asked them to match a spectral yellow with a mixture of red and green. The anomalous subjects fell into two classes: the protanomalous, whose red-plus-green mixture is too red by normal standards, and the deuteranomalous, whose mixture is too green.

Some have considered that the protanomalous need more red in their mixture because, although they have normal cone pigments, the red signals are too weak. Perhaps there is too little erythrolabe in their cones; perhaps the erythrolabe is mixed with chlorolabe; perhaps the generator of red signals is too feeble; perhaps there is some abnormality in the nerve-processing network. We shall argue that none of these things can account for the condition; it must be due to an abnormal cone pigment.

The necessary condition for a red-plus-green mixture to match a given spectral yellow is for each of the cone pigments to receive the same quantum catch from the yellow field as from the red-plus-green field. Clearly if neither pigment can distinguish the two fields, nothing else can, and they will be indistinguishable. Thus a Rayleigh match that satisfies normal subjects must produce an equal quantum catch from both fields in both of the normal cone pigments. Therefore if anomalous trichromats have normal pigments (whether they are separated or partly mixed), these pigments too will have equal quantum catches from the two fields of a normal Rayleigh setting. If the pigments cannot distinguish the fields, the subject cannot distinguish them, and he will say that the fields match. That, however, is precisely what anomalous trichromats deny; they make quite a different setting, which they consistently say is the match for them. Therefore they must have an abnormal pigment, and we should like to know what it is. It has proved quite a tricky problem to identify this anomalous pigment, but after some 10 years of trials we seem to have got the answer.

Our first hope, naturally, was to be able to harness reflection densitometry to discover anomalous cone pigments in anomalous eyes. This failed entirely to give the answer. The protanomalous gave results indistinguishable from those in the protanope. The protanomalous subjects had the chlorolabe of the protanope but no other detectable pigment. They obviously did have a second pigment, since they could distinguish red from green with certainty, but it was either too scarce to show up or too similar to chlorolabe for separation. The nature of the separation that was tried was to bleach either with deep red light or with bluish green light and to measure the resulting difference spectrum over the range of wavelengths that had proved successful in the normal eye. If the two pigments are quite similar, however, they will be affected almost equally by the two bleaching lights. And the small change in the proportions of similar difference spectra will give rise to only a second-order small change in the difference observed.

Results with deuteranomalous subjects were similar; only an erythrolabe difference spectrum could be seen. One pigment was the same as the pigment in the deuteranope; the other pigment eluded detection, either through its scarcity with respect to erythrolabe or through its similarity to erythrolabe. Since densitometry will not give the answer, we need a different approach. Diane Spitzer Powell, Keith White and I have developed such an approach in our laboratory at Florida State University.

A protanope can exactly match any green presented to him with any red, simply by adjusting the red's intensity. Hence if we neatly exchange the matching red for the given green, he will not know that the substitution has been made. The explanation, of course, is that his one cone pigment, chlorolabe, catches quanta from the two lights equally and experiences no change in the rate of catch when the colors are exchanged. Suppose now that the exchange of these same lights is presented to the protanomalous who possesses chlorolabe plus an anomalous pigment. He can distinguish red from green and sees the change in color. His chlorolabe cones, however, see no change, so that the detection must have been accomplished by the anomalous cones. We have now an effect that depends on the change in quantum catch of the anomalous cones only. How can we use this effect to find those cones' spectral sensitivity?

As is well known, if a flash is presented superposed on a luminous background, the flash becomes increasingly difficult to see as the background is made brighter. In other words, the threshold for seeing the flash rises in proportion to the background luminance; this effect is known as the Weber-Fechner law. In our experiments we verified that the threshold for our subject to detect the red-green exchange is likewise raised in proportion to the background luminance. We selected two backgrounds, a red one and a green one, and adjusted their luminance so that they both raised the exchange threshold equally. Then, since only anomalous cones are involved in this threshold detection, backgrounds that are equal in threshold-raising power must be equally absorbed by the anomalous cones.

With this procedure one can critically adjust the intensities of any two lights in the red-to-green part of the spectrum (say R_1, with a wavelength of 640 nanometers, and G_1, with a wavelength of 540 nanometers) so that they are equally absorbed by the pigment of the anomalous cones. By establishing the critical intensities for various pairs of wavelengths one can infer the spectral sensitivity of the anomalous pigment. A more reliable method, however, is to make perfect Rayleigh matches for lights of various wavelengths. The threshold rise due to the change in background luminosity depends on nerve organization, but as we have seen color matches depend only on pigments. Thus the method of matches is more secure in interpretation besides being more accurate in execution. Any spectral light of some wavelength, lambda, between R_1 and G_1 can be matched in color by a suitable mixture of R_1 and G_1, the critically adjusted pair.

In making these Rayleigh matches we use an apparatus [*see illustration on page 71*] in which R_1, polarized vertically, and G_1, polarized horizontally, are brought together to form a single beam that is passed through a polarizer. By rotating the polarizer one can vary the ratio of R_1 to G_1 in the common beam. When rotation of the polarizer increases the green component, there is an equivalent decrease in the red component. Since the anomalous pigment catches as many quanta of R_1 as of G_1, the red and green lights exactly compensate, so that the catch is the same in this pigment whatever the R/G proportion.

The R/G mixture appears in the right half of the bipartite field the subject sees in our apparatus. The color in the left half of the field is simply the monochromatic light of wavelength lambda. The subject trims the brightness of lambda by adjusting the position of an extinction wedge so that the two halves appear identical. From calibration of the instrument we know the intensity, E_λ, of the lambda light at the setting that matches the R/G mixture. Then, since the light mixture on the right side is always absorbed by the anomalous pigment to a fixed extent, the monochromatic light E_λ on the left is also absorbed by this pigment to that same fixed extent and E_λ is the intensity required. Thus we know at each wavelength the light intensity resulting in a fixed quantum catch in the anomalous pigment, and this gives us at once that pigment's spectral-sensitivity curve.

The method for protanomalous subjects can be applied in the same way to deuteranomalous subjects and to normal ones. The exchange between red and green lights that look identical to deuteranopes is the exchange that cannot be detected by erythrolabe cones. It is therefore detected only by the green-sensitive cones of deuteranomalous or normal subjects. The intensity of a red background that raises the thresholds for these green cones as much as a fixed green background does is the red intensity that is absorbed by these green cones as much as the fixed green background is. We then make Rayleigh matches equating various monochromatic wavelengths in turn to a red-green mixture, with the primaries G and R set for equal absorption. The energy found to make the perfect match at each wavelength is the reciprocal of the spectral sensitivity of the green-sensitive pigment at that wavelength.

The green pigment in the normal eye, when it is measured in this way, was found to coincide with the chlorolabe, the only pigment in the protanope (in the red-green range), as determined by the fixed bleaching-rate technique. In the same way the red-sensitive normal pigment was found to coincide with erythrolabe in the deuteranope. This justifies Koenig's second assumption; hence we can accept with some confidence his brilliant conjecture that the normal green-sensitive and red-sensitive cone pigments are respectively the chlorolabe of the protanope and the erythrolabe of the deuteranope.

In this account it was assumed for purposes of clarity that each kind of cone contains a single visual pigment. It was pointed out that if the normal pigments were mixed in the cones, one could not account for the abnormal vision of anomalous trichromats. If the eye intends to distinguish colors, it would certainly seem a bad start for it to mix together in one cone the discriminating pigments. The eye, however, often looks deeper than physiologists do and exhibits unexpected tricks of organization. The fact remains that, as Edward F. MacNichol, Jr., and his colleagues at Johns Hopkins University have shown, the pigments are not mixed; each cone contains its own single pigment [see "Three-Pigment Color Vision," by Edward F. MacNichol, Jr.; Scientific American, December, 1964].

The work of MacNichol and his colleagues was a brilliant technical achievement. A microscopic light beam was passed entirely through the pigment of the outer segment of single cones obtained from human or monkey retinas. The alignment could not be made in visible light or the pigment would have been bleached away before the experiment started. Moreover, the measuring beam had to be so weak that the separation of the signal from the noise was a formidable problem. Ultimately the Johns Hopkins workers obtained three spectral curves, one for each pigment: a blue-sensitive pigment with a peak at about 445 nanometers, a green-sensitive pigment with a peak at about 535 nanometers and a red-sensitive pigment with a peak at about 575 nanometers. The three curves are shown in the top illustration on page 68. The colored dots superposed on the green-sensitive and red-sensitive curves represent, on the same scale, our photosensitivity measurements of chlorolabe and erythrolabe in the normal human eye. In our experiments the pigments were single, and in the Johns Hopkins experiments the cones were single; the coincidence of results means that single cones contain single pigments.

The illustration on page 72 shows the log spectral-sensitivity curves for both the normal and the anomalous pigments in the red-green range. "Protanolabe" [*curve* 2] is the anomalous cone pigment that protanomalous subjects possess in addition to the normal pigment chlorolabe. "Deutanolabe" [*curve* 3] is the anomalous cone pigment that deuteranomalous subjects possess in addition to the normal pigment erythrolabe. Color discrimination depends on the differential quantum catch in the two cone pigments. If these pigments are well separated, as they are in normal people, the differential catch varies considerably with change of wavelength. Hence the color appearance changes greatly, particularly in the yellow-green region. The pigments of the anomalous trichromats, however, have curves much closer together, and as a result the difference in their ordinates changes little with change in wavelength. Thus it is hard for these subjects to discriminate colors even when the fields are juxtaposed; when they have to rely on the memory of these difficult perceptions (as in naming colors), the task is harder still.

It is likely that we have oversimplified the classification of anomalous vision, since some protanomalous subjects do not accept the color matches made by others. Hence the anomalous pigment, protanolabe, may vary somewhat from subject to subject. If it does, we now have a method for measuring it and for studying how it varies.

Thomas Piantanida and Harry G. Sperling of the University of Texas Graduate School of Biomedical Sciences, working simultaneously and independently from us, have obtained results identical with ours. Their method, which was quite different, involved not color matches but thresholds with backgrounds of different colors and intensities. The fact that we both reached the same conclusions about the nature of the abnormal pigments by different paths strengthens confidence in those conclusions.

The experiments described here seem to establish the following main conclusions. (1) In all the defectives considered there is a loss or an anomaly in a cone pigment. The condition cannot be caused by the abnormal processing of responses from normal pigments. (2) Koenig's two basic hypotheses have been substantiated: protanopes and deuteranopes have lost one cone pigment in the red-green range, and the remaining cone pigment is normal. (3) The pigments chlorolabe and erythrolabe that we can measure by densitometry in dichromats are the pigments with which they see, since lights of different wavelengths that bleach equally fast look equally bright. Therefore these are also the pigments responsible for cone vision in the normal eye. (4) The abnormal cone pigments in anomalous trichromats have at last been measured. In each type of anomalous trichromat the sensitivity curve of the anomalous pigment lies very close to the sensitivity curve of the normal pigment in that eye, thus explaining why color discrimination is difficult for these subjects.

Seafloor oil wells.

To boost America's energy production, Exxon is taking a giant step down.

Today, under the Gulf of Mexico, Exxon is testing the prototype of the Submerged Production System —or SPS, for short. It's a system of seafloor wells that may someday allow us to reach deeper into the ocean than ever before for oil and natural gas.

Exxon saw the need for such a system in the mid-1960's—years before the term "energy crisis" entered our vocabulary. Now, after $30 million of design, engineering and construction, the first SPS unit is undergoing operational testing.

Why it's needed

Today, nearly all offshore oil and gas come to us via wells drilled from huge steel towers known as production platforms.

While these platforms are practical for production in relatively shallow water, the costs and technological problems involved make them less attractive in deep water. This is where the SPS may take over.

The operation

Once the SPS has been lowered to the seafloor and anchored, a drilling ship will complete a number of wells through openings on the unit and then leave.

Upon completion of the wells, the SPS is designed to pipe oil and gas to shore—or to a tanker overhead.

All SPS functions will be monitored and remotely controlled by an operator aboard a ship or in a control room on shore. At the push of a button, the operator will start up or shut down parts of the SPS unit.

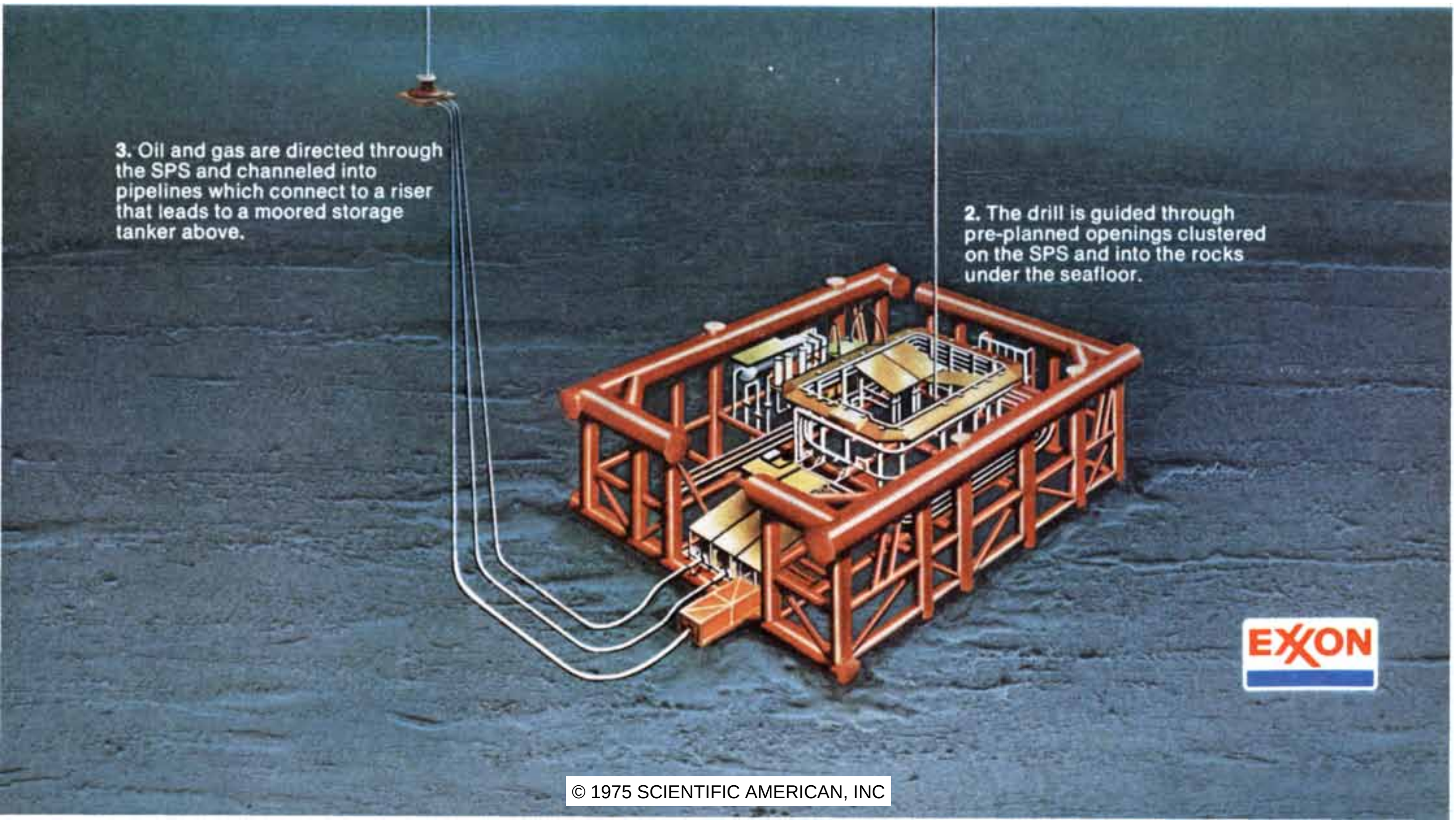
If repairs or servicing are needed, a maintenance device—the Manipulator—will be lowered to the SPS unit. The Manipulator removes the old part, inserts a replacement, and tests its work, all by remote control.

Environmentally safe

Besides having the potential to unlock billions of barrels of undersea oil—and trillions of cubic feet of natural gas—the SPS also offers many promising environmental innovations.

SPS units are designed with special equipment to prevent leaks—and special equipment to trap escaping fluids if a leak does occur.

Deeper beneath the sea than ever before—this is one of the places Exxon is going to bring you gas and oil.



The Role of Wax in Oceanic Food Chains

Copepods, ubiquitous marine crustaceans, store energy in the form of wax. It now seems that half of the organic matter synthesized by the sea's primary producers is converted for a time into waxes

by Andrew A. Benson and Richard F. Lee

Waxes are known to have important functions in a wide variety of plants and animals, but until recently they were not regarded as playing a major role in the energy economy of the animal world. It was well known that waxes on the surface of leaves and fruits protect them against damage and drying. Honeybees secrete a wax for the building of the honeycomb. The human skin produces sebum, a mixture of waxes, oils and fragments of dead cells that serves to lubricate the skin and make it supple. The head cavities of the sperm whale are filled with a mixture of liquid waxes that is believed to mediate the conduction and focusing of the high-frequency sound the animal uses for communication and echo location.

In the 1960's marine biologists began to find waxes in a wide range of marine animals. The most populous group of fishes, the bristlemouths or cyclothones, are rich in waxes. Appreciable amounts of wax are also found in cod, mullets, croakers and sharks. Mollusks, shrimps, decapods, squids, deep-sea worms, sea anemones, corals and even marine birds accumulate wax. All of this indicated that wax was stored by these animals in order to fill some need, but what it was remained a mystery. Then in 1967 Judd C. Nevenzel of the University of California at Los Angeles isolated a wax from the black deep-sea copepod *Gaussia princeps*. Copepods are small marine crustaceans that are enormously abundant; indeed, in most oceanic areas they are the largest single component of the zooplankton (animal plankton). Further investigation quickly established that many species of copepods synthesize and store waxes for use as a reserve metabolic fuel when food is not available.

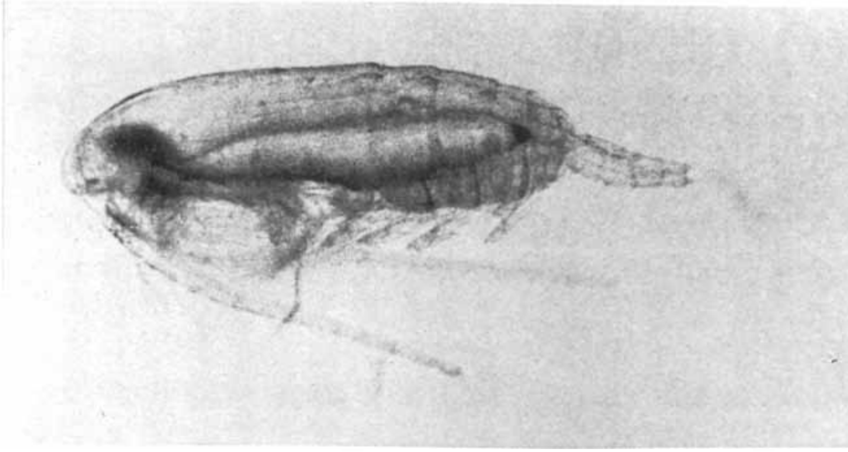
Copepods graze on the phytoplankton, the tiny plants that are the primary producers in the food chains of the ocean. These plants, which are algae of various types, build carbohydrates, fats and proteins by photosynthesis, and the copepods convert some of the fat into wax. Many higher forms of marine life in turn feed on the copepods [*see top illustration on page* 83]. The discovery that copepods are the main producers of the waxes in the marine food chain led to the realization that wax is a major medium of energy storage in animals of the sea. It is estimated that on a worldwide basis at least half of all the organic substance synthesized by the phytoplankton is converted for a time into wax.

The order Copepoda includes more than 2,000 species. Most copepods are herbivorous and feed only on the phytoplankton of the surface waters; some are omnivorous and feed on both phytoplankton and zooplankton; some, particularly the deep-sea species, must be carnivorous, feeding only on the zooplankton. Many species migrate vertically as much as 500 meters every day, descending to the depths in daylight and ascending to the surface at night, where they can feed on the phytoplankton that have flourished during their time in the sun.

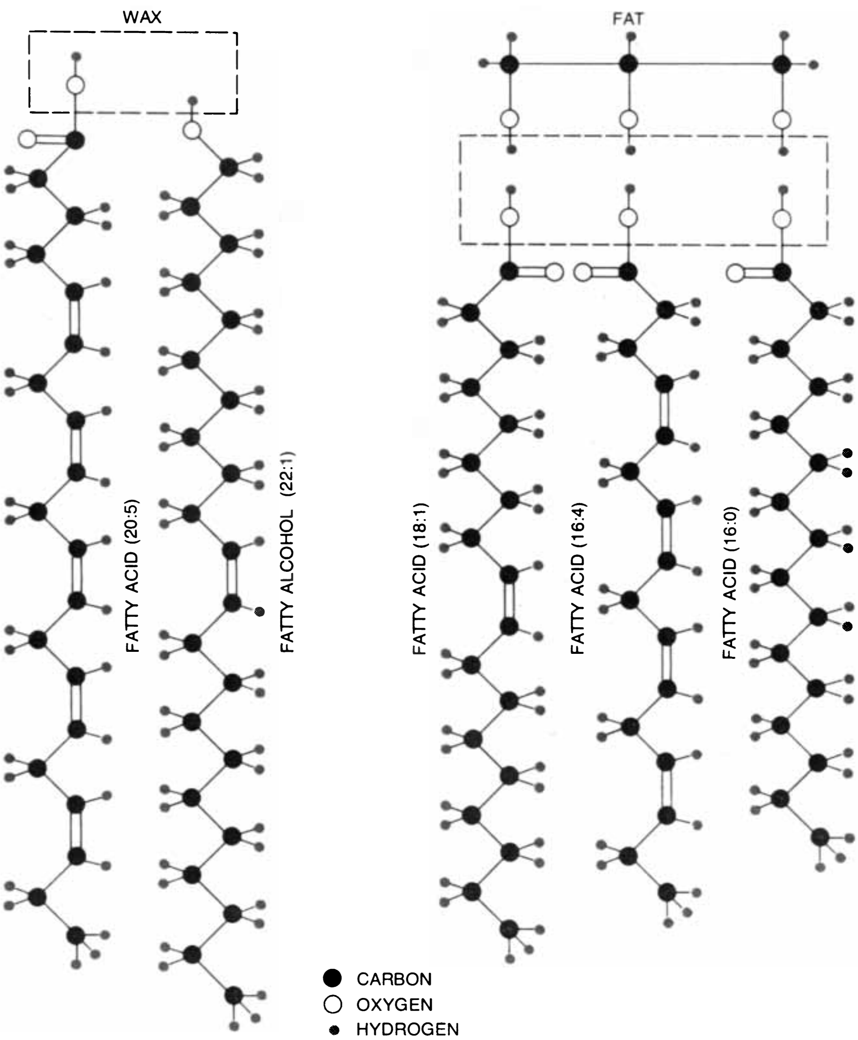
Copepods store both fat and wax as metabolic fuel. The fats are generally distributed throughout the body and provide the animal with energy for short-term purposes. The waxes are found in more specific locations. The calanoid copepods of temperate and polar waters, for example *Calanus plumchrus*, have a large sac in which wax is stored [*see top illustration on next page*].

Both fats and waxes are lipids; they also belong to the class of compounds known as esters. Fats are called triglycerides because fat molecules consist of three long chains of fatty acid linked to a molecule of glycerol. Wax molecules have a simpler structure, consisting of a long-chain fatty acid linked to a long-chain fatty alcohol [*see bottom illustration on next page*]. A shorthand notation is employed to indicate the length of the chain and the number of double bonds in it. For example, "oleic acid (16 : 1)" refers to a chain that is 16 carbon atoms long and has one double bond between two of the carbons. When the carbon chain in a fatty acid or a fatty alcohol has one double bond or more, it is called unsaturated. The great-

FEMALE COPEPOD carrying a cluster of bright blue eggs is enlarged some 20 diameters in this photograph by Fritz Goro. Copepods are the most abundant zooplankton (animal plankton). This copepod is *Euchaeta japonica*, a species common in the North Pacific. It is omnivorous, feeding both on other zooplankton and on phytoplankton (plant plankton). The large transparent cavity visible within the animal contains a store of both wax and fat. The wax accounts for as much as 50 percent of the dry weight of the adult female. When the female produces eggs, much of the wax and fat is transferred to the eggs to provide food for the larvae. The eggs are carried by the female to protect them from predators. Red color around copepod's mouthparts and blue color of eggs are both due to same pigment: astaxanthin, which gives rise to different colors when associated with different proteins.



LARGE OIL SAC filled with pure liquid wax is clearly visible in this photograph of a Stage V copepodite of the species *Calanus plumchrus*, the predominant copepod in the Strait of Georgia in British Columbia. On a dry-weight basis 50 percent of the copepodite is wax.



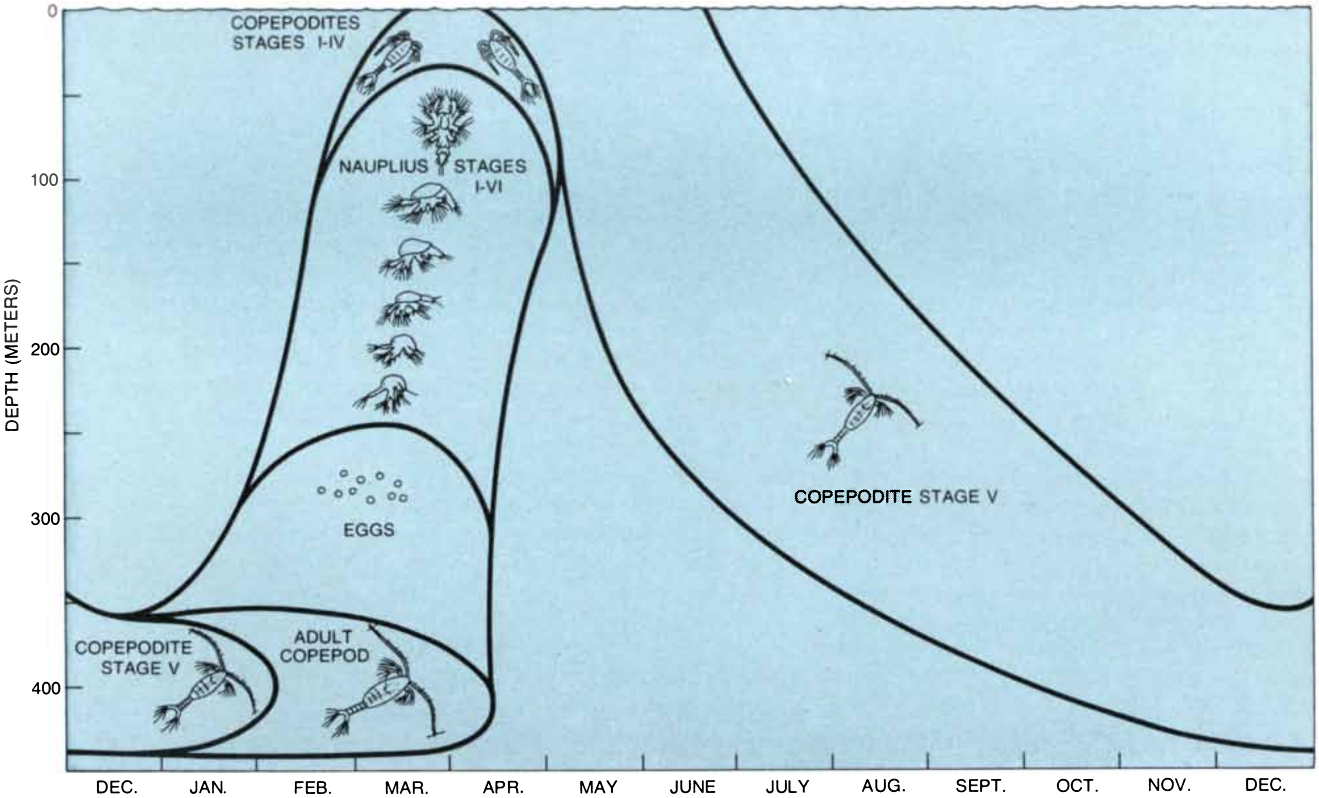
WAXES AND FATS are both esters, substances produced by the reaction of alcohols and acids. Wax is formed from a long-chain alcohol and a single fatty acid. The numbers in parentheses are a shorthand notation for identifying the carbon chain; for example, the numbers 20:5 refer to a chain that has 20 carbon atoms and five double bonds between the carbons. When a wax is formed, one molecule of water is released (*rectangle at left*). Fats are called triglycerides because they consist of three fatty acids linked to the three oxygen atoms of glycerol. Three molecules of water are liberated in the formation of triglyceride.

er the unsaturation in a fat or a wax, that is, the greater the number of double bonds, the more likely it is that the substance will be liquid at low temperatures. The fats synthesized by phytoplankton in cold waters are highly unsaturated, or polyunsaturated, and the waxes synthesized by the copepods that feed in those waters are similarly unsaturated. These waxes remain liquid even when the temperature of the water is close to freezing. In general, organisms adapted to low temperatures store more polyunsaturated lipids than organisms that live in warmer regions. Unsaturation therefore is a biological adaptation to life at low temperatures: it allows a plant or an animal to function well in a cold environment.

The fats and waxes in copepods are much alike in their physical properties. Both have nearly the same density, a similar caloric value per unit volume and similar compressibility. They do, however, differ in their coefficient of thermal expansion. Compared with a droplet of fat, a droplet of wax in a copepod expands more and becomes more buoyant as it is warmed. This change in buoyancy may assist copepods in their daily vertical migration from the cooler water of the depths to the warmer water of the surface.

Many organisms require for their metabolism fatty acids that they cannot synthesize for themselves. Such fatty acids must be obtained from food, and they are termed essential fatty acids. For man the essential fatty acids are linoleic acid (18 : 2) and linolenic acid (18 : 3). For copepods these fatty acids are also essential, particularly for the synthesis of docosahexaenoic acid (22 : 6), a crucial component of the phospholipids in the cell membranes of copepods, fishes and apparently all other marine animals. Herbivorous copepods synthesize docosahexaenoic acid by the addition of two-carbon units to unsaturated fatty acids obtained from phytoplankton. Such highly unsaturated fatty acids oxidize rapidly when they are exposed to air, giving rise to the characteristic odor of rancid fish. The polyunsaturated fatty acids are carefully conserved by the animal. For example, when the copepod *Gaussia princeps* is starved, its fat store is depleted in about two weeks, but there is very little change in its content of wax. Although the fat disappears, its valuable polyunsaturated chains are transferred to phospholipids in the cell membranes, where they cannot be depleted.

In most land animals fat is the only type of fuel that is stored for long-term



LIFE CYCLE of *Calanus plumchrus* copepods, based on the studies of John Fulton and Robin LeBrasseur of the Fisheries Research Board of Canada, is depicted. In late winter the adult female begins to lay eggs. The eggs rise slowly and hatch in about two and a half days. The newly hatched larvae, called nauplii, live off the wax stored in their eggs. They go through six molting stages and then become copepodites. The copepodites feed on the phytoplankton near the surface and go through five molting stages. By the time they have reached the fifth stage copepodites have large stores of wax. They migrate down to the deeper waters and during the winter become adult copepods fully equipped for mating. The adult has no feeding mouthparts and must live off its stored wax.

use. Although copepods also store fat, in many species the major reserve fuel is wax; indeed, as much as 70 percent of the animal's dry weight can be wax. This alternative food-storage system provides the copepod with a food reserve that can be controlled separately from its day-to-day metabolism. The control appears to rely on the relative activity of two enzymes: triglyceride lipase and wax lipase. The triglyceride lipase, which catalyzes the metabolism of fat, is normally active in the animal at all times. For example, in freshly caught copepods triglyceride-lipase activity can readily be detected, but the wax-lipase activity is virtually zero. It appears that while other food is available the enzyme that catalyzes the metabolism of wax is inactive, and that the enzyme is activated only under the stress of starvation. This provides the control that prevents the early depletion of the wax reserve. Thus copepods that accumulate wax do so because they are unable to use the wax as fuel until all their fat reserve is gone.

The utilization of waxes is strongly dependent on the nature of the copepod's environment. Working with Jed Hirota and Arthur M. Barnett, we captured 85 species of copepods from various depths in the subtropical and temperate regions of the Pacific and compared them with specimens collected under an ice island in the Arctic Ocean. In general larger amounts of wax are found in copepods as the temperature of the water in which the animals live decreases or as the depth of the habitat increases. Copepods in arctic or antarctic waters, where phytoplankton bloom only during two or three months of summer sunlight, store the greatest amount of wax. In warm tropical waters, where the supply of phytoplankton is sparse but relatively constant, copepods in the surface water accumulate very little wax. All copepods (and most other animals) from depths below 1,000 meters contain a considerable amount of wax.

In the subtropical waters off San Diego we found that for most copepod species the total lipid accounted for only 18 percent of the dry weight, and that there was usually only a small amount of wax in the lipid. In copepods taken at depths below 600 meters the total lipid accounted for 40 percent of the dry weight, and well over half of the lipid was wax. We also found that there was a transition zone where about half of the copepods we caught had stores of wax and half had stores of fat. The explanation is that the daily vertical migration of surface species of copepods can result in mixtures of animals.

John Fulton and Robin LeBrasseur of the Fisheries Research Board of Canada have studied the life patterns of copepods in the Strait of Georgia on the Canadian West Coast. One important species studied was *Calanus plumchrus,* which is a major source of food for young salmon arriving from inland streams and rivers. During the winter the adult copepods remain at a depth of about 400 meters. In late winter and early spring the female begins to lay eggs, which contain droplets of oil and are therefore buoyant. They rise slowly from the depths where they were released, hatching in about two and a half days and swimming to the surface. The newly hatched larva, called the nauplius, is barely able to swim and is not yet equipped to capture phytoplankton. Fortunately it has a supply of wax that provides much of the energy it needs. After the nauplius has grown and molt-ed, it must find phytoplankton of a size it can eat if it is to survive through its six molting stages. It then becomes a copepodite, which also molts as it grows. At about this time nutrient-rich water from the spring flood of the Fraser River gives rise to a great bloom of phytoplankton in the Strait of Georgia. The copepodites feed on the phytoplankton as they go through their molting stages. By the end of the bloom the copepodites have accumulated large stores of pure wax.

The copepodites migrate down to the deeper waters during the late summer months. At that time they feed very little, if at all. During the midwinter months they become adult copepods. The adults of *Calanus plumchrus* cannot feed and must rely on the wax formed the preceding spring for all their energy requirements during the last seven months of their life cycle. Most of the wax reserve of the female goes into the production of the eggs she discharges. The oil stored in the eggs surprisingly consists mostly of fat; it contains very little wax. Before the female breeds nearly 50 percent of her dry weight is in the form of lipid, and 90 percent of the lipid is wax. After the female has discharged between 300 and 800 eggs her lipid content has dropped to 4 percent, with only traces of wax remaining.

Arctic copepods, for example *Calanus hyperboreus,* go through an almost identical life cycle. The wax is stored during the months of July, August and September, when the phytoplankton bloom in this area. For the other nine months the copepods must survive on their reserve of lipid. The females shed eggs in January and February, and presumably they do not live much longer.

The surface-feeding copepod *Calanus helgolandicus* produces lipid amounting to 20 percent of its dry weight before its final molt into the adult stage. Although the lipid of the female of this species is 50 percent wax, the eggs she discharges are 60 percent fat and slightly denser than water. In two days nauplii having neither a mouth nor an anus emerge and valiantly struggle toward the surface, molting twice in two days before they are at last capable of feeding on small phytoplankton.

The eggs of the omnivorous deep-water copepod *Euchaeta japonica* are 58 percent wax. The female carries the eggs, which have a characteristic bright blue color [*see illustration on page 76*], to protect them from predation by other zooplankton. The early naupliar stages do not feed and must rely on the stored wax in the egg for their energy requirements as they grow and molt. In some instances there is enough wax in the egg to carry *E. japonica* through the entire six naupliar stages to the first copepodite stage.

Copepods are a colorful group of creatures: they can be orange, red, black, blue or white. White copepods, however, always have streaks or spots of orange pigment. Whatever the color, the pigment is always the same. It is astaxanthin, which copepods produce from the yellow xanthophylls of phytoplankton. This carotenoid pigment is also seen in shrimps, lobsters and salmon. When the molecule of the pigment is associated with certain protein molecules, it assumes different configurations and gives rise to different colors. Astaxanthin can be extracted from the blue eggs of copepods or from the animal's body with an organic solvent such as chloroform, whereupon the color of the solution is always a reddish orange.

Much has been learned about the metabolism and the nutritional requirements of copepods by raising them in the laboratory. G.-A. Paffenhöfer and M. M. Mullin of the Scripps Institution of Oceanography succeeded in raising *Calanus helgolandicus* from eggs to adulthood. It was the culture of the newly hatched nauplii that required the greatest care. Paffenhöfer kept them alive (with the help of an alarm clock) by feeding them every few hours with algae. He then succeeded in bringing the nauplii through their six molting stages and the copepodites through their five molts to the adult stage.

The composition of copepod wax is altered not only by starvation but also by changes in the concentration of the food available and by the species of phytoplankton that are ingested. When laboratory-reared copepods were fed the alga *Skeletonema costatum,* which is rich in fats but contains negligible amounts of wax, the fatty acids of the wax synthesized by the copepods resembled the fatty acids found in the fat of the alga. The copepod fatty alcohols, however, did not resemble the algal fatty acids.

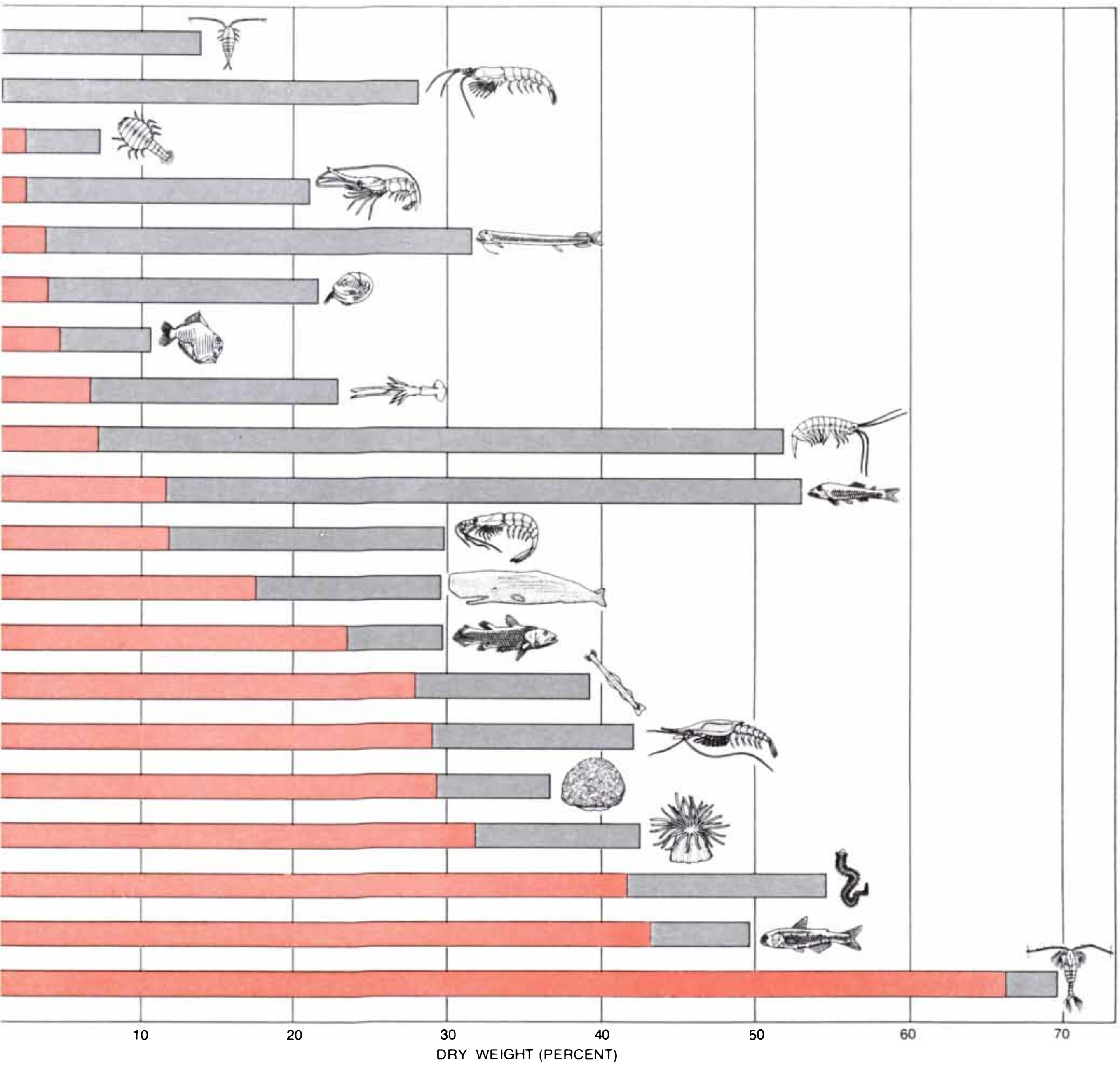
Since algae do not contain any long-chain alcohols, these compounds must be synthesized by the copepod and then combined with fatty acids to make wax. The wax of copepods that live in surface waters is rich in 20- and 22-carbon alcohols, whereas the wax in copepods that live in the depths is characterized by 16-carbon alcohols, which remain liquid at low temperatures.

The synthesis of wax, mediated by enzymes isolated from copepods, has been accomplished by John R. Sargent of the Institute of Marine Biochemistry in Aberdeen in Scotland. He found that fatty acids that had been radioactively labeled were converted into fatty alcohols and waxes by cell extracts and cell suspensions of copepods supplemented with coenzyme A, adenosine triphosphate (ATP) and reduced pyridine nucleotides. In our laboratory at the Scripps Institution, R. Barry Holtz and E. D. Marquez separated the organelles from various copepod cells and found that the biosynthesis of wax is accomplished on plasma membranes, possibly membranes associated with the oil sac.

AMOUNT OF WAX in marine animals varies considerably. All animals that live at

Man-made oil slicks at sea have become commonplace in modern times, but it is not generally realized that there are substantial biological oil slicks as well. On April 7, 1971, crewmen aboard the U.S. Coast Guard cutter *Minnetonka* saw a red oil slick surrounding the ship, which was on station between Hawaii and Japan. The slick extended in all directions to the limits of visibility (six nautical miles), and it was seen repeatedly over the next two weeks as the ship proceeded to other stations in a grid 10 miles by 60 miles. A sample of the slick was collected and immediately frozen. It contained the remnants of copepods, identified as a species of *Calanus*. Chromatographic analysis of the lipid fraction revealed that it consisted of 82 percent wax, 6 percent triglyceride, 4 percent cholesterol, 2 percent phospholipid, 2 percent astaxanthin and 3 percent hydrocarbons. The principal fatty alcohols of the oil slick were found to be similar to those found in the wax of *Calanus plumchrus* and other *Calanus* species. The presence of highly unsaturated fatty acids, which oxidize rapidly, suggested that the massive copepod kill had been fairly recent. It is believed the copepods had died in the Japan Current to the west, and that the remains of the animals, including their wax, had been borne to the location of the Coast Guard cutter by the easterly drift of the North Pacific Current.

Wax has also been observed on the surface in Bute Inlet, a deep fjord 120 miles northwest of Vancouver. In cold winters large quantities of wax accumulate on the shores of the inlet. The source of the wax was a complete mystery, and it was hypothesized that it came from a submarine deposit of "fossil wax." The predominant copepod in the inlet is *C. plumchrus*. Comparison of the length of the carbon chains in the inlet



depths below 1,000 meters contain some wax. Each bar represents the total lipid content (wax and fat) of the animal as a percent of its dry weight. The colored portion of the bar represents the average amount of lipid in each species that is in the form of wax.

wax and in the copepod wax suggests that the copepods were probably the source. The fatty acids and fatty alcohols of the inlet wax, however, are much more saturated than those of the copepod wax. That is probably the result of the degradation of the more unsaturated components by bacteria. The evidence points to some catastrophic destruction of a large population of copepods at some time in the past and the consequent sedimentation of their wax.

In the deep waters of the ocean wax is stored by many marine animals other than copepods. Indeed, most of the animals that live at depths below 1,000 meters contain wax. The coelacanth and the castor-oil fish are examples of fishes with a large wax content. The hagfish and some sharks are unique in having wax in the lipoproteins of their blood serum. (In man and most other vertebrates the same purpose is served by cholesterol esters.) Squids, chaetognaths, amphipods, ostracods and decapods also store wax in varying amounts [*see illustration on preceding two pages*].

Copepods are the principal food at some stage in the life of many fish species that are economically important to man. Anchovies, sardines and herrings prey on copepods. These fishes appear to produce a bile-activated wax lipase, the enzyme that catalyzes the metabolism of wax. The site of wax-lipase production is in the diffuse pancreas of the pyloric cecum, a mass of tubular appendages leading from the junction of the stomach and the small intestine. Neither waxes nor fatty alcohols have been found in the blood of these fishes. The fry of the chum salmon devour so many copepods that they become engorged, and their fecal pellets are so rich in oil that they float. The fry do not completely metabolize the wax, but even so no wax passes into their circulating blood.

As we have noted, copepods that live in the warm surface waters of the Tropics apparently do not store wax. We have recently discovered, however, that waxes play an important role in the food chain of tropical coral reefs.

Corals are half animal and half plant. On the surface of their intricate limestone skeletons reside colonies of polyps (the animal), and within the polyps live dinoflagellate algae (the plant). The algae, with their green chlorophyll *c* and orange carotenoid pigments, give corals their olive brown color.

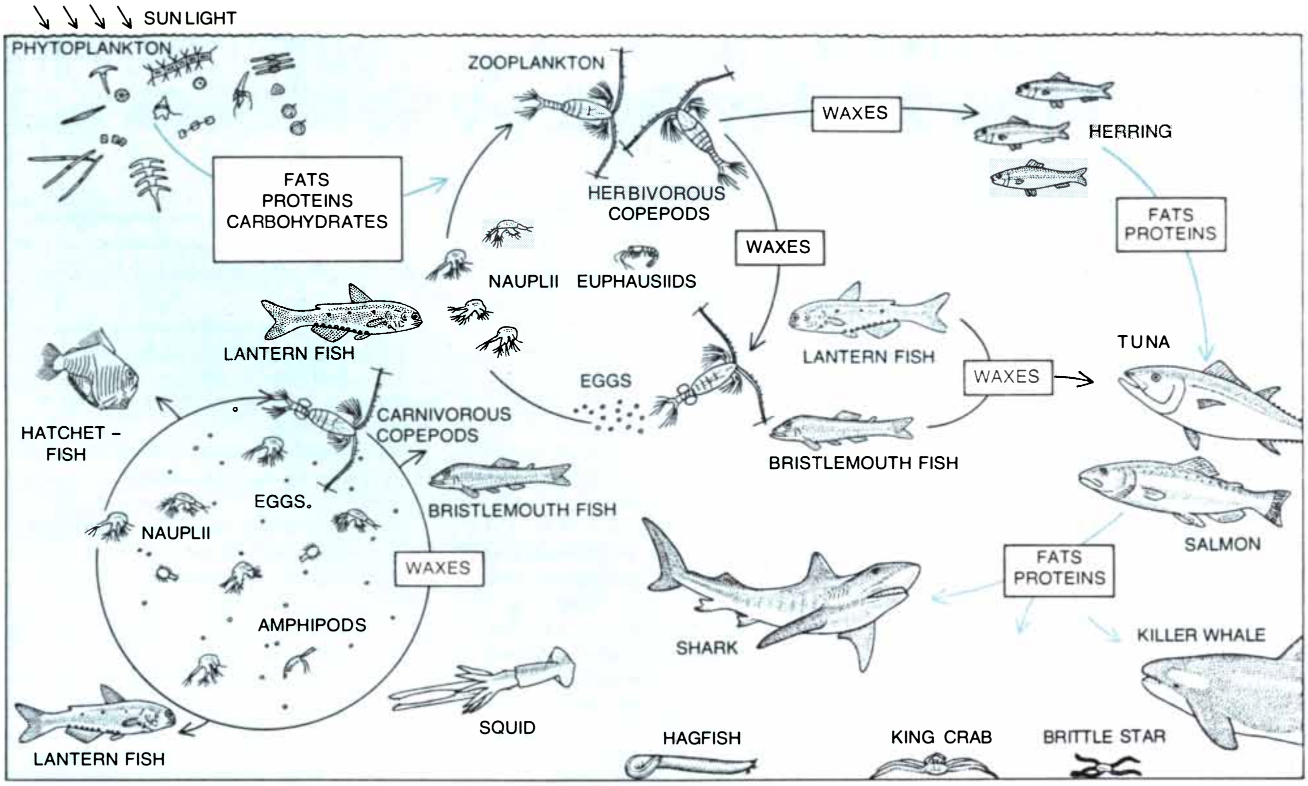
When the algae of coral are separated from the polyps and suspended in seawater, they behave like other algae, conserving the sugar and other substances they manufacture by photosynthesis. (They are related to the algae of "red tides" in temperate and polar waters.) When the alga is in contact with the proteins of the polyp, however, its cell wall becomes leaky. Glycerol, glucose and the amino acid alanine leak out of the alga and into the polyp. The polyp also "steals" fatty acids from the membranes of its algae.

Leonard Muscatine of the University of California at Los Angeles has shown that the algae avidly scavenge the waste products of the polyps, carbon dioxide and ammonia, which they utilize for the synthesis of amino acids. The algae in turn evolve oxygen for the polyps as a by-product of photosynthesis. By keeping these products within a closed, symbiotic system, both the alga and the polyp have evolved a solution to the problems of survival in the nutrient-poor tropical seas.

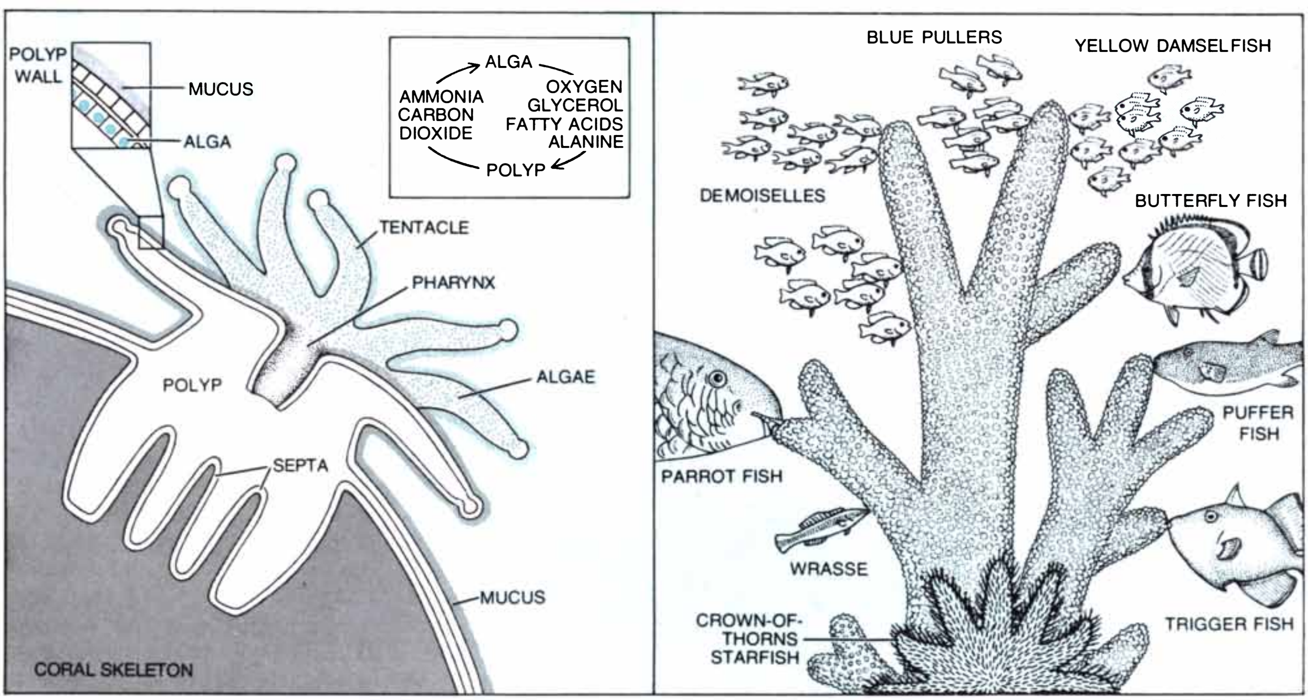
Corals are not generally subject to direct predation by fishes, although some tropical fishes such as the parrot fishes and puffers gnaw at the surface of boulderlike corals and bite off the tips of the more delicate species. Corals do, however, exude a slimy mucus that is ingested by many small fishes. When we brushed mucus from corals into the water, hordes of tiny, colorful coral fishes avidly gathered to eat it. These fish can also be found sucking or nibbling at branched corals. Other fishes, such as the butterfly fishes with their elongated snouts, eat mucus out of the grooves of brain coral.

It had been thought that the fishes were only feeding on tiny zooplankton enmeshed in the coral mucus. The number of zooplankton in the mucus is small, however, and the mucus would be a poor food if it were not for another ingredient in it. We found that coral mucus contains a wax that is similar to the wax of copepods. It should not have been surprising to find that coral mucus contains wax, since the main metabolic energy store of corals is the saturated wax cetyl palmitate. It seems likely that the fishes ingest the mucus for its energy-rich wax content, but the real extent of this food transfer from coral to fishes has not been measured. The fact remains that the transfer occurs, and we may yet find it is a crucial link in the food chain of tropical reefs.

The much publicized predator of Australia's Great Barrier Reef, the crown-of-thorns starfish *Acanthaster planci,* thrives on corals and has reduced miles of productive reef to lifeless rubble. Many biologists are concerned about the reefs' ability to recover. This huge star-



ROLE OF WAX IN THE FOOD CHAIN in Temperate Zone oceans is depicted. The food cycle begins near the surface, where the phytoplankton utilize sunlight to manufacture carbohydrates, proteins and fats. Copepods feed on the phytoplankton and store large amounts of wax. The copepods in turn are the primary food of fishes such as anchovies, sardines and herrings. These fishes have an enzyme that enables them to metabolize wax, but they do not store it. Larger fishes such as tuna prey on the smaller, copepod-eating ones. The deep-water food chain differs in that the copepods are carnivorous: they feed only on other zooplankton. Considerable amounts of wax are stored by these copepods and by deep-water fishes such as bristlemouth fish, lantern fish and hatchetfish.



WAX FOOD CHAIN IN CORAL REEFS has only recently been discovered. Corals are part animal (polyps) and part plant (algae). The polyp and the alga have a symbiotic relationship (*illustration at left*). The alga utilizes the carbon dioxide and ammonia waste products of the polyp, and the polyp obtains oxygen, glycerol, glucose, amino acids and fatty acids from the alga. The coral exudes a mucus that is rich in wax. Numerous small coral fishes (such as damselfish, blue pullers, puffer fish, butterfly fish and small wrasses) feed on the mucus. Larger fishes such as trigger fish and parrot fish break down and devour the coral as well as the mucus. The digestive system of the crown-of-thorns starfish, *Acanthaster planci*, has been found to be adapted to digest the wax in the coral.

Introduction to Contemporary Psychology

Announcing a New Text

by EDMUND J. FANTINO and GEORGE S. REYNOLDS,
University of California, San Diego

An instructor's guide and study guide will be published along with the text. A free test item file will be sent to teachers upon class adoption of the text.

"Although we feel that there is not necessarily any incompatibility among prevailing psychological points of view, we also feel that the introductory student deserves a fair look at, and a sure feel for, all of them."—The Authors

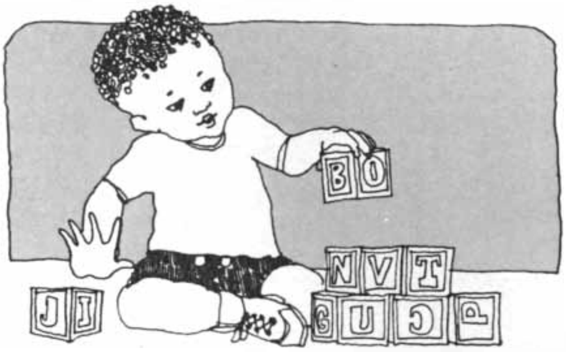
Many students take an introductory psychology course to satisfy an area requirement, others take it as a prerequisite to more advanced courses in psychology, still others take it because they find the subject fascinating. *Introduction to Contemporary Psychology* is a text that will serve all these students well.

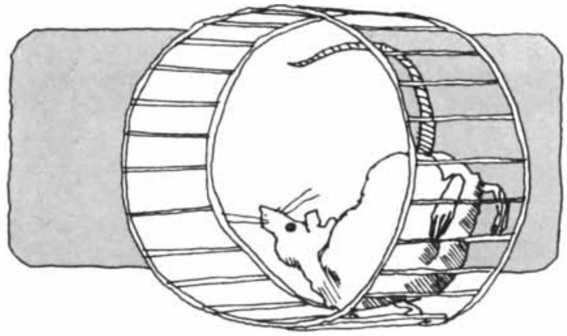
This book takes a new approach to introductory psychology—it is an intentionally eclectic text, treating the subdisciplines and many schools of thought that constitute modern psychology from the point of view of the psychologists actually involved in them. Such an approach gives students an appreciation of the great scope of the science and prepares them for subsequent specialization.

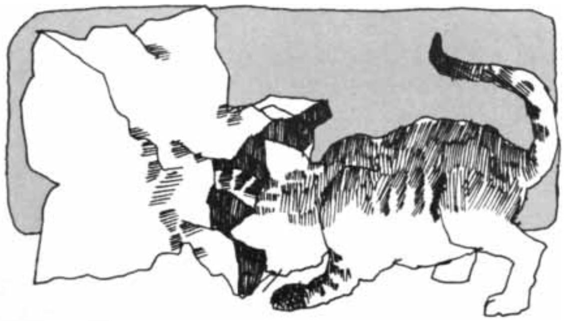
Introduction to Contemporary Psychology is divided into two major parts. The first half of the book deals with such basic psychological processes as learning, remembering, thinking, perception, motivation, instinct, and heredity. Whenever appropriate, the authors have included a discussion of the physiological correlates of behavior. The second half of the book discusses the major subdisciplines of psychology and the important theoretical controversies of the science.

Publication date: February 1975

Good Reading for Anyone Interested in the Many Dimensions of Psychology







Psychology in Progress

Readings from SCIENTIFIC AMERICAN

With Introductions by

RICHARD C. ATKINSON, Stanford University

Psychology in Progress is a collection of 39 *Scientific American* articles, selected to give readers an appreciation of the scope of modern psychology, an acquaintance with the frontiers of psychological research, and clear explanations of basic psychological principles. Some of the topics included are "Pleasure Centers in the Brain," "Teaching Language to an Ape," "Schizophrenia," "The Effects of Observing Violence," and "Intelligence and Race." The book includes a study guide and helpful introductions to the articles, making this valuable supplementary reading for an introductory course.

Publication date: Spring 1975, cloth $12.00, paper $6.25

A Guide to Psychologists and Their Concepts

VERNON J. NORDBY and CALVIN S. HALL

This book belongs in the library of every reader with more than a passing interest in psychology. Here are clear and accurate discussions of the lives and works of 42 men and women who have helped shape the modern behavioral sciences. The book includes a portrait, biography, summary of major ideas, and list of important writings for each psychologist.

1974, 187 pages, 42 illustrations, cloth $8.00, paper $3.50

A Primer of Drug Action

ROBERT M. JULIEN, University of California, Irvine

Here is a sensible new approach to drug education. Organized along the lines of a textbook of pharmacology, this book presents a medically accurate discussion of the major classes of psychoactive drugs—the sedatives, tranquilizers, stimulants, opiates, and psychedelics—at a level easily comprehended by college and advanced high school students. The style throughout is objective, allowing readers to form their own conclusions about the social and moral issues of drug use and abuse.

Publication date: February 1975, cloth $10.00, paper $4.95

A Teacher's Manual is available.

A Primer of Psychobiology

Brain and Behavior

TIMOTHY J. TEYLER, Harvard University

A Primer of Psychobiology is a brief, informative introduction to what is known about the structure and function of the nervous system and how these relate to behavior. Free of jargon and specialized terminology, this book is designed to be readily understood by the introductory level psychology student and the general reader.

Publication date: January 1975, cloth $6.95, paper $3.50

Conceptual Blockbusting

A Guide to Better Ideas

JAMES L. ADAMS, Stanford University

Conceptual blocks are mental walls that get between a problem and its solution. *Conceptual Blockbusting* is a freewheeling but serious guide to identifying and eliminating these blocks. The book is also a good introduction to the classical theoretical writings on creativity—by Freud, Jung, Maslow, and others—and to the modern techniques of applied creativity such as brainstorming and synectics. The book includes exercises and problems designed to improve the reader's "idea-having" abilities.

1974, 137 pages, 42 illustrations, cloth $5.95, paper $3.45

Some Must Watch While Some Must Sleep

WILLIAM C. DEMENT, Stanford University

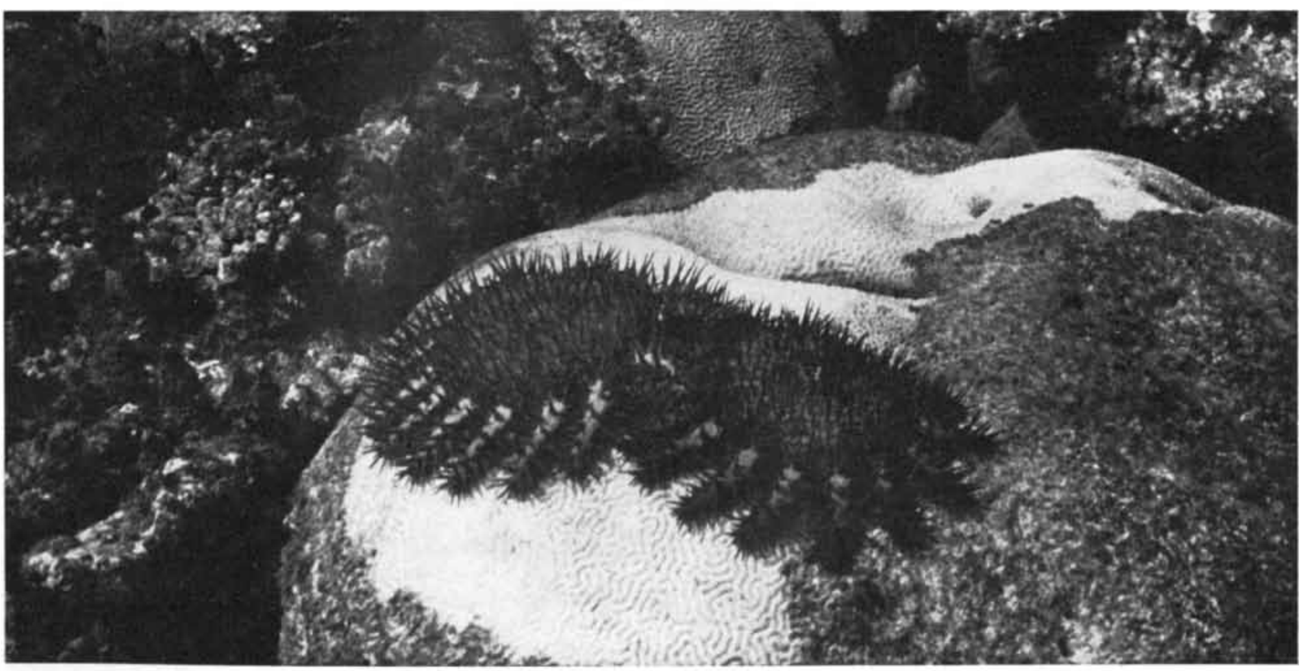
This brief, fact-filled book is an exciting introduction to the modern scientific study of sleep, dreams, and sleeping disorders. Sleep appears to be a highly dynamic state with a complexity rivaling waking consciousness. Among the topics discussed are: the "jet lag" phenomenon; the amount of time spent dreaming; sleep learning; sleep deprivation; and how sleeping pills make insomnia worse.

1974, 148 pages, 30 illustrations, cloth $5.95, paper $2.95

W. H. FREEMAN AND COMPANY
660 Market Street, San Francisco, California 94104
58 Kings Road, Reading, England RG1 3AA

fish (specimens more than two feet across have been collected) everts its stomach over the surface of the coral and exudes strong digestive enzymes; when the starfish moves away, the coral is completely bare. The digestive enzymes of *Acanthaster planci* are highly efficient in their action on the wax components of coral. No other animal we have examined has so thoroughly adapted its digestive system to wax nutrition. Other starfishes lack this ability; they have enzymes only for digesting proteins and carbohydrates. It can be deduced that the crown-of-thorns starfish gets much of the energy it needs for growth and reproduction by metabolizing wax from the coral. This being the case, it might even be possible to develop wax-based baits impregnated with the appropriate hormones for controlling the reproduction of *Acanthaster planci.* (Poisons would harm the reef fishes that feed on waxes.) In this way one might be able to avoid the excessive growth of starfish populations and excessive damage to valuable coral reefs.

POLYPS ON BRAIN CORAL retreat into their holes when they are touched. The surface of the coral is covered with mucus that is secreted by the polyps. This photograph was made by Fritz Goro on the Great Barrier Reef on the west coast of Australia.



TWO CROWN-OF-THORNS STARFISH feeding on brain coral in a bay on the island of Guam are shown in this photograph made by Thomas F. Goreau. The heavy-spined starfish, which can grow to a size of two feet or more, everts its stomach over the living coral and digests it. When the starfish has finished feeding, only the limestone skeleton of the coral remains. The wax stored in the coral animal is an important part of the diet of the starfish. In the late 1960's the crown-of-thorns starfish began to multiply rapidly throughout the South Pacific, threatening the complete destruction of the coral reefs that surround and protect many of the islands in the region.

Compare TI's NEW SR-51....

Mean, variance and standard deviation. Automatic linear regression. 20 programmed conversions. Percent and percent difference. Random number generator. 3 accessible memories. And much more for only $224.95

FUNCTION	SR-51	HP-45
Log, ln	yes	yes
Trig (sin, cos, tan, Inv)	yes	yes
Hyperbolic (sinh, cosh, tanh, Inv)	yes	no
Degree-radian conversion	yes	yes
Deg/rad mode selection	yes	yes
Decimal degrees – deg-min-sec	yes	yes
Polar-rectangular conversion	yes	yes
y^x	yes	yes
e^x	yes	yes
10^x	yes	yes
x^2	yes	yes
$\sqrt{x}$	yes	yes
$\sqrt[x]{y}$	yes	no
1/x	yes	yes
x!	yes	yes
Exchange x with y	yes	yes
Metric conversion constants	13	3
% and Δ%	yes	yes
Mean and standard deviation	yes	yes
Linear regression	yes	no
Trend line analysis	yes	no
Slope and intercept	yes	no
Store and recall	yes	yes
Σ to memory	yes	yes
Product to memory	yes	yes
Random number generator	yes	no
Automatic permutation	yes	no
Preprogrammed conversions	20	7
Digits accuracy	13	10
Algebraic no tion (sum of products)	yes	no
Memory (other than stack)	3	9
Fixed decimal option	yes	yes
Keys	40	35
Second function key	yes	yes
Constant mode operation	yes	no

More math power for the money. More than log and trig and hyperbolics and functions of x . . . the SR-51 has these and also has *statistical* functions . . . like mean, variance and standard deviation, random numbers, factorials, permutations, slope and intercept, and trend line analysis. Check the chart above – compare it. With the HP-45 or any other quality calculator. Then try it – at no risk. We're sure you'll agree that the SR-51 offers extraordinary value.

Test the SR-51 at no risk. Full refund if not satisfied. Master Charge or Bank Americard accepted.

TO: Texas Instruments Incorporated
P.O. Box 22013, M/S 358, Dallas, Texas 75221

Try it 15 days. Return with all accessories if not satisfied.
Enclosed is my ☐ check, ☐ money order, ☐ company purchase order for $ ________ for the purchase of ____ SR-51(s).
Please add state and local taxes where applicable.*
Please charge this order to my ☐ Master Charge. Or, ☐ BankAmericard
My Card Number is: ☐☐☐☐ ☐☐☐ ☐☐☐ ☐☐☐
With the expiration date: ____________

SA3
If Master Charge is used indicate 4-digit Bank Number appearing on card just above your name ☐☐☐☐

Name ____________
Street ____________
City ________ State ________ ZIP ________

Authorizing Signature (If billed to Credit Card Must be signed by card holder)

*AL, AZ, CA, CO, CT, FL, GA, IA, IL, IN, KY, MA, MD, MI, MN, MO, NC, NM, NJ, NY, PA, TN, TX, UT, VA, WA, WI. Offer good only in U.S.

TEXAS INSTRUMENTS
INCORPORATED

82016

The Most Poisonous Mushrooms

Certain fungi of the genus Amanita contain toxins of extraordinary virulence. They are ring-shaped molecules made up of amino acids, and they disrupt the membranes and the nuclei of cells in the liver

by Walter Litten

Wild mushrooms are a prized delicacy, but dining on them is notoriously hazardous. Their reputation for danger is not unjustified; the risk, at least to the naïve forager, is documented in the case histories of those whose adventurous appetites were restrained by neither prudence nor training. There are many mushrooms that are toxic to man, including some common and widely distributed ones, and there are a few that must be regarded as lethal. Furthermore, there is no simple test for distinguishing the edible species from the poisonous ones. The counsel of folklore is useless; for example, whether or not a cooked mushroom blackens silver has no bearing on its edibility. Safety consists in knowing exactly what one is eating, and that knowledge can often be attained only by discerning rather subtle features of fungal anatomy.

Among the poisonous mushrooms almost all the most deadly ones are members of a single genus: *Amanita.* In Europe from 90 to 95 percent of all deaths from mushroom poisoning have traditionally been attributed to a single species of that genus, *Amanita phalloides.* Mushroom gathering is more popular in Europe than it is in the U.S., and *A. phalloides* is a common European species. In North America *A. phalloides* has long been considered rare, but in the past few years it has been implicated in at least two incidents of mushroom poisoning, including two deaths. At the same time there is evidence to suggest that in Europe the incidence of poisoning by amanitas other than *A. phalloides* has been underestimated.

The substances that make the amanitas so dangerous are quite different from those encountered in most other poisonous mushrooms. Compared with most other poisons, *Amanita* toxins are rather complex molecules, and they may form aggregates of even greater complexity. Their physiological effect, in man and in other vertebrates, has only recently been explored in detail; it involves, in addition to disturbances of metabolism, the disruption of the nuclei of certain cells. For these reasons the amanitas are of interest not only to mycologists, the investigators who study mushrooms and other fungi, but also to biochemists and to biologists generally. In addition they are, of course, a matter of concern to the physician. Mushroom poisoning is hardly a major cause of human mortality, but an antidote to the *Amanita* toxins has long been sought.

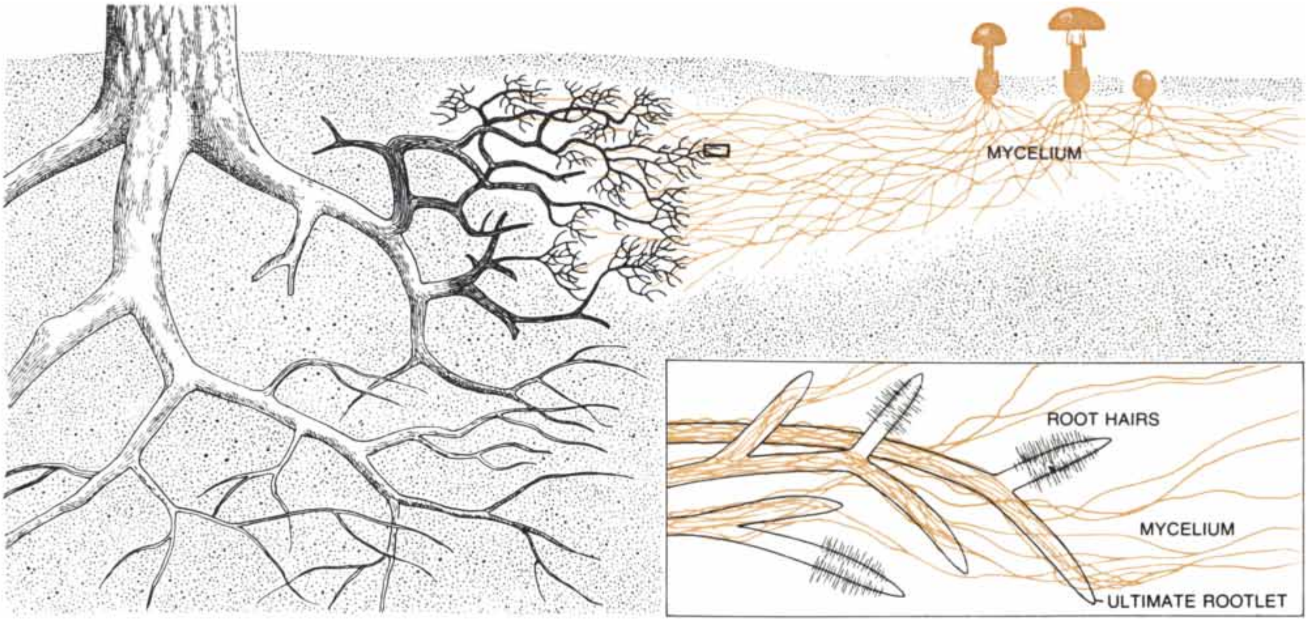
A mushroom is the reproductive apparatus of a fungus; the rest of the organism consists of a network of filaments only one cell wide. The network is the mycelium; the individual filaments are hyphae. The mushroom is often referred to as the fruit body of the fungus, and its relation to the mycelium is analogous with that of an apple to an apple tree [*see top illustration on next page*].

The mycelium of a fungus can extend through the topsoil for hundreds of feet, or through a fallen log, a carpet of leaves or conifer needles or a living tree trunk. It may live for many years, growing into new territory each season and dying off where it has exhausted the supply of the nutrients it requires. The fruit body, on the other hand, typically functions for only a few days, then withers. It begins as a knot of hyphae that soon grows to form a fleshy button. At this stage of development some mushrooms, including those of the genus *Amanita,* are enclosed by a membrane called the universal veil. As the fruit body grows and emerges from the soil, the universal veil ruptures, exposing the familiar structures of the mushroom: the stipe, or stalk; the pileus, or cap, and the gills, the vertical plates that radiate from the stalk on the underside of the cap. In some species the gills are at first protected by another membrane, the partial veil. After it has ruptured, remnants may cling to the stalk as a ring, the annulus.

The surfaces of the gills are covered with cells called basidia, which produce the spores of the fungus; in most cases each basidium bears four spores. When the mushroom is mature, the spores are ejected and carried away by the wind; those that land in a suitable place can germinate and establish a new mycelium. Picking the mushroom does not destroy the fungus; it may help to perpetuate the species by dispersing the spores. Because the spores are the least variable and most durable part of the fungus, the determination of their shape, color and other characteristics is often important in the identification of species.

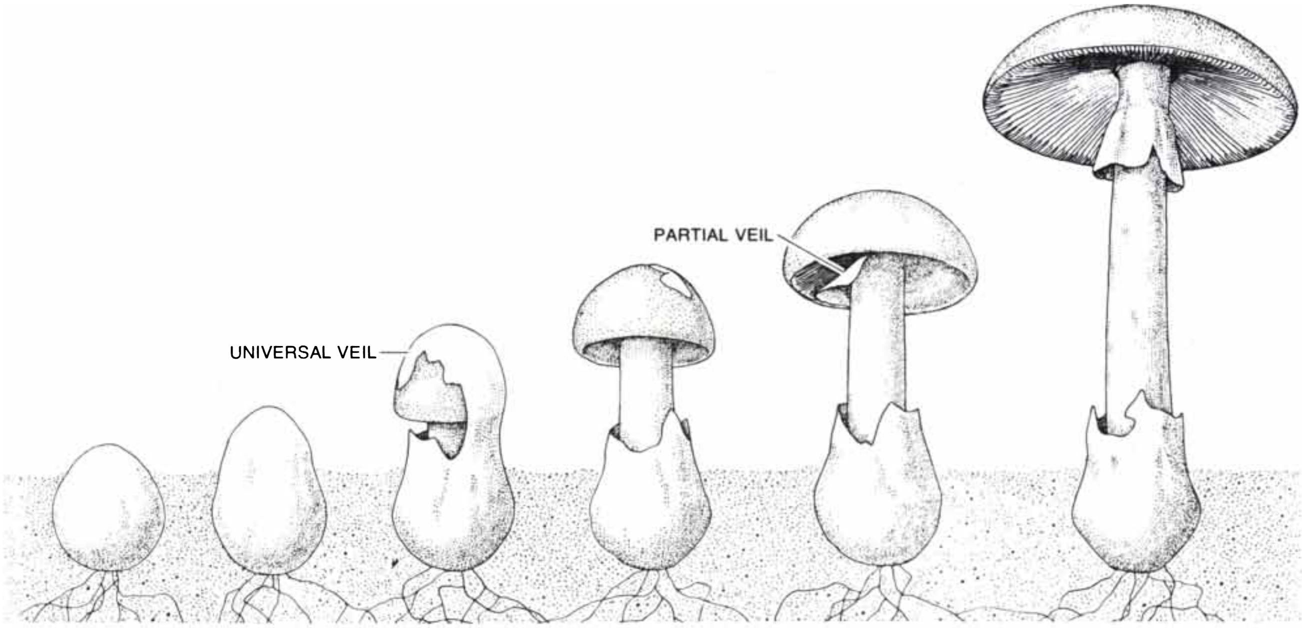
Many fungi are saprophytic, that is, they obtain their food by decomposing dead organic matter. Some are parasitic and invade the wood of living trees.

GREEN-CAPPED MUSHROOM is a fruit body of the fungus *Amanita phalloides.* It is extremely poisonous, causing damage to the liver and kidneys that often results in death. The species has long been a menace to European mushroom gatherers, but until recently it was considered rare in North America. This specimen was found in 1971 in a stand of pine trees in Cape May County in New Jersey. A year earlier six people were poisoned after eating mushrooms picked under the same trees, and two died. The photograph, which enlarges the mushroom, was made by Donald M. Simons; dark background is an artifact of the lighting.



STRUCTURE OF A FUNGUS includes not only the mushroom but also the mycelium: a network of hyphae, or filaments, that permeate the soil or some other medium. Some mycelia participate in a symbiotic association with higher plants in which the hyphae (*color*) and the ultimate rootlets of a tree (*black*) form a joint structure called the mycorrhiza (*detail drawing*). Many amanitas, and perhaps all of them, form mycorrhizae. The mycelium can live for many years, invading new territory as it grows; the mushroom, on the other hand, is an ephemeral reproductive body that appears in season and decays in a few days or weeks. Its only function is to produce and disperse microscopic spores that, if they come to rest in a suitable habitat, can germinate to establish a new mycelium.

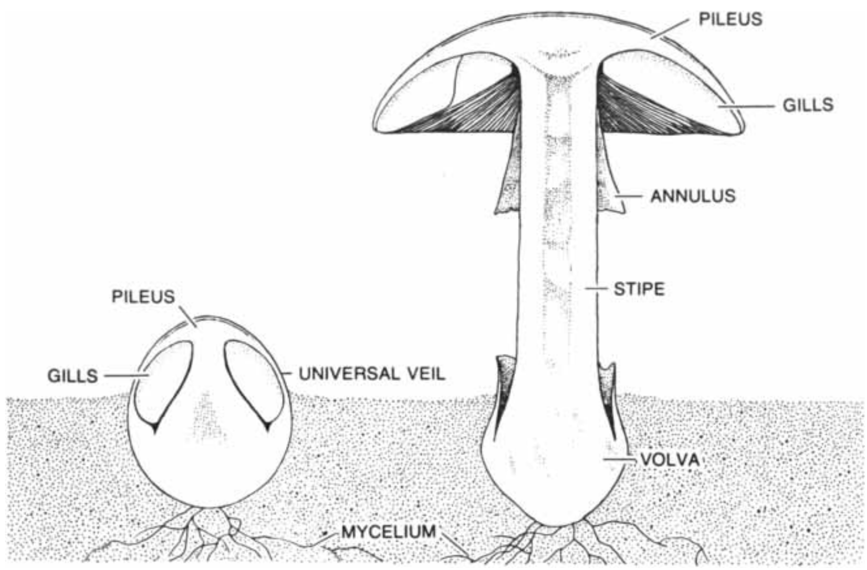
There are also a large number that form symbiotic associations with green plants. The hyphae of these fungi and the ultimate rootlets of a tree form a joint structure called the mycorrhiza. The fungus apparently derives part of its nutritional requirements from the mycorrhiza, and some species may be entirely dependent on it. The nature of the benefits the tree receives from the association is uncertain, but it has been observed that some trees die young if the appropriate fungus is not present. Some mycorrhizal partners require a particular symbiont, whereas others can accept any one of a large number of species. Many amanitas are known to form mycorrhizal associations, and some mycologists believe all members of the genus share this capability. If *A. phalloides* is mycorrhizal, it is not very selective; in Europe it grows



GROWTH OF A MUSHROOM begins with the formation of a knot of hyphae in the soil. Mushrooms of the genus *Amanita* are initially enveloped by a thick membrane, the universal veil; when this membrane ruptures, a portion is left surrounding the base of the stalk as a jagged cup. The gills, on the underside of the cap, are protected by a second membrane, the partial veil, which also breaks away. As the mushroom reaches maturity, the cap flattens and the outer edge may turn upward. The deadly amanitas are poisonous in all stages of development; because they change in form and color as they grow, identification in the field requires discernment.

near both hardwoods and conifers.

The fruit bodies of *A. phalloides* generally appear in late summer or fall. They vary widely in size: mature specimens may stand from three to eight inches tall. The coloration is also somewhat variable. The cap ranges from light yellow to greenish brown; the stipe is usually lighter, from a light greenish yellow to pure white. The identification of *A. phalloides* in the field demands attention to crucial details. Less virulent species (and one esteemed edible mushroom, *Tricholoma equestre*) have a similar coloration, habitat and season and may share the same mycorrhizal partners. The appearance of the fruit bodies can change rapidly as they mature [*see bottom illustration on opposite page*]. Relying on a single photograph for identification could have grave consequences.



ANATOMY OF THE AMANITAS is distinguished by two main features: the volva and the annulus, which are remnants respectively of the universal veil and the partial veil. The coloration and texture of the stipe, or stalk, and the pileus, or cap, are also identifying characteristics. The spores are produced by specialized cells on the surface of the gills.

One of the principal identifying marks of the deadly amanitas is the volva, a cuplike remnant of the universal veil at the base of the stalk. In *A. phalloides* the edges of the volva are jagged and stand well out from the stalk. Pieces of the universal veil are also sometimes found adhering to the top of the cap, where they form irregular white patches. The annulus, the remnant of the partial veil, hangs loosely from the stalk below the gills like a white dinner napkin.

The presence of well-defined traits such as these might suggest that identifying the deadly amanitas is a simple and straightforward matter. Actually there are variations from the norm and intergradations between species that can make the task a perilous one. Even accomplished mycologists have erred; indeed, confusion persisted for at least three decades over the status of *A. phalloides* in North America.

In the early years of this century an American fungus was identified as *A. phalloides* by Charles Horton Peck, a distinguished mycologist who was state botanist of New York for more than 50 years. Other American mycologists concurred in the identification. It was subsequently discovered, however, that the American mushroom differs from European specimens in that it has a more fragile volva and a sharp-rimmed, bulbous enlargement at the base of the stalk. In 1918 it was recognized as a distinct species by G. F. Atkinson of Cornell University. Atkinson called attention to another distinguishing trait of the American species: it tends to develop brown stains on the stalk where it has been bruised by handling, and he therefore named it *Amanita brunnescens.* This mushroom is poisonous, but its toxins differ chemically and in physiological effect from those of other amanitas.

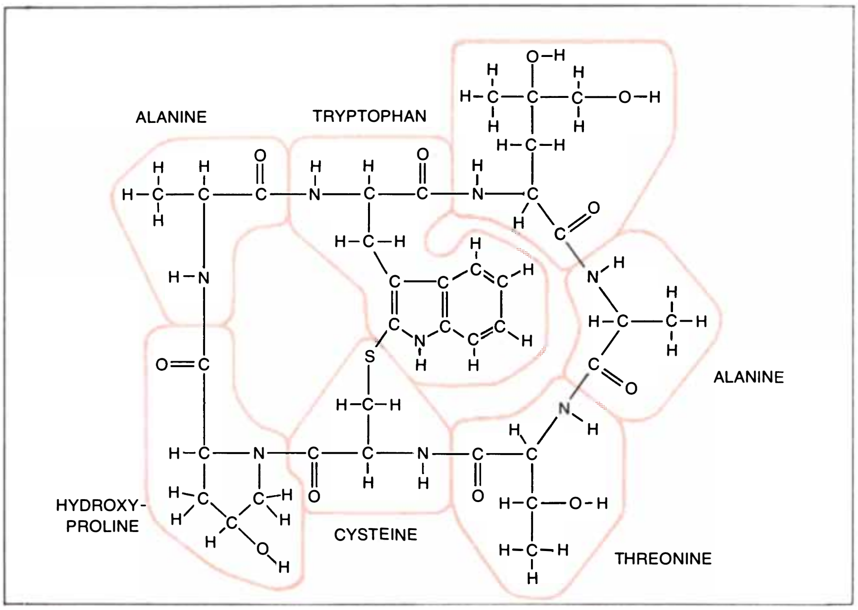
The difficulty of telling these two mushrooms apart is compounded by the presence of another species, *Amanita rubescens,* which resembles *A. brunnescens* but when bruised stains red instead of brown. Moreover, *A. rubescens* is edible; in fact, it is considered a delicacy, particularly in Britain, where it has been nicknamed "the blusher." Still another amanita, *A. citrina,* common both in the eastern regions of North America and in Europe, has a coloration that could be confused with that of *A. phalloides* and a bulbous base with a sharp rim like that of *A. brunnescens* (although it does not stain when it is handled). *A. citrina* was long regarded as a poisonous species under the name *Amanita mappa.* Its edibility was demonstrated at the 1925 banquet of the Mycological Society of France, when it was served without ill effect.

Another large group of amanitas is distinguished principally by lack of coloration. *Amanita virosa,* for example, is pure white throughout; it is less common in Europe than *A. phalloides,* but it grows in many parts of eastern North America. Equally common in the U.S. (even more so in my experience) is *Amanita bisporigera,* which differs from *A. virosa* in that each basidium bears only two spores instead of the usual four. A species indigenous to the southern U.S., *Amanita tenuifolia,* can best be identified by the shape of its spores: they are narrowly elliptical, whereas those of similar species are globose. Finally, *Amanita verna* can be identified by testing it with a drop of a weak solution of potassium hydroxide; *A. virosa* and *A. bisporigera* stain yellow in response, whereas *A. verna* shows no reaction. The entire group of white amanitas should be regarded as deadly.

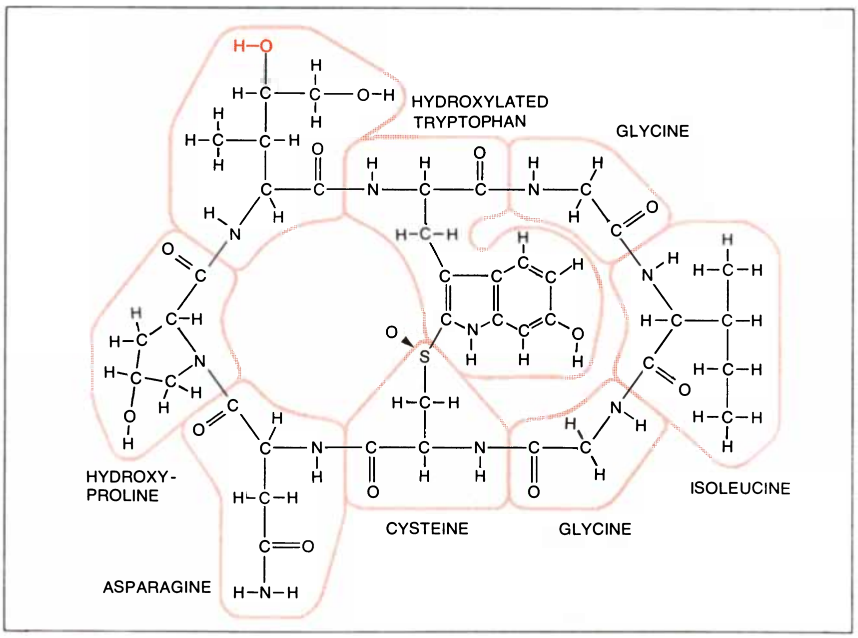
Still another group of fungi has recently been found to contain some of the toxins present in the deadly amanitas. They are members of the genus *Galerina* and are small, brown mushrooms that would probably not ordinarily be gathered for the table. One of them, *Galerina venenata,* was first collected on a lawn in Portland, Ore.

The question of the occurrence of *A. phalloides* in the U.S. was not settled with Atkinson's description of *A. brunnescens* in 1918. In many years of fieldwork Alexander H. Smith of the University of Michigan was able to verify the collection of only a few specimens of the true *A. phalloides* in the U.S., all of them found in the Western states. By the 1970's, however, it had become evident that the species is more widely distributed than had been believed and that it can be found in the East. From the standpoint of the unfortunate diner the issue is perhaps a trivial one, since the white amanitas are at least as dangerous as the green one. Regardless of taxonomic distinctions between mushrooms of the two groups, the consequences of eating them have proved equally difficult to overcome.

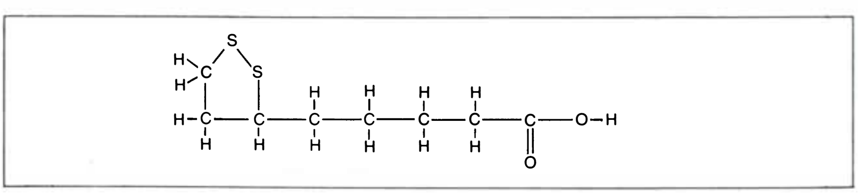
Most poisonous fungi other than the amanitas produce symptoms soon after



PHALLOTOXINS are cyclic, or ring, molecules made up of seven amino acid units (*colored outlines*). A sulfur atom connects the side chains of two amino acids on opposite sides of the ring. Shown is phalloidin; in related compounds substitutions are made in peripheral groups. Only amino acids that are among the 20 commonly found in proteins are identified.



AMATOXINS have eight amino acid units instead of seven. Again a sulfur atom joins two side chains; hydroxyl group in color is essential for toxicity. The toxin shown is alpha-amanitin. This formula and the one at the top of the page were worked out by Theodor H. F. Wieland and his co-workers at Max Planck Institute for Medical Research in Heidelberg.



THIOCTIC ACID, a possible antidote to amanita poisons, consists of a ring of carbon and sulfur atoms and a chain of carbon atoms. The molecule shown is 6-thioctic acid, or alpha-lipoic acid; other isomers have more of the carbon atoms in the ring and fewer in the chain.

ingestion. The toxins in certain species of the genera *Clitocybe* and *Inocybe*, for example, take effect in an hour or two. Some of these fungi are found on lawns (although not as commonly as harmless species), and they must be taken into account when a child has picked and eaten a mushroom. They contain muscarine, a toxin that causes blurred vision, sweating, labored breathing, increased peristalsis, reduced heart rate and blood pressure, and in severe cases convulsions. Fortunately a specific antidote to muscarine is known: atropine, the alkaloid obtained from the belladonna plant.

Muscarine is named for the fungus *Amanita muscaria*, from which it was first extracted in 1869. *A. muscaria* is now known to contain much less muscarine than the *Clitocybe* and *Inocybe* species, but it contains other physiologically active substances as well. Among these is muscimol, an intoxicant for which *A. muscaria* was allegedly employed in the rituals of certain Asian peoples. Muscimol is a degradation product of ibotenic acid, an insecticide, which is also present in the fungus. The mushroom attracts flies and kills them, and is popularly known as the fly amanita.

Another mushroom toxin is psilocybin, found in certain species of the genera *Psilocybe*, *Panaeolus*, *Stropharia* and *Conocybe*. It is a hallucinogenic agent, and its effects become apparent about 15 minutes after the mushrooms are eaten. Monomethylhydrazine, a relatively simple organic compound, causes occasional fatalities from the ingestion of *Gyromitra esculenta*, a "false morel" that is taxonomically remote from gilled fungi such as the amanitas. None of these toxins is related chemically to those that characterize *A. phalloides*, nor are the symptoms of poisoning by these substances similar to those that ensue when *A. phalloides* is eaten. The less virulent toxins are rarely fatal.

Rumor has it that *A. phalloides* makes a tasty dish. In most cases it is not swallowed innocently by a child but is eaten deliberately by a mushroom gatherer. Typically it is encountered and collected in anticipation of an unusual delicacy, tested to see if it blackens a silver spoon or coin, and pronounced fit for the table.

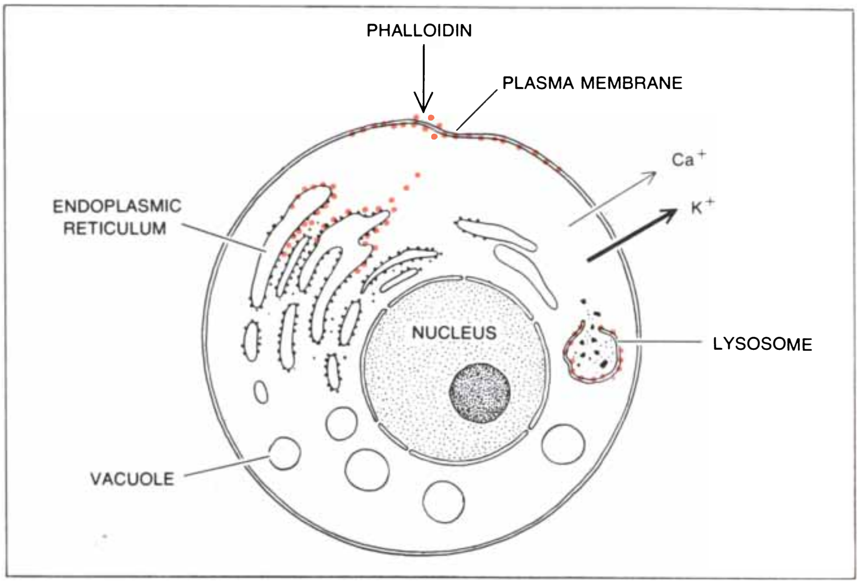
The symptoms of poisoning do not appear for many hours after the mushrooms have been eaten. A full day may pass without ill effect; another meal or two may have been eaten and the unusual mushrooms may have been forgotten. The symptoms begin with abdominal pain, followed by diarrhea and violent emesis, even though the mushrooms have by then been completely digested. If appropriate medical care is not provided, and if more than one cap of average size has been consumed, the gastrointestinal distress can be expected to intensify and eventually cause death.

With the supportive measures available in a modern hospital the gastrointestinal disturbance can be alleviated and the dangerous dehydration it leads to can be corrected. The patient is made comfortable, his condition apparently improves and a complete remission of symptoms seems imminent. After another day or two, in about a third of the recorded cases, the patient dies. An autopsy discloses extensive damage to the liver and the kidneys.

The chemistry of the toxic agents in *A. phalloides* has been under investigation for many years. In 1937 Feodor Lynen and Ulrich Wieland of the University of Munich Chemical Laboratories crystallized from the fungus a toxic substance they named phalloidin. In 1941 Heinrich O. Wieland and Rudolf Hallermayer of the same laboratory isolated a somewhat more toxic constituent they called amanitin. The two toxins have been the subject of a long series of subsequent investigations by Theodor H. F. Wieland and his co-workers, first at Johann Wolfgang Goethe University in Frankfurt am Main and since 1968 at the Max Planck Institute for Medical Research in Heidelberg. Theodor Wieland has shown that phalloidin can be fractionated into six compounds, termed the phallotoxins, and that amanitin consists of another five or more compounds, the amatoxins. (The dangerous *Galerina* species contain amatoxins but not phallotoxins.)

All the phallotoxins and amatoxins are compounds of the type called cyclopeptides. They consist of amino acids linked together by peptide bonds to form a continuous ring. An amino acid is an organic compound that has both an amino group ($-NH_2$) and a carboxyl group ($-COOH$); the peptide bond joins these groups by the elimination of a molecule of water forming the linkage $-CO-NH-$. Of the amino acids in the phallotoxins and amatoxins some are among the 20 that make up all proteins, but some are less familiar.

In the phallotoxins the cyclopeptide ring has seven amino acids; in the amatoxins it has eight. In both types of compound a sulfur atom connects side chains of amino acids on opposite sides of the ring, thus creating a bridge across the structure [*see top and middle illustrations on opposite page*]. Wieland has established that certain portions of the molecules are crucial to their toxicity. For example, in the amatoxin amanullin the elimination of a single hydroxyl group from one amino acid abolishes the toxicity of the molecule. Generalizations about the functional significance of the molecular structure cannot yet be reliably made, however; the removal of the corresponding hydroxyl group in the phallotoxins has little effect.



EFFECTS OF PHALLOIDIN on an isolated liver cell consist mainly of the disruption of membranes. The plasma membrane that surrounds the cell is attacked, and the cell begins to "leak" ions. Calcium ions escape first, followed by a massive efflux of potassium ions. Phalloidin also binds to some membranes inside the cell, such as those of the endoplasmic reticulum. If the membranes of the lysosomes are ruptured, the organelles release enzymes that destroy the cell. Experiments suggest that the number of vacuoles, which are also bounded by membranes, increases on exposure to phalloidin. In spite of these manifestations of toxicity phalloidin does not appear to be responsible for clinical amanita poisoning.

Caution is also required in any attempt to compare the results of laboratory experiments employing the purified toxins with the clinical symptoms presented when *A. phalloides* is prepared and consumed as food. In laboratory studies the influence of differing rates of absorption of the various toxins is usually circumvented, yet these differences may have significant consequences. Moreover, Wieland has demonstrated that various animals differ widely in their sensitivity to the injected toxins. In proportion to body weight the frog and the toad, for example, can tolerate about 10 times as much of the amatoxin alpha-amanitin as the rat, and the rat in turn is 10 times as tolerant as the mouse. The guinea pig, the dog and the rabbit, on the other hand, are more sensitive than the mouse.

In the laboratory phalloidin administered intraperitoneally has an almost immediate effect. When the animal is given a sufficiently large dose, it dies within an hour or two. In contrast, the amatoxins do not cause death in less than 15 hours, no matter how large the dose. The toxicity of alpha-amanitin is nevertheless from 10 to 20 times greater than that of the phallotoxins, in spite of its slower action. If they are delivered to the bloodstream, both the phallotoxins and the amatoxins attack the liver, but by different mechanisms.

The almost instantaneous effect of phalloidin on liver tissue has been demonstrated by Werner Jahn of Wieland's laboratory in an experiment employing a rat liver rendered almost transparent by the removal of all red blood cells. The toxin was introduced into a fluid with which the liver was perfused, so that it quickly reached all parts of the tissue. Within two or three minutes the tissue lost some of its transparency because of the formation of vacuoles within the liver cells. Simultaneously oxygen consumption increased and calcium ions were released; shortly afterward there was a massive exodus of potassium ions.

These events suggest a destructive mechanism involving cellular membranes, and phalloidin is known to bind to the plasma membrane of liver cells. Membranes enclosing some of the cell's organelles are also attacked, including those of the lysosomes, which contain digestive enzymes. If these membranes are ruptured, the cell is killed. Membrane damage can be reduced by a natural antagonist to phalloidin, found in low concentration in *A. phalloides* along with the toxins. The antagonist, called antamanide, is a cyclic peptide of 10 amino acid units.

The points of attachment to which phalloidin binds are the protein molecules on the surface of the cellular membrane. Wieland and his co-workers have recently shown that an interlaced network of filaments, each with the appearance of a string of beads, is formed when phalloidin becomes bound to these proteins [*see illustration below*]. They suggest that the filaments consist of actin, a protein with an essential role in muscle contraction and in the construction of certain intracellular structures. Cells functioning normally are able to form and destroy actin filaments as they are needed; in liver cells phalloidin may inhibit the destructive pathway.

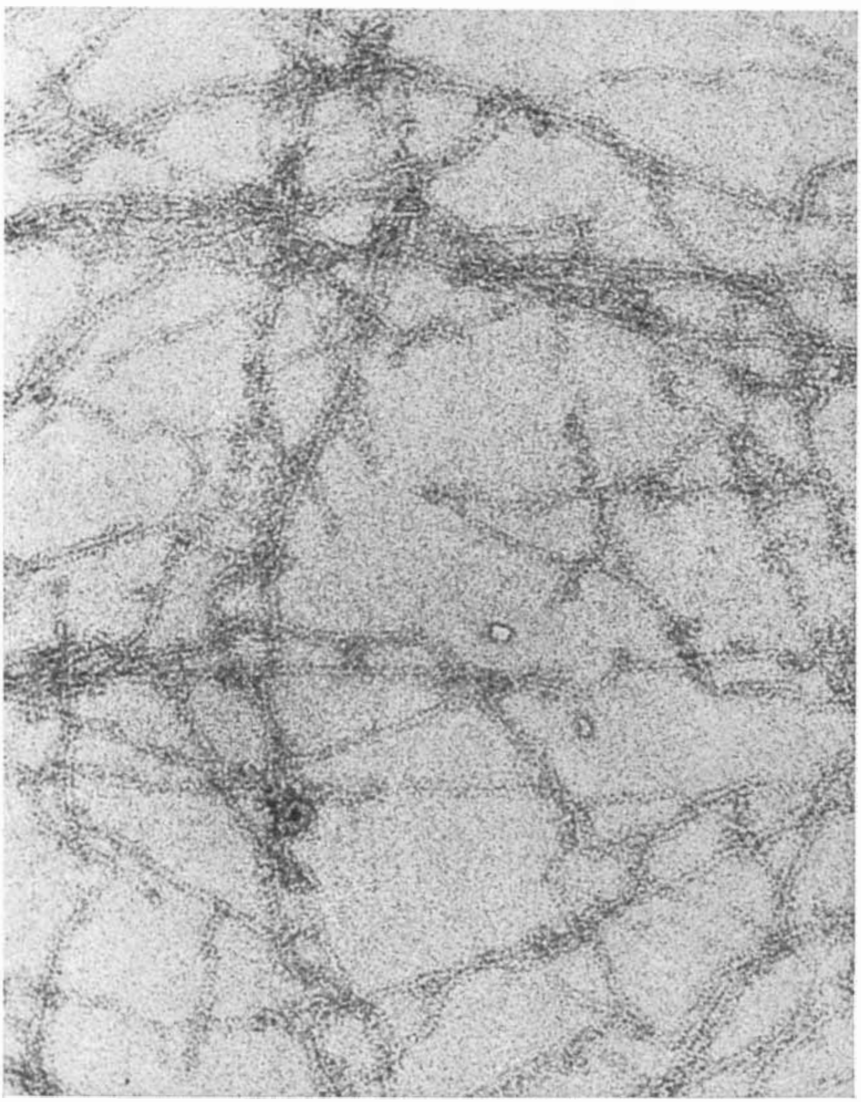
The mechanism of liver-cell destruction in phalloidin poisoning has also been studied by Max Frimmer and his colleagues at Justus Liebig University in Giessen in West Germany. In isolated rat liver cells Frimmer found that phalloidin produced protrusions in the cell membrane in about 10 minutes. The protrusions were apparently caused by the pressure of the cytoplasm inside the cell on weakened sections of the membrane. In cells tested in the intact liver, where the pressure is higher outside the cell than it is inside, invaginations developed instead of protrusions, again suggesting that the toxin weakens sections of the membrane. In both experiments the stretched membranes leaked potassium ions [*see illustrations on opposite page*].

Sensitivity to phalloidin is diminished in the immature liver and in one that has been damaged by agents such as carbon tetrachloride. The reduced susceptibility may result from a deficiency or a defect in the proteins that bind phalloidin. Similarly, some substances that have been shown to provide a degree of protection against the phallotoxins are believed to compete with them for binding sites on the membrane. Frimmer has found that liver cells can be made insensitive to phalloidin by treatment with trypsin, a digestive enzyme. Trypsin degrades many kinds of protein, and in liver cells it apparently disrupts the proteins that bind phalloidin.

In spite of the proved toxicity of phalloidin to liver tissue, several investigators now doubt that it participates at all in human mushroom poisoning. Its potency when injected into the bloodstream has not been questioned, but the efficiency of its transport from the stomach to the blood is in doubt. Wieland has pointed out that no animal has been demonstrably poisoned by the oral administration of phalloidin. Moreover, when phalloidin is administered by injection, death results in an hour or two, whereas those who have eaten deadly amanitas experience no symptoms at all for at least eight hours. Finally, studies conducted by Sami H. Abdel-Malak at the University of Maine suggest that *A. rubescens* contains phalloidin; since this species is known to be edible, it seems unlikely that phalloidin is the agent of human mushroom poisoning.

The toxic properties of the amatoxins are quite different from those of the phallotoxins: damage caused by alpha-amanitin is first observed not in the cell membrane but in the nucleus. Furthermore, there is no question of the toxicity of the amatoxins, or indeed of their lethality. They are effective whether they are given by injection or by mouth, and when they are administered orally, at least to some experimental animals, the complex of symptoms characteristic of amanita poisoning develops in the expected sequence and in approximately the expected time.

Luigi Fiume and his co-workers at the University of Bologna have shown that when alpha-amanitin is perfused directly into liver tissue, the nucleoli in the liver-cell nuclei begin to disintegrate in



TANGLED FILAMENTS may be the agent of cell destruction in experimentally induced phalloidin poisoning. In this electron micrograph, made by Anneliese M. Lengsfeld of the Max Planck Institute for Medical Research in Heidelberg, they are seen in a homogenate of membranes from rat liver cells. The filaments probably consist of the polymeric protein actin. The cell can ordinarily polymerize actin or degrade it as necessary; phalloidin is thought to inhibit the degradation. As a result the filaments accumulate in the cytoplasm.

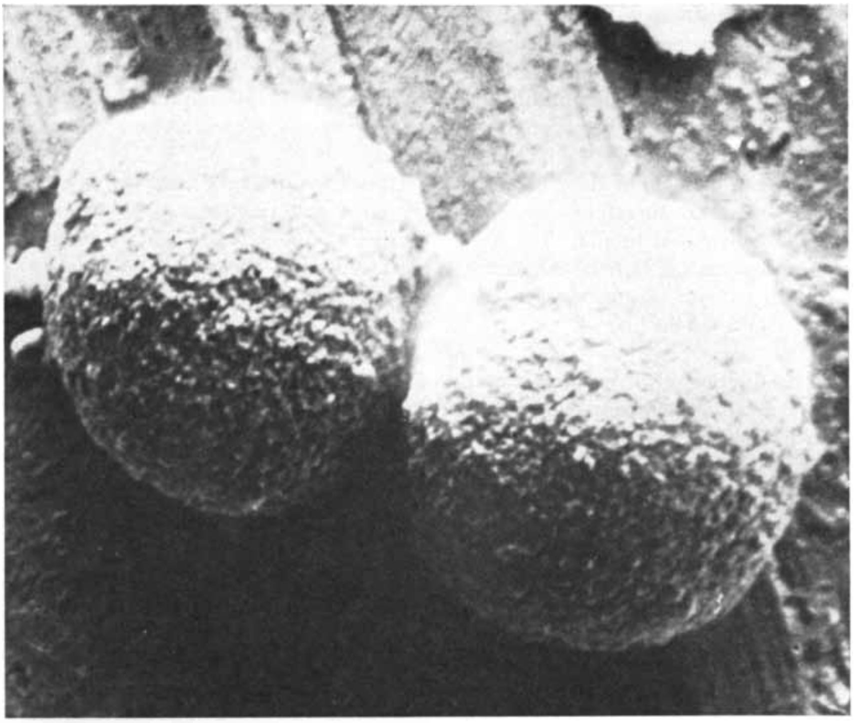
minutes [*see illustration on next page*]. The nucleoli are the organelles in which the ribosomes are built out of protein and ribosomal RNA. The ribosomes are subsequently exported to the cytoplasm of the cell, where they attach themselves to membranes and serve as the site of protein synthesis. The nucleoli contain a high concentration of RNA; indeed, most of the RNA in the nucleus is in the nucleoli. Fiume and his colleagues have found that within an hour after exposure to alpha-amanitin the nucleus loses much of its RNA.

Fiume and F. Stirpe have identified a specific enzyme in the cell nucleus that is inhibited by alpha-amanitin. It is an RNA polymerase, an enzyme that catalyzes the synthesis of RNA, but curiously it is not the enzyme that is involved in the synthesis of ribosomal RNA. The amanitin-inhibited enzyme, which is characterized by a requirement for manganese, directs the synthesis of messenger RNA during the transcription of DNA: the fundamental process in which the genetic information of the cell is "read" to produce a template for protein synthesis. Another RNA polymerase, which is dependent on magnesium rather than manganese, mediates the synthesis of ribosomal RNA in the nucleolus; amanitin does not directly inhibit this enzyme. Because of alpha-amanitin's ability to inhibit a single crucial enzyme it is of interest as a research tool in molecular biology.

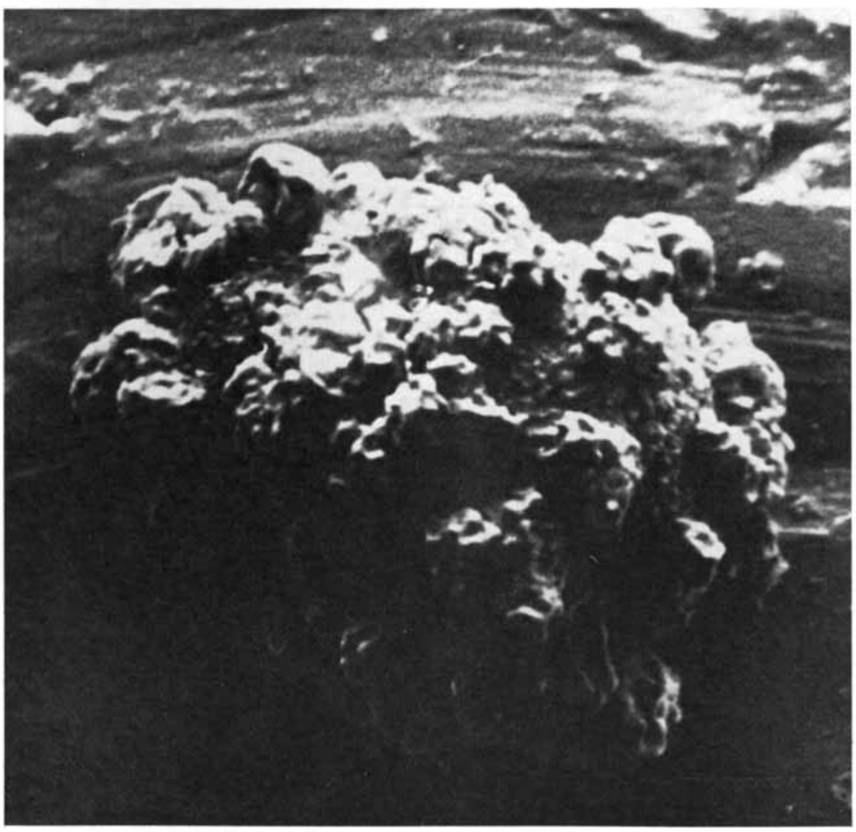
Another site in the body that is susceptible to amanitin poisoning is the kidney. The effects of the toxin in the kidneys cause direct damage to those organs, and they also lead indirectly to additional liver damage.

In man and apparently in some other mammals alpha-amanitin is filtered out of the blood in the glomeruli of the kidneys. Each of these dense balls of capillaries discharges the toxin, and many other substances, into the long duct called the convoluted tubule. Many substances are ordinarily reabsorbed from the convoluted tubules, but toxins such as amanitin should not be. The toxin attacks the tubule itself, however, and thereby reenters the blood and eventually returns to the liver [*see top illustration on page* 99].

An exception to this pattern is found in the rat. The convoluted tubules of the rat kidney are not destroyed by amanitin, and the toxin is therefore excreted efficiently. This capability probably accounts for the rat's relatively high tolerance for amatoxins, since at moderate dosages the liver should be able to replace injured cells as the toxin is elimi-



NORMAL LIVER CELLS in isolation are regular in form and are surrounded by a plasma membrane with a relatively smooth surface. This scanning electron micrograph and the one below were made by E. Weiss of the Institute for Veterinary Pathology in Giessen.



LIVER CELL TREATED WITH PHALLOIDIN displays gross distortions. The pressure of the cytoplasm creates protrusions where the plasma membrane has been weakened. When the cells are treated in the liver, where the pressure is greater outside the cell than inside it, invaginations develop instead of protrusions. In both cases the changes in the strength of the membrane are accompanied by an increase in its permeability to potassium ions.

nated. In other animals Fiume and his co-workers have been able to prevent kidney injury by binding the relatively small molecule of alpha-amanitin to the rather large molecule of the protein serum albumin. The complex is not filtered out of the bloodstream by the glomeruli, and the amanitin therefore does not reach the convoluted tubules. The kidneys are thus spared, but the constantly recirculating toxin causes greatly increased damage to the liver.

The destruction of liver and kidney tissue by the amatoxins can explain the delayed illness that is usually the fatal episode in amanita poisoning. There remains to be explained the gastrointestinal distress that is usually the first symptom. A recent series of experiments performed by Fiume and his colleagues suggests that this effect too can be attributed to the amatoxins. Earlier investigations employing both isolated toxins and a whole extract of *A. phalloides* had failed to induce the gastrointestinal symptoms in mice, rats and guinea pigs. Fiume repeated the experiment with dogs as experimental subjects and found that the characteristic diarrhea and vomiting began from nine to 11 hours after an injection of alpha-amanitin. A histological examination revealed lesions in cells of the stomach and several regions of the intestines.

A thorough understanding of the mechanism of amanita poisoning may eventually lead to the development of a specific antidote. The possible antidotes that have been proposed so far have been discovered by a more empirical method.

One of the potential antidotes is cytochrome *c,* an enzyme involved in the transport of electrons in cellular respiration. George L. Floersheim of the University of Basel has reported a survival rate of about 30 percent for mice given large doses of cytochrome *c* eight hours after injection with alpha-amanitin. Curiously the enzyme was effective only in female mice; male mice and both male and female rats showed no response to it. In addition it has proved successful only against isolated amanitin and has not been able to counteract the effects of an extract of *A. phalloides.*

Another proposed antidote is thioctic acid, a small organic molecule consisting of eight carbon atoms, terminating in a carboxyl group and with a disulfide group linking two of the carbon atoms [*see bottom illustration on page 94*]. Several structural isomers are possible; the one in which the sulfur atoms form part of a five-atom ring is alpha-lipoic acid, which participates in the citric acid cycle of cellular respiration.

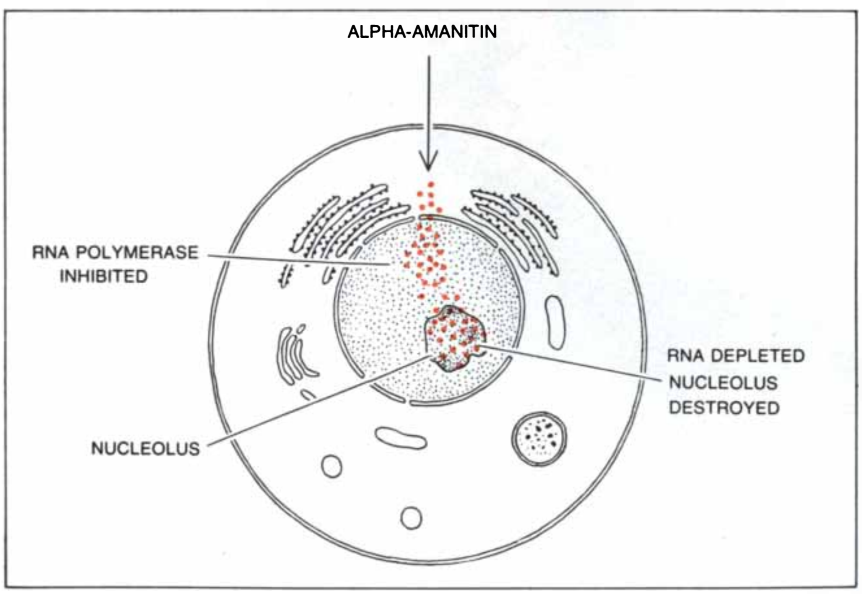
In the 1950's thioctic acid had been proposed as a remedy for liver damage caused by heavy-metal poisoning; the suggestion to try it for *A. phalloides* poisoning was first made by Josef Herlinka of Czechoslovakia in 1958. Ten years later Jiří Kubička, a physician at Třebon in Czechoslovakia, reported that after observing almost 100 fatalities attributed to amanitas during a period of 20 years, he was now able to save almost all such patients with large doses of thioctic acid. Kubička suggested that the extent of liver damage could be estimated by monitoring the level of the enzyme transaminase in the blood serum and that this indicator could serve as a guide to the administration of the drug. Thioctic acid and transaminase are involved in the same metabolic pathway: the oxidation of pyruvic acid.

In addition to Kubička's report successful laboratory tests of thioctic acid were announced in the German medical literature and clinical successes were reported in Italy. On the other hand, Floersheim has since evaluated thioctic acid as an antidote to alpha-amanitin and found it ineffective.

In 1970 thioctic acid was given an unplanned clinical trial in the U.S. On Sunday, November 15, two families from Goshen, N.J., shared a stew containing wild mushrooms they had gathered two days earlier in a pine woods of Cape May County, in the southern part of the state. About nine hours after the meal Clayton Brown awoke with severe gastrointestinal symptoms; three hours later Gary Bartee was sick also, as were Mrs. Brown and two of the Browns' daughters, aged 13 and eight. Another daughter, aged six, who had eaten less than a teaspoonful of the stew, had only mild cramps 24 hours after the meal, and a five-year-old, who had eaten even less, had no symptoms.

Bartee and the Browns and their two elder daughters were admitted to a Cape May County hospital. The attending physician had little doubt that they had consumed *A. verna,* the white amanita of which the American mushroom manuals warn. The state poison-control centers of New Jersey and New York were not able to offer him any information on antidotes.

Bartee's case presented the classic pattern of amanita poisoning. His dehydration was successfully corrected, and by Thursday, four days after he had eaten the mushrooms, he was well enough to eat again. The next day, however, he became stuporous and reported hallucinations. While he was being moved to another hospital he died. The Browns' 13-year-old daughter also died. Her sister experienced hallucinations for 24 hours but recovered, as did Mrs. Brown.



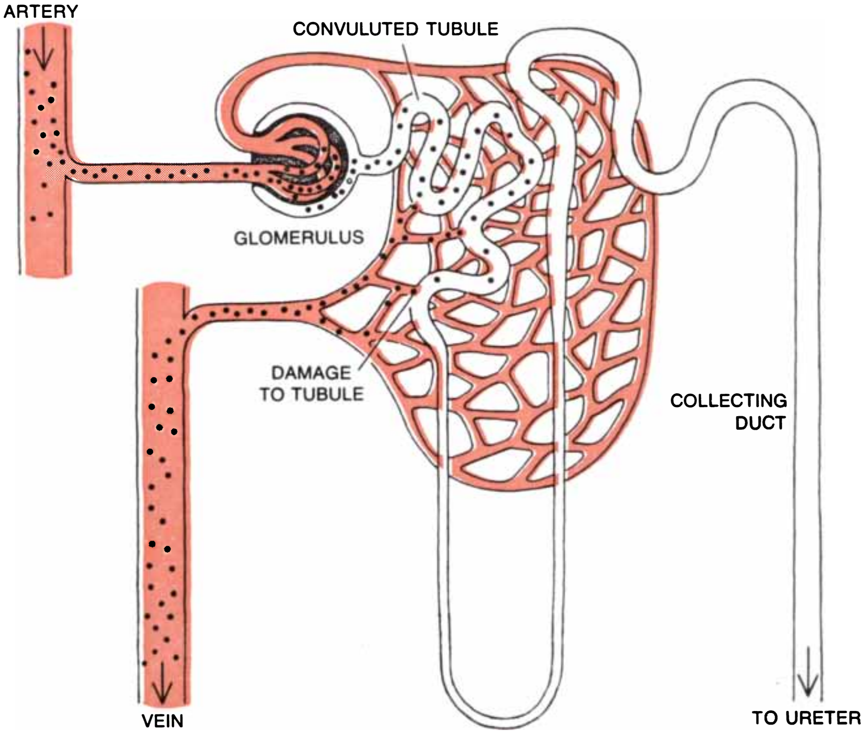
EFFECTS OF AMANITIN on the liver cell are concentrated in the nucleus. Minutes after a cell is exposed to the toxin the nucleolus begins to disintegrate. The nucleolus (where the ribosomes employed in protein synthesis are made) contains most of the RNA in the nucleus, and within an hour after treatment with amanitin most of the nuclear RNA is depleted. Amanitin is also known to inhibit an RNA polymerase, an enzyme that directs the synthesis of messenger RNA. Messenger RNA differs from the RNA of the nucleolus, but the effect of stopping its production is the same: the eventual cessation of protein synthesis.

Clayton Brown, whose liver function was rapidly deteriorating, was transferred to a Philadelphia hospital. The physicians treating him there were alerted to the reported European successes with thioctic acid by Donald M. Simons, a chemist with E. I. du Pont de Nemours & Company whose avocation is the study of fungi. Thioctic acid was not then certified as a pharmaceutical in the U.S., so that treatment was begun with a quantity procured from a chemical supply company and continued with material flown from Italy. Brown was in and out of a coma when the therapy began; six days later, when it was discontinued, he was well enough to tolerate a low-protein diet. He was discharged on December 18 as perhaps the first person in the U.S. to have been cured of amanita poisoning by thioctic acid. A biopsy showed necrosis of the liver with repair. Since then a Fresno, Calif., man with severe liver damage from amanita poisoning has recovered after treatment with thioctic acid. The Food and Drug Administration has approved the drug for use on an experimental basis in emergencies.

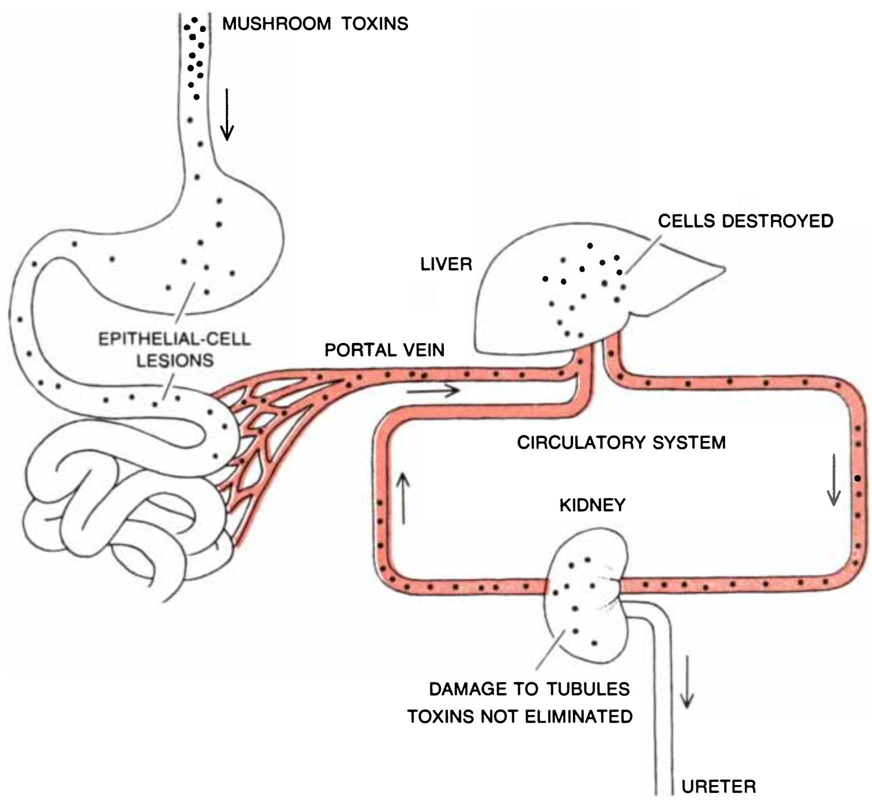
The identity of the mushrooms that poisoned Bartee and the Browns surprised many American mycologists. While Brown was still receiving thioctic acid treatment Simons examined a specimen from the Cape May County stand. It had begun to decompose, and in that condition it did not contradict the original diagnosis of *A. verna.* When Simons returned to the same stand of white pine a year later, however, he found more than 100 mushrooms that were a dark olive green to greenish yellow; there was considerable variation in color, but none of the fruit bodies were the pure white that is characteristic of *A. verna.*

Three weeks later another fruiting of what appeared to be the same species was encountered near a picnic shelter under oak trees in Durand-Eastman Park in Rochester, N.Y. They were collected by the Reverend James Wolf, a student in a beginner's class in mushroom identification. The instructor, Leo J. Tanghe, a chemist retired from the Eastman Kodak Company, had hunted mushrooms in the same place for more than 30 years without encountering the species. During the next month more than 100 fruit bodies were observed in the area.

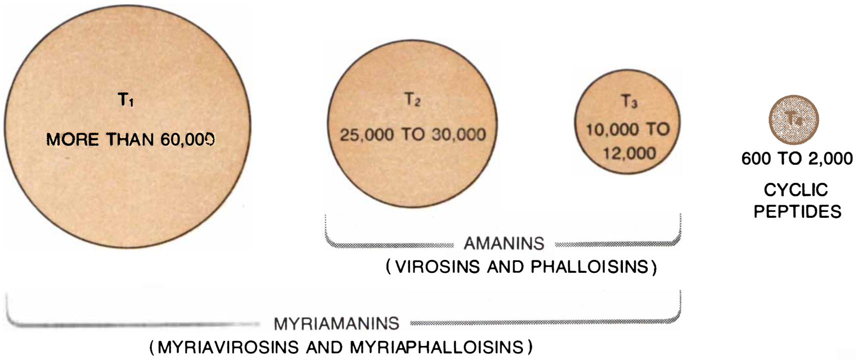
Simons visited the Rochester site and concluded that the mushrooms were identical with the ones he had seen in New Jersey. They also matched specimens he had collected a year earlier near



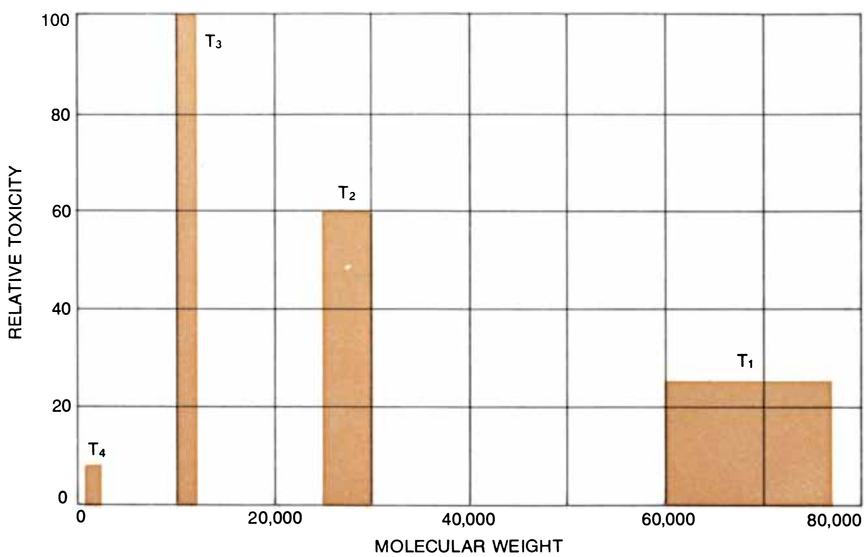
KIDNEY is susceptible to attack by amanitin in most mammals. The toxin is filtered out of the blood by the capillaries of the glomeruli. Before it can be eliminated in the urine, however, it damages the walls of the convoluted tubules and is reabsorbed by another network of capillaries. As a result the kidney itself is injured, and the toxin returns to the bloodstream to cause additional liver damage. The kidneys of rats are resistant to the toxin.



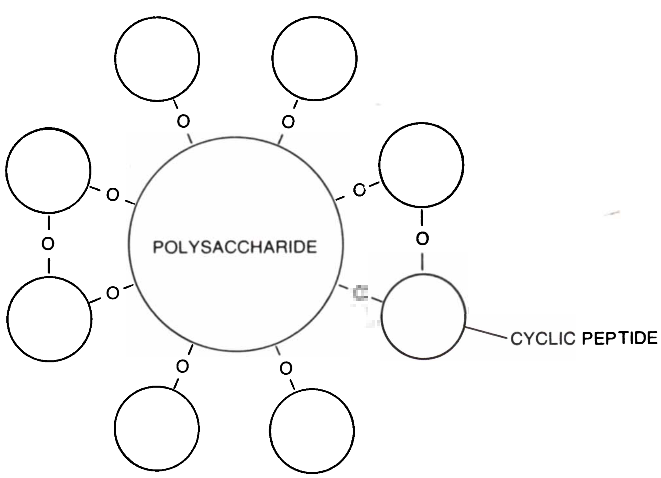
AMATOXIN POISONING destroys tissue in the gastrointestinal tract, the liver and the kidneys. The initial symptoms of poisoning develop when the toxins produce lesions in cells lining several regions of the stomach and the intestines. These symptoms can usually be alleviated; fatalities most often result from liver damage exacerbated by kidney dysfunction.



LARGE TOXIN MOLECULES called myriamanins were extracted from amanitas by Michel Courtillot and Thadée Staron of the National Institute for Agronomic Research in Versailles. The largest molecule, designated T_1, is thought to be a polymer composed of multiple T_2 or T_3 units. T_2 and T_3 when derived from *A. virosa* are called virosins; when from *A. phalloides*, phalloisins. The T_4 fraction is made up of cyclopeptides of about the same molecular weight as the phallotoxins and amatoxins characterized by other investigators.



RELATIVE TOXICITY of substances extracted by Courtillot and Staron varies by a factor of 10. The cyclopeptides (T_4), which are least toxic, may be fragments of a larger molecule.



TENTATIVE STRUCTURE of the T_3 molecule was proposed by Courtillot and Staron. Cyclopeptides are attached through oxygen atoms to a polysaccharide that makes up from 25 to 30 percent of the mass of the molecule. Some peptides may be bonded to one another.

a reservoir in Newcastle County, Del. Another collection made near Williamsburg, Va., also appeared to be the same species.

In attempting to determine the identity of this species emphasis was placed not on the fleshy parts of the fungus but on the spores. For comparison with the American collections the late Albert Pilat of the National Museum in Prague supplied Tanghe with five herbarium specimens collected during a period of 30 years in several regions of Czechoslovakia. Spores from both the European and the American mushrooms had the elliptical shape ascribed to *A. phalloides* and *A. verna,* rather than the globose form of *A. virosa* and *A. bisporigera* spores. Tanghe and I were surprised to observe, however, that the ellipticity of the spores varied markedly among the five collections from Czechoslovakia. One had spores that were virtually globose.

Seeking an explanation for the variation, Tanghe collected two more specimens of the greenish mushroom from the Rochester park. One was dug up intact, with a large clod of moist soil, and was kept covered to maintain humidity. Spores from this specimen and from a severed cap were collected several times a day. Tanghe found that the ellipticity of the spores changed as the tissues they dropped from aged; the ratio of length to width briefly increased to a maximum, then slowly decreased until the mushroom had exhausted its supply of spores. Having found that spore shape is variable, and since the pigmentation of the mushroom is also a mutable characteristic, the distinction between *A. phalloides* and *A. virosa* becomes somewhat vague. Tanghe and Simons nevertheless concluded that the specimens they had collected, and the species responsible for the New Jersey poisoning, were probably *A. phalloides.*

Just as the distinctions between the *Amanita* species have been obscured, the identity and nature of the toxins have also recently been questioned. Michel Courtillot and Thadée Staron of the French National Institute for Agronomic Research in Versailles have suggested that the cyclopeptides isolated by Wieland may be mere fragments of a more complex toxin, or artifacts of the methods used to extract the substances. They believe that by employing gentle methods of extraction and separation they have preserved fragile molecular assemblages that would ordinarily be destroyed. They have also found that, in spite of morphological similarities between *A. phalloides* and *A. virosa,* the two species contain distinct families of toxins.

The toxins extracted by Courtillot and Staron include substances of much higher molecular weight than those isolated by Wieland and his colleagues. These large molecules they have named myriamanins; those from *A. phalloides* are called myriaphalloisins and those from *A. virosa* are called myriavirosins.

Mice given the whole toxic extract die in from a few hours to four days. The extract of *A. virosa* has a strong odor, and even mice that have been deprived of food and water will not willingly eat it. Paradoxically the substance responsible for the odor is an antagonist to the toxins; when the antagonist has been removed from the preparation, and its toxicity has thus been increased, the mice are no longer repelled by it. Barbara Courtillot-Wielezynska of Staron's laboratory has found that the toxicity of the extract is also reduced by incubating it with a homogenate made from horse stomach muscle or mouse stomach.

The myriamanins can be separated into four fractions by filtration through a series of permeable membranes. A fraction with a molecular weight of more than 60,000, designated T_1, is precipitated by chloroform; it is probably a polymer of toxins present in solution as smaller molecules. Another fraction, T_2, has a molecular weight of 25,000 or 30,000. Still another, T_4, consists of relatively small molecules, with molecular weights between 600 and 2,000. It is this fraction that Courtillot and Staron have identified with the cyclopeptides. Significantly they have found that it is present only in low concentration except under strongly acidic or alkaline conditions; these findings suggest that the cyclopeptides may be degradation products of a larger molecule.

The most toxic fraction of the extract, which Courtillot and Staron designate T_3, has a molecular weight of from 10,000 to 12,000. They suggest that it consists of a polysaccharide (a molecular chain made up of sugar units) to which several cyclopeptides are connected by bonds through oxygen atoms. Other oxygen atoms or pairs of sulfur atoms are believed to link some of the cyclopeptides together [*see illustrations on opposite page*]. Courtillot and Staron speculate that it may be possible to develop an immunological defense to amanita poisoning; it would depend on an immunological attack on the polysaccharide portion of the T_3 fraction.

In general, preparations of T_2, T_3 and T_4 from *A. phalloides* are less toxic than those from *A. virosa.* They can be distinguished by the techniques of chromatography and electrophoresis, and the symptoms they elicit when they are injected into mice also differ. In mice thioctic acid has proved ineffective against *A. virosa* extracts.

Wieland and Heinz Faulstich have recently provided further evidence that the distinction between the green and the white deadly amanitas is real. They report that they have isolated another new toxin from fruit bodies of *A. virosa* collected in Italy and Sweden. It is of low molecular weight and undetermined structure; in potency and toxic action it resembles phalloidin.

Further investigation of the amanitas can be expected to define more precisely the distinctions between species and the nature of the toxins they contain. It appears that the habitat and the history of the individual fruit body may influence the abundance of the various toxins. Even apart from the problem of providing aid to those who have been poisoned, the amanitas lead toward inherently interesting areas of biology. On the one hand they take us to the cell membrane and inside the nucleus. On the other they call attention to a profound enigma in the theory of evolution: It is not at all apparent what selective advantage is conferred on the deadly amanitas by the poisonous substances they manufacture.

The traditional view that almost all European fatalities from mushroom poisoning are caused by *A. phalloides* has now been questioned by Wieland. He considers that the white amanitas are more likely to be mistaken for edible species than the green *A. phalloides* is. The white species are, of course, no less toxic.

At the same time *A. phalloides* must now be considered a suspect in cases of mushroom poisoning in the U.S. The mycelium in Durand-Eastman Park in Rochester has fruited each fall for three years, the species has been reported in other regions and it must now be regarded as being established. There remains the question of why *A. phalloides* suddenly appeared in the U.S. in 1970. It is possible that it has been growing unnoticed in North America for centuries. It seems more probable, however, that it immigrated sometime during this century as bits of hyphae on the rootlets of nursery stock from Europe and has taken time to bear its lethal fruit.

Galileo's Discovery of the Parabolic Trajectory

He showed that a body falling with a horizontal component of motion describes a parabola. It has been thought that he did so solely by principle, but it now appears that he conducted careful experiments

by Stillman Drake and James MacLachlan

Galileo, in the last part of his final work *Discourses on Two New Sciences* (1638), proved that the trajectory of a projectile traveling through a nonresisting medium is a parabola. The proof is simple and straightforward, once it is known that the vertical distance the object falls from rest is proportional to the square of the time elapsed and that the projectile's horizontal velocity will remain uniform. As a matter of fact, a similar proof had been published six years earlier in a book on conic sections by Bonaventura Cavalieri, a pioneer of the calculus who knew Galileo and who had studied mathematics under one of Galileo's pupils, Benedetto Castelli. Galileo was indignant when he first learned of Cavalieri's publication; he wrote to a friend that the proof was the fruit of studies he had begun 40 years earlier and that the least he deserved was the courtesy of first publication. No one knew better than he, he said, how hard it had been to make the discovery and yet how easy it was to work out the proof when the shape of the trajectory was known.

Cavalieri was much distressed when he learned of Galileo's wrath, and he wrote at once to say that he had credited both Galileo and Castelli in his book, a copy of which he sent to Galileo. He added that everyone knew the discovery was Galileo's and that he himself had believed Galileo had already published it. Galileo was satisfied and went out of his way in *Discourses on Two New Sciences* to praise Cavalieri as a new Archimedes.

Some historians of science have naturally wondered how much of this sequence of events could be believed. Was it really possible that a man who had made such a discovery would never have mentioned it in print over a span of three decades? On the other hand, would Cavalieri have said that he had believed the result to have been already published by Galileo if the discovery, and not just the proof, was really his own?

These questions can now be answered on the basis of Galileo's notes on motion, preserved at the National Central Library in Florence in Volume 72 of the Galilean manuscripts. It is certain that Galileo discovered the parabolic trajectory no later than 1608 and proved it mathematically early in 1609, although he did not mention it in print until 30 years later. The papers that support this conclusion were not published with Galileo's collected works; they are reproduced here for the first time.

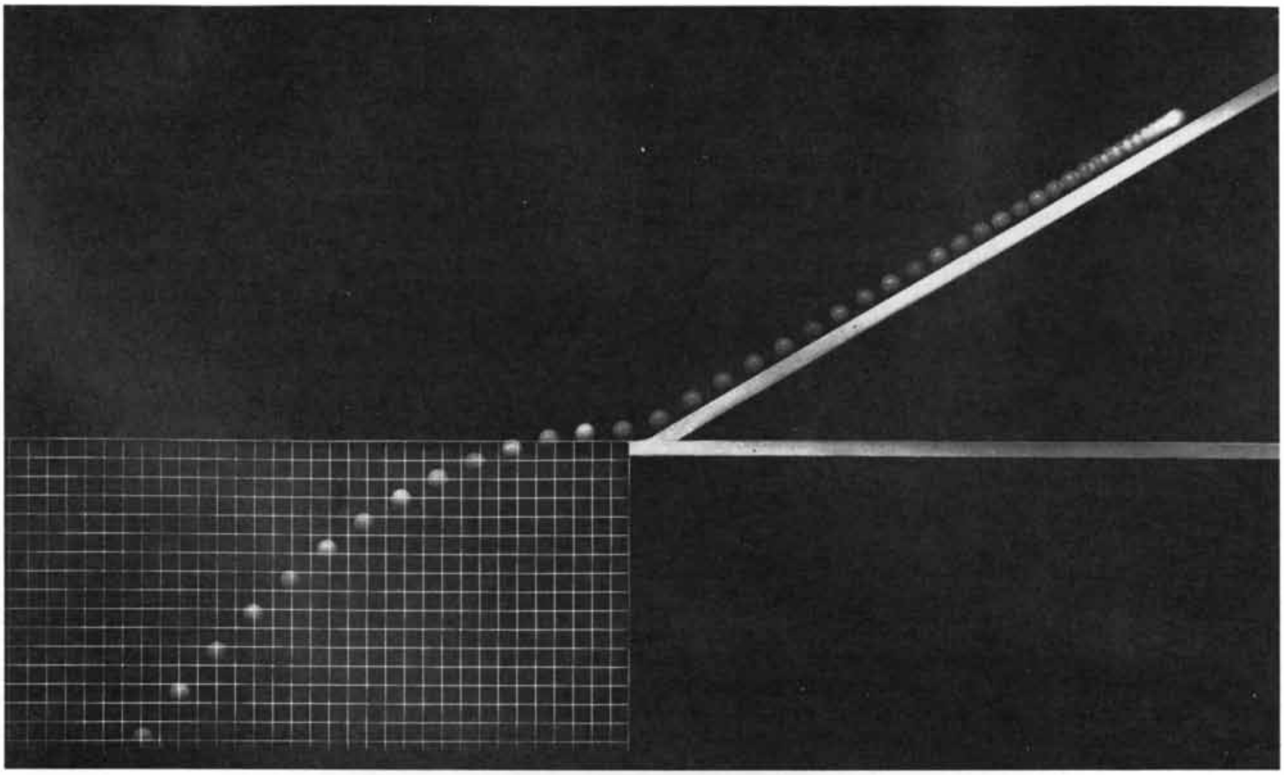
As we have mentioned, two laws must be known in order to derive the parabolic trajectory. One is the law of free fall, which states that the distance an object falls from rest is proportional to the square of the time elapsed, and which Galileo had discovered in 1604 by a combination of luck and mathematical reasoning [see "Galileo's Discovery of the Law of Free Fall," by Stillman Drake; SCIENTIFIC AMERICAN, May, 1973]. The other law needed was a restricted principle of inertia that gives a relation between the motion imparted to a body and the behavior of the body after it is free from the initial impulse. Such a principle was precisely what Galileo was testing when he discovered the parabolic trajectory. Luck played no part this time. Galileo's discovery of the parabolic trajectory was a case of serendipity: the discovery of something other than what the seeker had set out to find.

To a historian of science the interesting thing about Galileo's discovery is the remarkable extent to which he made use of experimental methods in science that we now take for granted but that were not standard procedure in the 17th century. It is difficult to tell from Galileo's published works how much use he really made of careful experiments. Those that he did describe as having been actually carried out have been repeated in recent years, and they work very well. Galileo did not, however, describe many such experiments, and he did not give his results in numerical form, as became the custom among his successors. It is only in his private notes that traces of his experimental data survive.

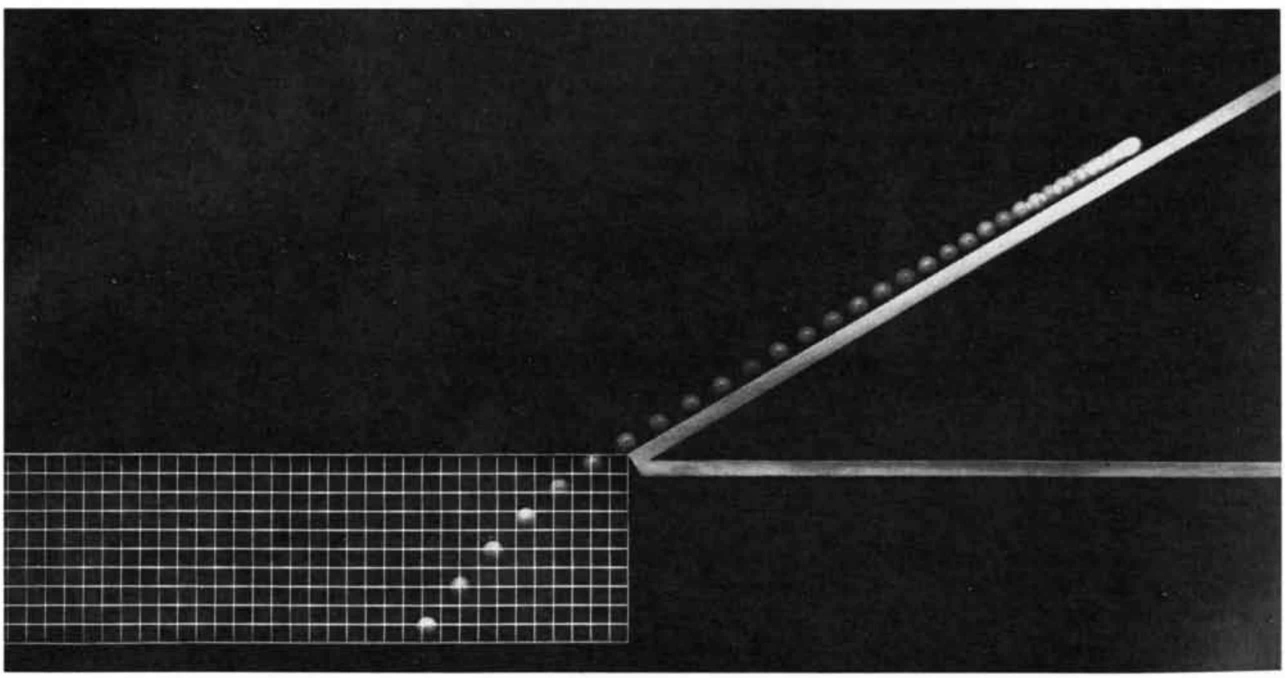
Naturally we must be cautious about interpreting those data. In the case of Galileo's discovery of the parabolic trajectory, however, one set of seven three-digit numbers (253, 337, 395, 451, 495, 534 and 573) exists that he could not have obtained in any way other than by careful measurement. We shall reconstruct the steps that led to those numbers and then describe an experiment we have conducted that nearly duplicates them. Although we may be wrong about details, the evidence in favor of our overall picture seems to be convincing.

As early as 1590 Galileo had proved mathematically that in theory a ball on a level plane will be set in motion by any force, no matter how small. By 1607 he was teaching his student Castelli that once any motion has begun it needs no force for its continuation. It was accordingly reasonable for him to believe that on a level plane a ball would continue rolling indefinitely with a uniform motion if it was not impeded.

During the years 1602 through 1608 Galileo developed many theorems con-



PARABOLIC TRAJECTORY of a billiard ball deflected horizontally into the air after rolling in a groove down an inclined plane is clearly shown in this multiple-flash photograph made by Ben Rose. The plane and table, built by SCIENTIFIC AMERICAN, duplicate the apparatus used by the authors to reconstruct the experiment conducted by Galileo before 1608. The experiment is recorded in the document designated *f. 116* in Volume 72 of the Galilean manuscripts in the National Central Library in Florence. The ball was released from a height of 828 *punti* (77.7 centimeters) above the horizontal plank and dropped 828 *punti* after being deflected. In this photograph and the one below grid lines in background are 50 *punti* apart; time between images of ball is 30.5 milliseconds.



SIMILAR EXPERIMENT without the curved deflector duplicates the experiment recorded by Galileo on the page of his notes designated *f. 114* (*see bottom illustration on page 108*). Here the ball was released from a height of 800 *punti* and the final drop was only 500 *punti*. Galileo began these experiments in order to establish a restricted principle of inertia that gives a relation between the motion imparted to a body and the behavior of the body after it is free from the initial impulse. In the course of this investigation he noted the shape of the ball's trajectory. The restricted principle of inertia was one of the two laws he needed in order to prove that the trajectory was indeed a parabola; the other law was the law of free fall, which he had discovered four years previously.

cerning motion in vertical free fall, along inclined planes and in combinations of the two. It was during this period that he experimented with an actual level plane. In the document labeled *f. 117* in Volume 72 of the Galilean manuscripts a series of diminishing numbers is seen along a central horizontal line [*see illustration on page 107*]. The numbers probably represent the distances traveled by a ball in successive equal times after it had been given an initial push along a groove on a level plane. Obviously it would have been unsafe for Galileo to conjecture from such data alone that the ball was slowing down entirely because of friction and other external influences. It might have been that the initial impetus simply "got tired" and would have done so even on an ideally flat and frictionless plane. Hence it would have been natural for Galileo to wonder what would happen in the absence of contact between the ball and the plane.

Galileo's own law of free fall and his knowledge of how it is modified on inclined planes had put into his hands a means of answering that question. We know that a body heavy enough not to be much influenced by the friction of the air takes the same time to fall a given distance whether or not its motion is straight down. Its horizontal progress during the time it falls should therefore depend only on its horizontal speed at the moment it begins to fall. Then if one could control the ratios of different horizontal speeds, it would be possible to test how uniformly the horizontal motion is continued in the absence of friction. The ratios of the terminal speeds at the end of motions from different heights along a fixed inclined plane could be calculated by using any single actual measurement and the law of free fall. Such a set of calculations is found together with Galileo's measurements in the document labeled *f. 116*.

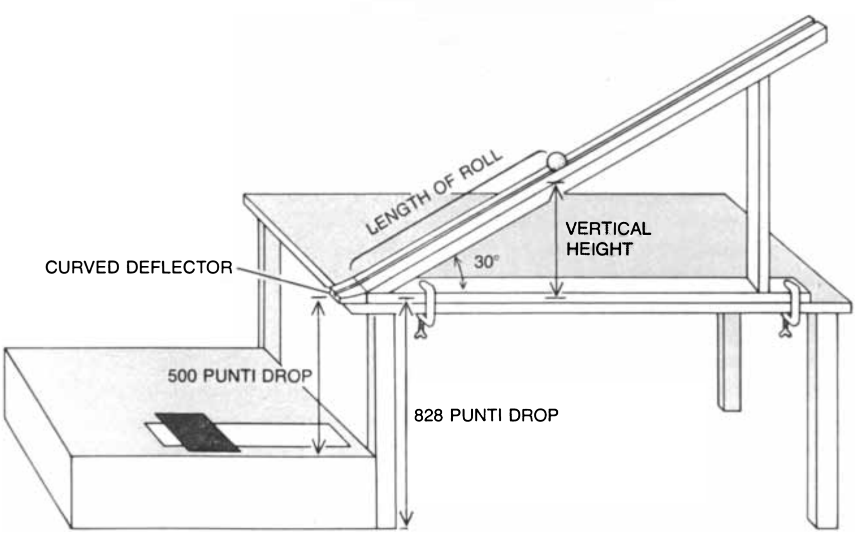
Galileo rolled a ball from different heights down a grooved plane supported on a table and recorded the heights above the table with the numbers 300, 600, 800 and 1,000. He deflected the ball horizontally at a height above the floor recorded as 828. For each of the four original heights he marked the point at which the ball hit the floor. The distances from the edge of the table were 800, 1,172, 1,328 and 1,500. For a reason that we shall explain below he also rolled the ball from a height of 828 above the table, and he recorded the horizontal distance traversed as 1,340.

After obtaining these measurements he used his law of free fall to relate the height of the plane to the horizontal distance traversed. The first pair of measurements, 300 units of height to 800 units of horizontal distance traversed, he took as a standard ratio. On the basis of the standard ratio and his times-squared law of free fall he calculated the horizontal distances the ball should have traversed. The calculated distances for the heights of 300, 600, 800 and 1,000 came out as 800, 1,131, 1,306 and 1,460. He also calculated a horizontal advance of 1,329 or 1,330 for the added case when the height of the ball on the plane above the table (828 units) was equal to the drop from the table to the floor (828 units). The fact that the calculated horizontal distances so nearly agreed with the measured distances established the restricted principle of inertia that Galileo had been seeking and that he needed. Such a principle can be stated concisely: In the absence of any appreciable resistance the horizontal motion is uniform and of indefinite duration.

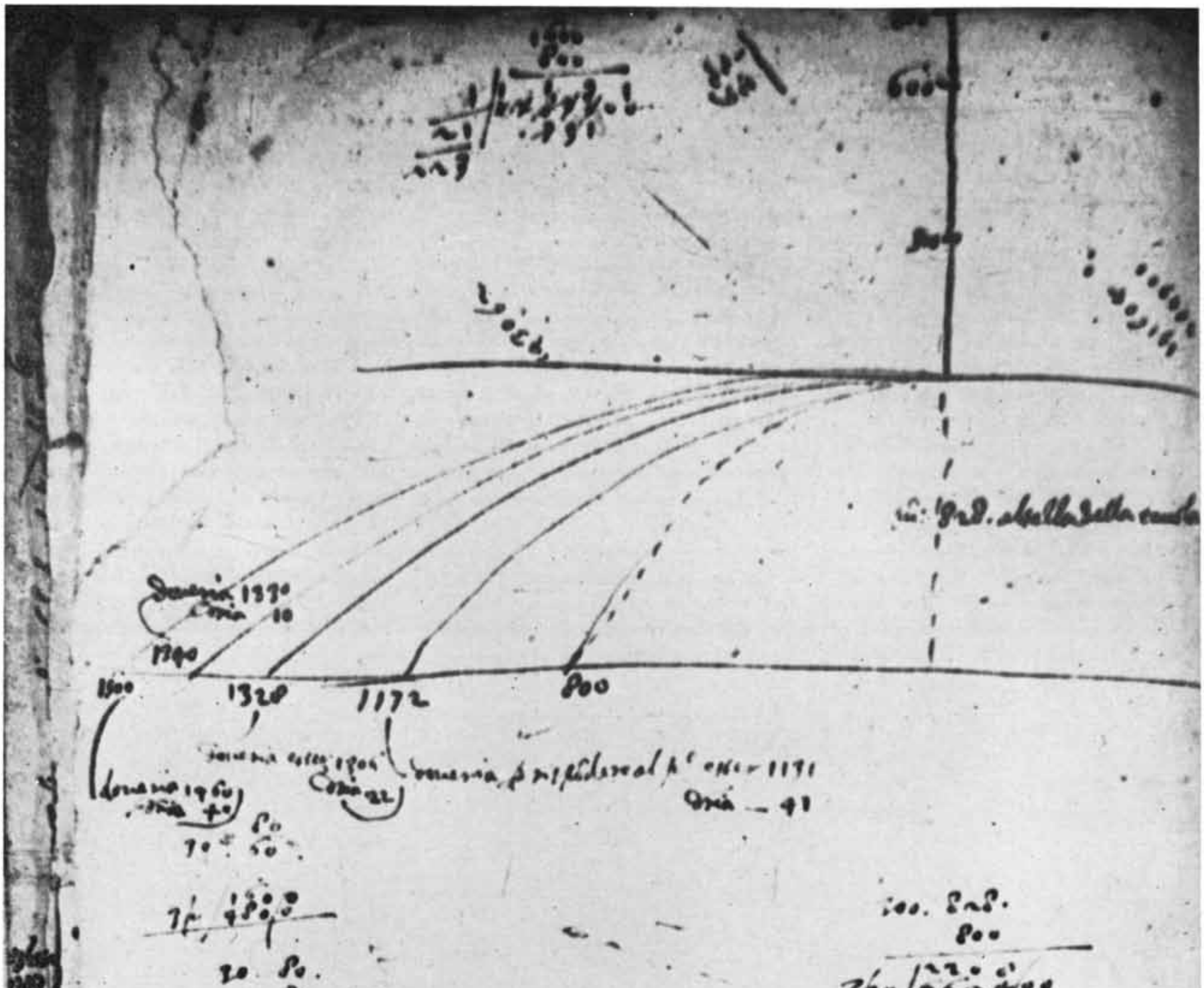
In a third document, *f. 175*, which is clearly associated with these experiments and which bears the same watermark as a letter Galileo wrote dated late in 1608, he drew a steep inclined plane equipped with a curved deflector designed to convert all the ball's motion into horizontal motion before the drop from the table to the floor. In practice he could not have used such a steep plane because it is very difficult to make the ball roll along it uniformly or even to make it always travel in the same straight line. In fact, our analysis of Galileo's data shows that his plane could not have been steeper than 64 degrees because the ball's acceleration implied by the data is less than the acceleration a ball would attain even if it rolled perfectly on such a steep plane. Moreover, it is quite difficult to get a perfect roll even on slopes much shallower than 64 degrees.

One might reasonably ask how we can determine the acceleration of the ball in Galileo's experiment, since we have said nothing so far about his units of distance. It happens that in other notes Galileo gave the length of certain lines in what he called *punti*, or points. These measures turn out to agree very nearly with the small units engraved on the linear scale of Galileo's proportional compass preserved at the Museum of the History of Science in Florence. On that instrument the unit is 169/180 millimeter, which we have used even though the drawings in some of his notes put one *punto* at 58/60 millimeter.

In *f. 116* Galileo wrote *pū. 828 altezza della tavola*, or "828 points height of the table." On the basis of this remark we have taken the height of his table as being 777 millimeters in order to attain the times, terminal velocities and rates of acceleration stated in his measurements. Thus the table of Galileo's ex-



DRAWING OF APPARATUS used by the authors shows the equipment they constructed to duplicate Galileo's experiments. The inclined plane was a plank of wood measuring two inches by four inches by six feet raised to an angle of 30 degrees from the horizontal. A steel ball was rolled down a square groove from specific vertical heights above the table. When it reached the end of the incline, it fell onto a sheet of carbon paper resting on top of a paper strip. The impact of the ball left a spot on the paper whose distance could be measured to the edge of the table. Measurements were then compared with Galileo's. Curved deflector was used in duplication of the experiment in *f. 116*, and the ball was allowed to drop 828 *punti* to the floor. The deflector was removed in the duplication of the experiment in *f. 114* and a box was placed below the inclined plane to decrease drop from 828 *punti* to 500 *punti*.



UNPUBLISHED MANUSCRIPT records one of the experiments conducted by Galileo by which he discovered the law of the parabolic trajectory in 1608. It is reproduced by courtesy of the National Central Library in Florence, where it is preserved as *f. 116* of the Galilean manuscripts. The diagram at the top shows the data Galileo obtained by measuring the actual distance the ball moved from the edge of the table after being rolled down an inclined plane from specified vertical heights. The numbers at the bottom are his calculations of expected distance that ball should have traveled from different heights (*see illustration on next page*).

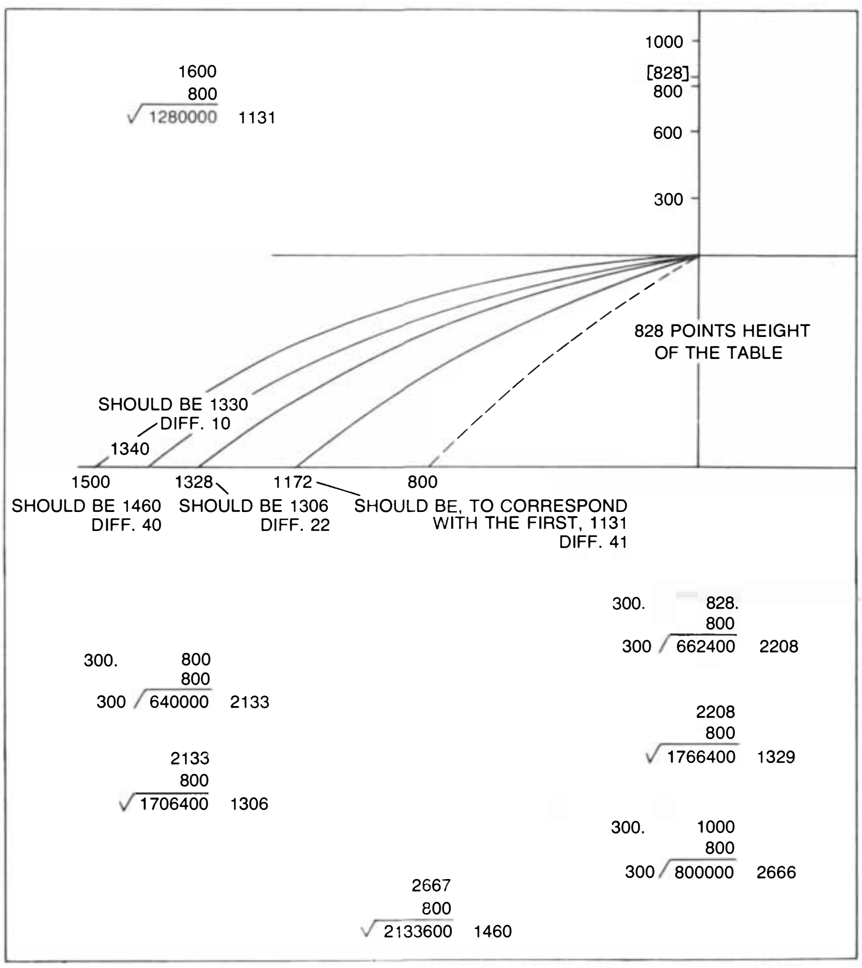
periments was about the standard height of modern tables, and the highest point from which he released the ball was about five feet eight inches above the floor. That is reasonable for a man of average height who wanted to be sure he had the ball exactly on the mark before he released it. Moreover, the same arrangement would have allowed Galileo to stoop and observe the ball in flight. He sketched the paths in his diagram, and we can hardly doubt that their form suggested the parabolic trajectory to him. Galileo knew a great deal about parabolas, having begun his scientific career with a study of the centers of gravity of solids with parabolic shapes.

Once the idea of the parabolic trajectory had suggested itself to Galileo, it was easy for him to derive it theoretically. If we look back at *f. 117,* we see that he began at once to design a parabolic trajectory and to sketch little parabolas with steps under them, each of equal horizontal length, corresponding to the height of the curve above the ground for each equal unit of horizontal distance traversed. On the other side of the page there are some related calculations. A later group of pages, written early in 1609, contains the basic theorems of the parabolic trajectory that he did not publish until 30 years later.

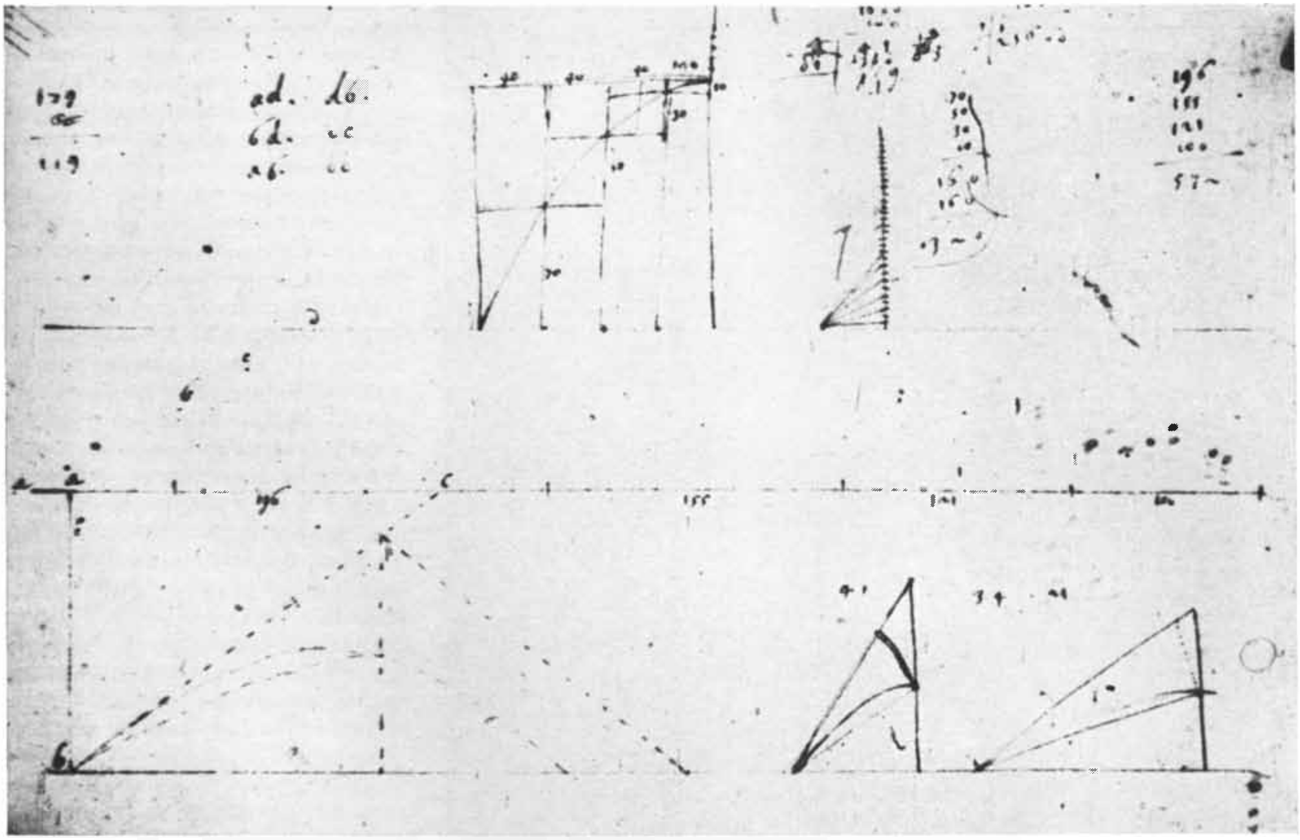
When we first analyzed Galileo's data in 1972, one of us (Drake) believed the inclined plane Galileo used for the experiment recorded in *f. 116* was probably tilted at an angle of 64 degrees to the table, the steepest angle that could be accounted for. Now, however, we believe that, for that experiment as well as for the continuation of the work we shall discuss below, Galileo employed a plane at an angle of only 30 degrees to the table. A plane at that angle is easy to set up with considerable precision, and it also lends itself to easy computation. All the experiments were probably made with a bronze ball covered with ink running in a groove and then quickly deflected horizontally; when the ball struck the floor, which was presumably marble, it would leave a well-defined ink spot. From such a spot it would be possible to measure the distance to a point directly under the edge of the table to an accuracy of within about a millimeter. Galileo's drawing of the ball's path in *f. 175* shows he was aware that the ball would rise after it was deflected, so that it would be above the level of the table for a short time before it again reached that level. How he took this rise into account experimentally is not recorded.

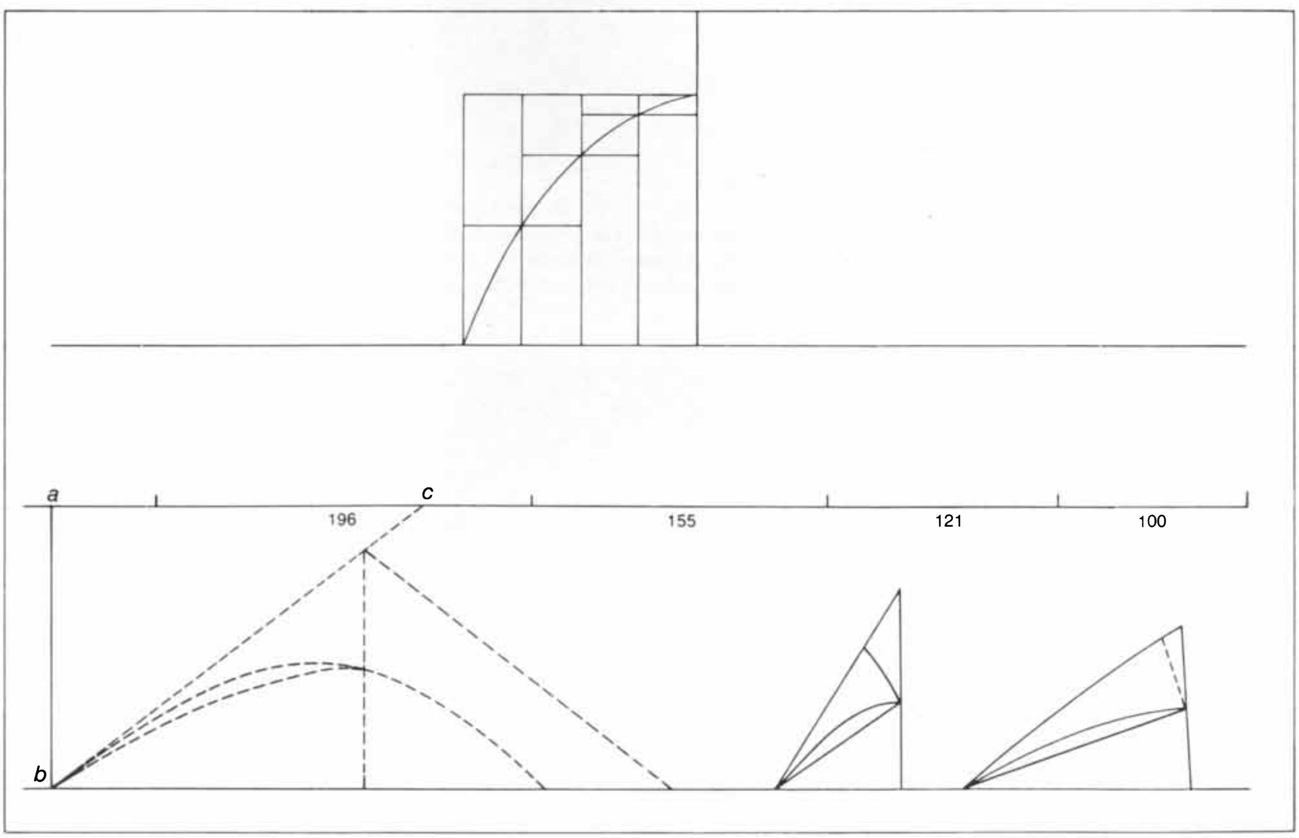
We have mentioned that Galileo included a drop of 828 *punti* to the table. His reason for adding this test is easy to guess. Galileo had proved mathematically, probably in 1604 or 1605, that a body deflected horizontally after it has been accelerated from rest for a certain time interval will, in another such time interval, traverse a distance double the distance it moved under the initial uniform acceleration. He had proved this double-distance rule for inclined planes as well as for vertical descent. Naturally Galileo wanted to test this result as well as the uniformity of horizontal motion. He supposed he could do so by having the ball start initially from exactly the same vertical height above the table as it would subsequently drop from the table to the floor. Such an arrangement, he thought, should make it easy for him to confirm his double-distance rule.

The result, however, must have surprised and disturbed him. First he computed the expected distance for this case exactly as he had computed the others, using the standard ratio and the times-squared law, and the result fitted with the others very well, giving a horizontal distance of 1,330 *punti* traversed as against the 1,340 *punti* actually measured. The double-distance rule still had to stand on its own feet, so to speak. According to Galileo's theory of inclined planes, the times consumed by a ball descending vertically and along any plane of the same height were proportional to the ratio of the height of the plane to the length of the plane. Now, that would be true if there were free, frictionless fall in both descents, as if a heavy block of ice were sliding on a hot track. The phenomenon of rolling, however, is a different matter. For Galileo's bronze ball two-sevenths of the energy from the attraction of gravity went into rotating the mass, leaving only five-sevenths to go into the motion of translation down the plane. Hence the ratio of the time it takes a ball to roll down a plane 828 *punti*

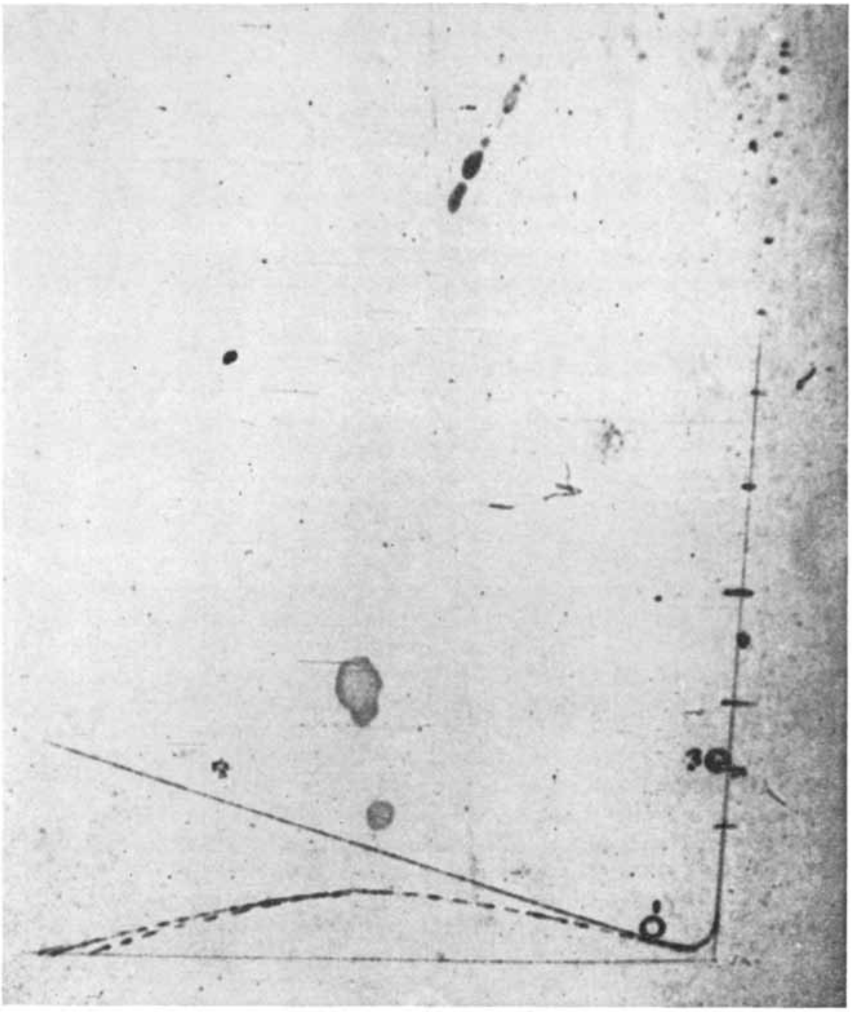


ENGLISH TRANSCRIPTION of *f. 116* shows Galileo's measurements of the distances the ball traveled horizontally and his calculations of distances he expected using law of free fall and restricted principle of inertia. Distances he expected were based on length of shortest roll and horizontal distance he measured for that roll. Galileo did not include the number 828 on the line above the table although he used that drop to test a special rule.

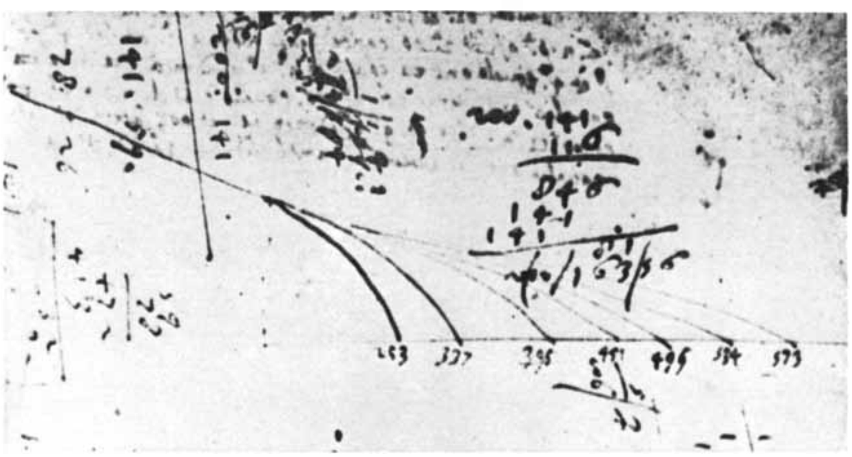




PHOTOGRAPH AND TRANSCRIPTION of *f. 117*, also in Volume 72 of the Galilean manuscripts, show Galileo's attempts to establish the restricted principle of inertia he needed. Such a principle, used in conjunction with his times-squared law of free fall, would enable him to prove that the trajectory of an object traveling through a nonresisting medium was indeed a parabola. The numbers on the horizontal line probably represent the diminishing distances traversed by a ball in successive equal times after it had been given an initial push along a level grooved plane. The little parabolas are preliminary sketches he made of likely trajectories. The steps under the larger parabola at the top represent the height of the ball above the ground at equal horizontal distances along its path.



BALL BOUNCED INTO THE AIR for a short distance when it rolled directly off the inclined plane onto a flat surface deflecting it horizontally into the air. Galileo was aware of this problem, as is shown by his sketch of the ball's path in the document designated *f. 175*. In order to reduce the bounce and make the transition smooth he devised curved deflector.



DOCUMENT *f. 114* shows Galileo's measurements of the distances the ball fell from the end of the inclined plane when it was not deflected horizontally. He had devised this experiment to try to resolve the apparent conflict of the data he recorded in *f. 116* for the horizontal distances the ball traveled after it fell through the vertical drop of 828 *punti* above the table. Galileo could not have computed the set of seven three-digit numbers shown (253, 337, 395, 451, 495, 534 and 573), and so he must have obtained them by careful measurements. Results from experiment duplicated by the authors are in close agreement with Galileo's.

high to the time it takes the ball to fall 828 *punti* is not the same as the ratio of the length of the plane to the height of the plane. Galileo did not know this, and so his separate calculations according to the double-distance rule would have led him to expect the ball to advance 1,656 *punti* horizontally (twice the vertical height of 828 *punti*), as against the actually observed advance of 1,340 *punti*.

That is a pretty serious discrepancy. Galileo could not have reasonably accounted for it except by supposing some of the ball's motion must be lost as a result of its sudden horizontal deflection at the end of the incline. That is probably why he devised a new experiment in which there would be no deflection and thus no artificial change in the direction of motion. The record of this experiment is found in the document *f. 114*, although it could not have been reconstructed without the information already gained from *f. 116*. In this experiment the ball was allowed simply to roll off the end of the inclined plane and drop to the floor. Data are given for the respective horizontal distances traversed after the ball rolled seven different lengths, but there are in this case no relevant calculations. On the face of it, it seems all we can say about this experiment is that the plane seems to have been at an angle of about 30 degrees and that that is the end of the matter.

Fortunately for us it is far from the end. We can further relate the experiment shown in *f. 114* to the experiment shown in *f. 116*. Our analysis indicates that Galileo's data are consistent with the horizontal distances traversed by a ball rolling six different times along a plane from heights having the ratios of 1 : 2 : 3 : 6 : 8 : 10, plus one ratio that is not expressible in whole numbers. These are precisely the same as the ratios of the vertical heights of 300, 600, 800 and 1,000 shown in *f. 116,* plus heights of 100, 200 and 450. Considering the complexity of the analysis, the agreement is not likely to be coincidence.

Next, it is highly probable that when Galileo elected the heights from which the ball was to be rolled in the experiment recorded in *f. 114,* he would have used round hundreds of *punti* for the sake of convenience, and that he would have done the same for the terminal drop. Our analysis shows that his recorded data are indeed consistent with the assumption that the lengths of the rolls along the plane, in *punti,* were 200, 400, 600, 900, 1,200, 1,600 and 2,000; the final vertical drop to the floor was 500 *punti.* The same calculations show

LENGTH OF ROLL ON PLANE (PUNTI)	VERTICAL HEIGHT (PUNTI)	TIME (SECONDS)	TERMINAL SPEED (PUNTI PER SECOND)	HORIZONTAL DISTANCE TRAVELED IN AIR (PUNTI) THEORY *f. 116*	GALILEO	PERCENT DIFFERENCE FROM THEORY	HORIZONTAL DISTANCE TRAVELED IN AIR (PUNTI) THEORY *f.114*	GALILEO	PERCENT DIFFERENCE FROM THEORY
200	100	.343	1,168				261	253	−3.2
400	200	.484	1,651				344	337	−1.9
600	300	.593	2,022	805	800	−.7	398	395	−.8
900	450	.727	2,477				456	451	−1.2
1,200	600	.839	2,860	1,139	1,172	+2.8	499	495	−.8
1,600	800	.969	3,303	1,315	1,328	+1.0	542	534	−1.4
1,656	828	.986	3,360	1,338	1,340	+.2			
2,000	1,000	1.083	3,692	1,470	1,500	+2.0	574	573	~0

MODERN THEORY AND GALILEO'S RESULTS for the experiments in *f. 116* and *f. 114* are compared in this table. In both cases the plane was inclined at an angle of 30 degrees. The final drop for the experiment in *f. 116* was 828 *punti* (77.7 centimeters) and the final drop for the experiment in *f. 114* was 500 *punti* (46.9 centimeters). At an angle of 30 degrees the distance the ball rolls down the plane is twice the vertical height of the plane above the table. Consistently negative departures of Galileo's results from modern theory for experiment in *f. 114* may be due to a slight difference in width of the groove, diameter of the ball or the slope of plane.

that the terminal speeds reached by the ball as it rolled from the four vertical heights originally mentioned in *f. 116* are very nearly the same as the terminal speeds calculated for four of the seven above-mentioned rolls we have assumed for the results in *f. 114*. We can accordingly consolidate both experiments by supposing the same tilt of 30 degrees was always used for the inclined planes, an assumption that is quite plausible. On the basis of our deductions we have compiled a table of data showing Galileo's experimental results.

The acceleration of a body falling under the influence of gravity at the latitude of Padua is 980.7 centimeters per second per second. Assuming that the ball rolled perfectly without slipping and touched the plane at only one point as it rolled, its acceleration down a plane at an angle of 30 degrees would be 350 centimeters per second per second. Its acceleration as it rolled along a groove in the plane, however, would be less, and the rate of acceleration is related to how much of the ball is above the top of the groove and how much is below. We shall be noting some experimental observations of this effect, but for our present purpose we assumed that in Galileo's experiment the ball accelerated at the rate of 320 centimeters per second per second. If in the second experiment the angle of Galileo's plane deviated very slightly from 30 degrees when it rested on a different table only 500 *punti* above the floor instead of 828 *punti*, such a departure would be sufficient to account for the consistently negative deviations of Galileo's measurements from modern theory [*see illustration above*].

The experiment recorded in *f. 114* is of much more use to historians of science than it turned out to be to Galileo himself. After he had obtained the seven measurements of the horizontal distances traversed he was unable to calculate how they compared with theory. It had been easy for him to combine uniform horizontal motion with accelerated vertical motion, but he did not see how he could combine a uniform "impetus" along a slope with vertical free fall. In due course he did learn how to determine the impetus of a projectile at any point along its path, and after that he could have interpreted the data in *f. 114*. Even though he preserved this scrap of paper with its experimental data, however, he appears not to have returned to analyze it. In a way that is fortunate for historians of science. If Galileo had been able to compute the figures we have been discussing, someone might now say that he had not conducted the experiments at all and had simply calculated the answers. As things stand there is no way to explain the series of numbers 253, 337, 395, 451, 495, 534 and 573 except by acknowledging that Galileo was a very careful experimentalist indeed, working with an inclined plane that was at an angle only imperceptibly different from 30 degrees and measuring distances to an accuracy of a millimeter.

Up to this point our interpretation of *f. 116* and *f. 114* as recording carefully planned experiments has depended entirely on paper work based on some diagrams and numbers. We have shown that Galileo's numerical data fit well with modern analysis that takes into account how much the linear acceleration is reduced by the energy involved in the ball's rolling, a factor of which Galileo was not aware. A random variation of the data in *f. 116* around the expected theoretical result makes it appear even more likely that the numbers were obtained experimentally and not calculated hypothetically. Still, if we could physically duplicate Galileo's experiments and not just reconstruct them on paper, the evidence would be even stronger. In short, what we have developed up to this point is itself a theory—not a physical theory but a theory of Galileo's procedures. It remains to test that theory as we would any physical theory, namely by actual experiment.

We obtained a plank that measured two inches by four inches by six feet and planed it smooth on one side. Two grooves were cut straight along its length. The plank was then mounted at an angle of 30 degrees from the horizontal along the edge of a level table that was 77.7 centimeters high. In the first trials we rolled a ball down one of the grooves and allowed it to strike a horizontal board before it took off in free flight. Although the ball bounced slightly, with the result that it stayed in the air longer than it would have if it had simply fallen from the edge of the plane, it nevertheless landed considerably short of the horizontal distances reported by Galileo for the same roll.

In an attempt to remedy the situation we installed a curved deflector at the base of the plane to convert the ball's motion from the 30-degree slope to the horizontal more smoothly. The results now obtained followed the pattern of Galileo's results more closely, although our measured distances were consistently somewhat greater than his. The ball fell on a sheet of carbon paper on top of a paper strip that was supported on a sheet of hard plastic resting on a smooth plank. The carbon paper marked the point of impact on the paper strip, and the distances from the edge of the table could be measured quite accurately.

When we duplicated the experiment shown in *f. 114*, no curved deflector was used, and a box supported the sheet of plastic and raised the point of impact to a vertical distance of 500 *punti* from the edge of the table instead of 828. Once again the measured distances followed the same pattern as Galileo's in *f. 114,* although again they were consistently slightly longer.

We used a steel ball four centimeters in diameter and rolled it in a groove 1.5 centimeters wide. We also tried rolling a second ball only 2.5 centimeters in diameter down a groove one centimeter wide. The results of the trials showed that the width of the groove significantly influences the rate of the ball's acceleration along the plane. The maximum acceleration of a ball rolling perfectly along a flat plane at an angle of 30 degrees is about 350 centimeters per second per second. By way of comparison, the four-centimeter ball in the 1.5-centimeter groove had an acceleration of about 325 centimeters per second per second; the 2.5 centimeter ball in the same groove had an acceleration of about 280 centimeters per second per second. The difference between the accelerations of the two balls was less for the narrower groove.

Our analysis of Galileo's data shows that 320 centimeters per second per second is a good intermediate value for his two experiments. The acceleration implied by the data in *f. 116* was larger and that implied by the data in *f. 114* was smaller. A good intermediate value for our two experiments is 330 centimeters per second per second, and the acceleration was somewhat greater for our duplication of the experiment in *f. 116* and somewhat less for our duplication of the experiment in *f. 114.* Thus our measured accelerations deviated from our intermediate value in the same way that Galileo's did from his intermediate value. Our experiments further indicate that Galileo did use a curved deflector at the end of his plane for the results shown in *f. 116.* Hence the diagram of his apparatus in *f. 175* was not entirely the fruit of his imagination but reflected some experimental design.

Two conclusions seem reasonable to us as a result of our experiments. First, the data we recorded are sufficiently similar to the data recorded by Galileo to verify the hypothesis that he experimentally obtained sets of numbers measured to three or four significant figures. Second, the experimental effects of the width of the groove provide a plausible explanation for the accelerations implied by Galileo's data being less than the theoretical acceleration of 350 centimeters per second per second that would hold only for a sphere rolling perfectly on top of a flat plane.

When Galileo published his discussion of the parabolic trajectory in 1638, he did not refer to any experiments. All he could derive was an ideal law that excluded the sources of actual variations. Hence he wrote: "I mentally conceive of some movable projected on a horizontal plane, all impediments being put aside. ... Equable motion on this plane would be perpetual if the plane were of infinite extent, but if we assume it to be ended, and [situated] on high, the movable (which I conceive of as being endowed with heaviness), driven to the end of this plane and going on farther, adds on to its previous equable and indelible motion that downward tendency which it has from its own heaviness. Thus there emerges a certain motion, compounded from the uniform horizontal and the naturally accelerated downward." That motion he showed to be parabolic.

Now that we have seen *f. 116* it is apparent that Galileo was describing as a mental conception something he had carefully observed with his own eyes 30 years earlier. The first historians of science jumped to the conclusion that that was what he had done. Recent historians of science, critical of their predecessors, have jumped instead to the conclusion that Galileo worked from pure mathematics without empirical evidence; faith in ideal Platonic forms rather than attention to physical detail, they say, opened the way to modern science. As far as Galileo is concerned, the earlier historians came closer to the truth. What they lacked in philosophical insight they made up for in common sense. To the conclusions of the recent historians we reply in the words of Salviati, who spoke for Galileo in his *Dialogue on the Two Great Systems of the World* (1632).

"What you refer to is the method Aristotle used in writing his doctrine, but I do not believe it to be that with which he [originally] investigated it. Rather, I think it certain that he first obtained it by means of the senses, experiments and observations, to assure himself as much as possible of the truth of his conclusions. Afterward he sought means to make them demonstrable. That is what is done for the most part in the demonstrative sciences.... And you may be sure that Pythagoras, long before he discovered the proof for which he sacrificed a hecatomb, was sure that the square on the side opposite to the right angle in a right triangle was equal to the squares on the other two sides. The certainty of a conclusion assists not a little in the discovery of its proof."

the Perfect Calculator

Space-age science has produced America's first portable display calculator with an optional tape.

Nobody's perfect! To be sure you're getting the correct answers from any calculator, you should run a tape. For only with a paper tape can you be absolutely sure.

Just flick a switch on the new Casio Mini-Printer, and your hand-held display calculator turns into a printing calculator recording every entry you make. Meanwhile, your entries appear on an 8-digit L.E.D. display, just like on any other pocket calculator.

A CASE FOR THE HORIZONTAL PAPER TAPE

As exciting as the Mini-Printer appeared to us, we couldn't get used to horizontal tape. But when we finally did, we actually liked it better. Here's why:

If you normally discard your tape after checking it, the horizontal tape has distinct advantages. It's easier to check because you read it as you naturally read—from left to right. It's easier to discard because it's smaller.

If you must attach your tape on a document, the horizontal tape is by far the neatest method because the tape, after mounting, looks like part of the document.

In short, once you discover the distinct advantages and the many new applications of the horizontal tape, you'll have no problem accepting its new format.

A CASE FOR THE LOW PRICE

The most inexpensive combination display/printing calculator with all the features of the Mini-Printer costs over $300 in today's calculator market. And it isn't portable. When you add the Casio's portability, the rechargeable batteries, the AC battery charger and the ease of operation, the Mini-Printer stands out as an incredible value. And therein lies the problem. The price of the Mini-Printer is so low that people tend to question the quality or look for the gimmick. There is no gimmick. The Mini-Printer is a "breakthrough product"—a unit whose advanced technology has totally revolutionized the printing calculator market.

THE STORY ABOUT THE UNIQUE PRINTING HEAD

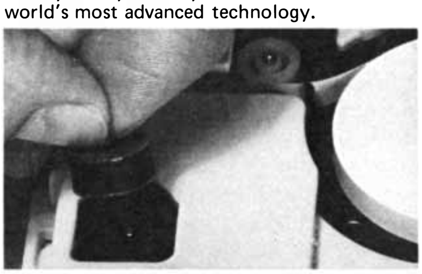
A calculator's printing head is the electromechanical device that prints the paper tape. On a standard 8-digit printing calculator with vertical tape, a printing head has eight rows of ten digits or a total of 80 separate digits. When calculations are performed, all eight rows must respond.

By using a single horizontal printing disc, the Mini-Printer requires only ten digits in one row. There is less power required, less wear, easier inking, and little or no maintenance.

This new printing head is so revolutionary that over 20 patents were applied for after its development. The system is so simple and has so few moving parts that it can be inexpensively manufactured with the most expensive materials.

The 5/16" tape is the most efficient form of record keeping. Simply attach it horizontally, exactly where you want it, with a piece of Scotch® brand transparent tape. Your papers are kept much neater. The tape is perfect for personal record keeping, homework, and documents where verification is required.

The Mini-Printer uses standard paper—readily available 5/16" ticker tape. The tape costs only 50 cents a spool and gives over 10,000 impressions. (In a standard printer, that's equivalent to 1,250 rows of eight digits each row.) The ink cartridge will last over 100,000 impressions or almost a year and costs only $2.00 to replace. In short, the Mini-Printer will save you money when you buy it and money every time you use it—all with the world's most advanced technology.



There's no messy ribbon to worry about. The Casio uses an ink cartridge which is replaced by slipping off the old one and slipping on the new one.

Part of its great design is its simplicity and minimal service requirements. It should last years without any service whatsoever, but should service be required, simply slip your 25 ounce Mini-Printer into its handy mailer and off it goes to one of several prompt national service centers. The Mini-Printer has already been proven by months of testing and actual consumer usage and is backed by a solid one year warranty.

OH YES, THE FUNCTIONS

Almost lost in the unit's excitement is what you actually get in functions. The Mini-Printer has square root, a fully addressable memory, a "Paper Advance" key and a "Result Only" key that prints only the results of each calculation.

Your eight digit L.E.D. display has a full floating decimal, overflow indicator and even a floating negative sign to signal negative balances. The unit performs the functions algebraically—exactly as you think. Power is provided by rechargeable batteries that will last years without replacement, and the AC charger actually signals when a full charge is reached.

The Mini-Printer measures only 1¾" x 4¼" x 7½" so you can now have the convenience of your own printing calculator without the desk clutter. In its display or printer mode, the calculator is totally silent. Only when a key is depressed does it generate any sound. This is one of the newer options now offered in only the very expensive printers.

JS&A IS AMERICA'S LARGEST SINGLE SOURCE OF ELECTRONIC CALCULATORS

When we took on the challenge of selling the Mini-Printer, our first reaction was simply, "How does one describe a miracle, and who would believe it?" And as objective as we have tried to be, we can't help but let our enthusiasm call the Mini—Printer—"the perfect calculator." As America's largest single source of electronic calculators, that's quite an endorsement.

The key to appreciating the Mini-Printer lies not just in understanding our description nor even in seeing it in person. It lies in actually using it at your home or office under your everyday operating conditions. That's why buying directly from our company makes sense. We'll give you the opportunity to use the Mini-Printer, without obligation, for two full weeks. If you're not absolutely convinced it's the best calculator investment you've ever made, return it and we'll gladly issue you a prompt and courteous refund.

TO ORDER BY MAIL

To order your Mini-Printer, simply send your check for $132.95 ($129.95 plus $3.00 postage, insurance and handling) to the address shown below. Illinois residents add 5% sales tax. Or use your credit card and call our toll-free number below. By return mail, you'll receive your Mini-Printer, three free rolls of tape, an AC charger/adapter and complete instructions. You may also order 10 rolls of extra tape and an extra ink cartridge for only $7.95 extra. Order your Mini-Printer at no obligation today.

FREE 2 WEEK TRIAL **$129.95**

JS&A NATIONAL SALES GROUP

DEPT. SCA-8 4200 DUNDEE ROAD
NORTHBROOK, ILLINOIS 60062

CREDIT CARD BUYERS CALL:

(800)323-6400

In New York City and Suburbs Call: (212)964-0690
In Illinois (312) 498-6900 ©JS&A Group, Inc., 1975

Shown actual size:

2×3=6....3×6=18....T24....2−3÷5×6+3=1•8....56−567=−511....144√12....5÷3=1•6666666

MEMORY TOTAL AND RECALL — CHAIN CALCULATIONS — NEGATIVE BALANCE — SQUARE ROOT — FLOATING DECIMAL

MATHEMATICAL GAMES

From rubber ropes to rolling cubes, a miscellany of refreshing problems

by Martin Gardner

The following problems, chosen for their variety and unfamiliarity, will be answered next month. Letters with solutions and comments are always welcome, although it will not be possible for me to respond to more than a fraction of them.

1. The Rubber Rope. A worm is at one end of a rubber rope that can be stretched indefinitely [*see illustration below*]. Initially the rope is one kilometer long. The worm crawls along the rope toward the other end at a constant rate of one centimeter per second. At the end of each second the rope is instantly stretched another kilometer. Thus after the first second the worm has traveled one centimeter and the length of the rope has become two kilometers. After the second second the worm has crawled another centimeter and the rope has become three kilometers long, and so on. The stretching is uniform, like the stretching of a rubber band. Only the rope stretches. Units of length and time remain constant.

Does the worm ever reach the end of the rope? If it does, estimate how long the trip will take and how long the rope will be. This delightful problem, which has the flavor of a Zeno paradox, was devised by Denys Wilquin of New Caledonia. It first appeared in December, 1972, in Pierre Berloquin's lively puzzle column in the French monthly *Science et Vie*.

2. The Sigil of Scoteia. One of James Branch Cabell's finest novels, *The Cream of the Jest,* involves a series of hypnotic dreams that Felix Kennaston induces by staring at half of the Sigil of Scoteia. At the end of the novel Kennaston discovers that the Sigil is nothing more than the broken top of a cold-cream jar designed by someone who "just made it up out of his head." It "is in no known alphabet. It blends meaningless curlicues and dots and circles with an irresponsible hand." The artist had sketched a crack across it "just to make it look ancient like."

A picture of the complete Sigil, which appears in most editions of the book, is reproduced in the top illustration on the opposite page. I remember puzzling over it many years ago and vainly trying to decode it. Designed by Cabell himself, it is not a cipher at all. Can the reader read it?

Cabell is still out of favor with most literary critics, but he has a loyal band of admirers who support the James Branch Cabell Society and its official organ, *Kalki.* (Inquiries can be addressed to the treasurer, William Jenkins, 172 Balsam Road, Wayne, N.J. 07470.) It was in the sixth issue of *Kalki* (1968) that I learned the Sigil's secret.

3. The Integer-Choice Game. Two mathematicians are drinking beer. After each has been served a glass they arrive at who pays for the round by simultaneously writing any positive integer on slips of paper. They compare the slips. Whoever wrote the larger integer has to pay for the round, unless the integer is larger by only 1. In that case the person who wrote the smaller integer pays for that round and the next as well. If both players have chosen the same integer, they play again.

To describe the game positively: The person who writes the smaller number scores a point, unless it is smaller by only 1. In that case the other player scores two points. For example, if the numbers were 12 and 20, 12 would win a point. If the numbers were 12 and 13, 13 would win two points.

The game is fair, but what strategy is best in the sense that no other strategy can beat it in the long run, and if any different strategy is followed, there is a counterstrategy to beat it?

The answer is surprising. I shall not have space next month for a proof, but I shall give the strategy and references to where the proof can be found. The game was devised by N. S. Mendelsohn and Irving Kaplansky, but it was Paul Halmos who called it to my attention.

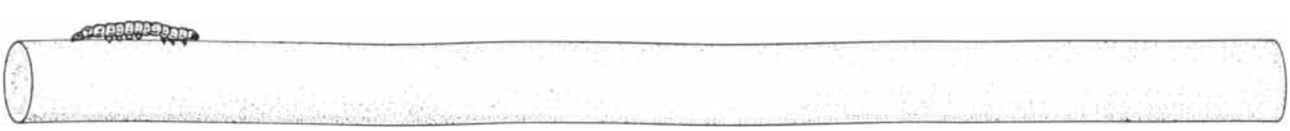
4. The Three Circles. Draw three nonoverlapping circles of three different sizes anywhere on a piece of paper. For each pair of circles draw their two common tangents. If you have not encountered this beautiful theorem before, you may be surprised to find that the intersections of the three pairs of tangents lie on a straight line [*see bottom illustration on opposite page*].

As one would expect, there are many ways the theorem can be proved by adding construction lines to the figure. As *Popular Computing* reported in its issue of December, 1974, however, the theorem lends itself to an elegant solution if one leaves the two-dimensional plane for a three-dimensional extension. Quoting from an earlier book in which they found the problem, the editors of the magazine report that when the theorem was shown to John Edson Sweet, a professor of engineering at Cornell University who died in 1916, he studied the picture for a moment and then said, "Yes, that's perfectly self-evident."

What was Professor Sweet's sweet solution?

Popular Computing, a monthly periodical that stresses the recreational aspects of computer mathematics, is now in its third year. Anyone interested in it can write for information to the publisher, Fred Gruenberger, Box 272, Calabasas, Calif. 91302.

5. The Mutilated Score Sheet. The top illustration on page 114 reproduces the score sheet of a chess game played in a German chess club in 1897. As you can see, a cigar or a cigarette has burned a hole in the sheet, so that Black's moves, which ended in Black's win on his fourth



A worm crawling on a stretching rubber rope move, have been obliterated. Can you reconstruct the game?

I wish to thank Randolph W. Banner, of Newport Beach, Calif., for passing along this amusing problem. He writes that he thinks he saw it in an English periodical published about 1920.

6. Kaprekar's Self-Numbers. D. R. Kaprekar is a 70-year-old mathematician, diminutive in body but large in brain and heart, who lives in India. For more than 40 years he has been doing highly original work in recreational number theory, at times aided by grants from Indian universities. He contributes frequently to Indian mathematics journals, speaks at conferences and has published some two dozen booklets written in broken English that can be purchased only by writing to him at 311 Devlali Camp, Devlali, India.

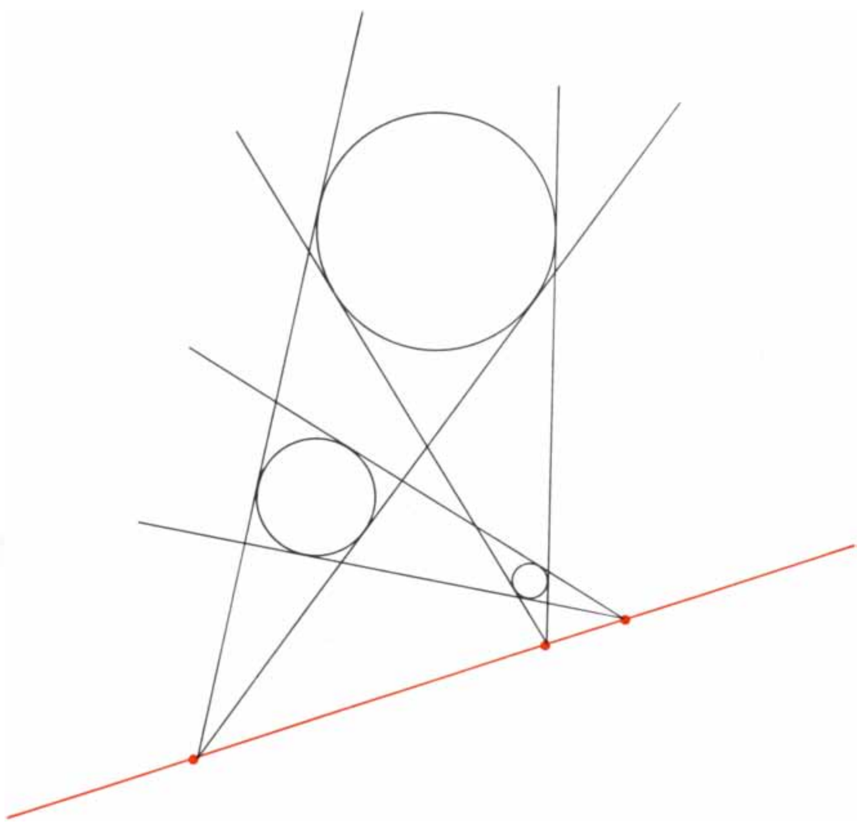
Kaprekar is best known outside India for his discovery more than 20 years ago of "Kaprekar's constant." Start with any four-digit number in which not all the digits are alike. Arrange the digits in descending order, reverse them to make a new number and subtract the new number from the first number. If you keep repeating this process with the remainders, you will (in eight steps or fewer) arrive at Kaprekar's constant, 6,174, which then generates itself.

Here we concern ourselves with a remarkable class of numbers called self-numbers, discovered by Kaprekar in 1949. He has written many pamphlets about them. They are virtually unknown outside India, although last year they turned up briefly (under another name) in an article in *The American Mathematical Monthly* (April, 1974, page 407). The article contains a proof that there is an infinity of self-numbers.

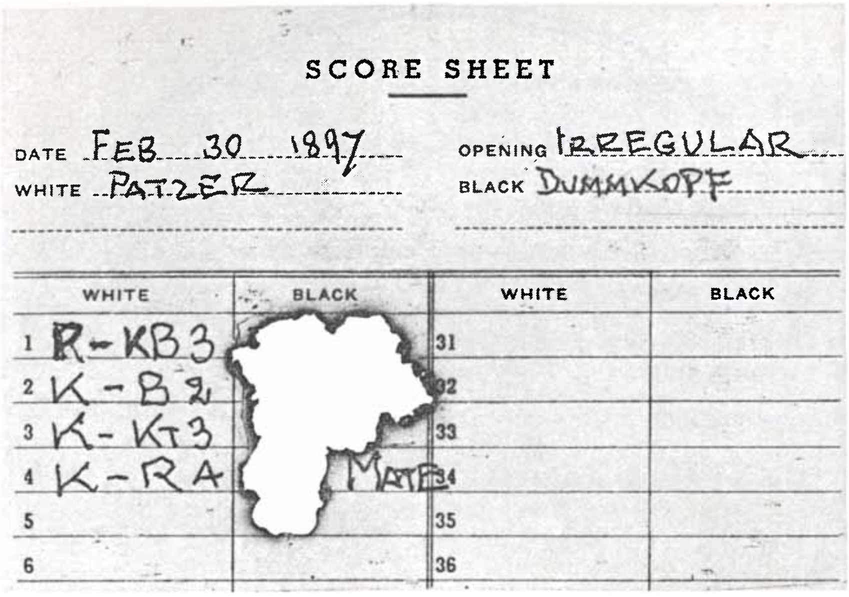
In explaining self-numbers it is best to start with a basic procedure that Kaprekar calls digitadition. Select any positive integer and add to it the sum of its digits. Take 47, for example. The sum of 4 and 7 is 11, and 47 and 11 is 58. The new number, 58, is called a generated number. The original number, 47, is the generator. The process can be repeated endlessly, forming a digitadition series: 47, 58, 71, 79, 95, ...

No one has yet found a nonrecursive formula for the partial sum of a digitadition series, given its first and last terms, but there is a simple formula for the sum of all the digits in a digitadition series. Simply subtract the first number from the last and add the sum of the digits in the last number. "Is this not a wonderful new result?" Kaprekar asks in one of his booklets. "The Proof of all this rule is

The Sigil of Scoteia



The three-circle problem



What were Black's moves?

very easy and I have completely written it with me. But as soon as the proof is seen the charm of the whole process is lost, and so I do not wish to give it just now."

Can a generated number have more than one generator? Yes, but not until the number exceeds 100. The smallest such number (Kaprekar calls it a junction number) is 101. It has two generators: 91 and 101. The smallest junction number with three generators is 10,000,000,000,001. It is generated by 10,000,000,000,000, 9,999,999,999,901 and 9,999,999,999,892. The smallest number with four generators, discovered by Kaprekar on June 7, 1961, has 25 digits. It is 1 followed by 21 zeros and 102. Since then he has found what he conjectures to be the smallest numbers with five and six generators.

A self-number is simply a number that has no generator. In Kaprekar's words, "It is self born." There is an infinity of such numbers, but they are much scarcer than generated numbers. Below 100 there are 13: 1, 3, 5, 7, 9, 20, 31, 42, 53, 64, 75, 86 and 97. Self-numbers that are prime are called self-primes. The familiar cyclic number 142,857 is a self-number (multiply it by the digits 1 through 6 and you always get the same six digits in the same cyclic order). The numbers 11,111,111,111,111,111 and 3,333,333,333 are self-numbers. Previous years of this century that are self-numbers are 1908, 1919, 1930, 1941, 1952, 1963 and 1974.

Consider the powers of 10. The number 10 is generated by 5, 100 by 86, 1,000 by 977, 10,000 by 9,968 and 100,000 by 99,950. Why is a millionaire such an important man? Because, answers Kaprekar, 1,000,000 is a self-number! The next power of 10 that is a self-number is 10^{16}.

No one has yet discovered a nonrecursive formula that generates all self-numbers, but Kaprekar has a simple algorithm by which any number can be tested to determine whether it is self-born or generated. Readers are asked to see if they can discover the procedure. If they can, it should be easy for them to answer this question: What is the next year that is a self-number?

7. The Colored Poker Chips. What is the smallest number of poker chips that can be placed on a flat surface so that chips of three different colors are required to meet the condition that no two touching chips are the same color? As the bottom illustration on this page shows, the answer is obviously three.

Our problem is to determine the smallest number of chips that can be placed flat on the plane so that chips of not three but four colors are necessary for meeting the same proviso.

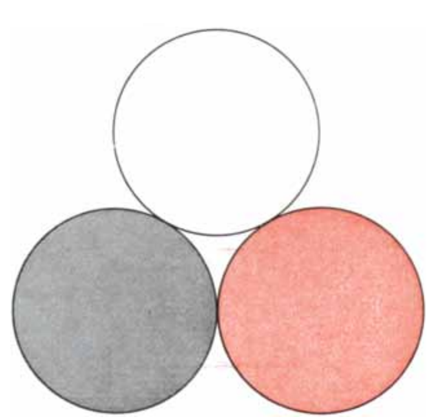
8. The Rolling Cubes. For this beautiful combinatorial puzzle, invented by John Harris of Santa Barbara, Calif., you must obtain eight unit cubes. On each cube color one face and make the opposite face black. (Of course, you may distinguish the two faces in any other way you like.) Place the cubes in a shallow 3-by-3 box (or on a 3-by-3 matrix) with the middle cell vacant and all cubes black on top [*see illustration on page 116*].

A move consists in rolling a cube to an empty cell by tipping it over one of its four bottom edges. The problem is to invert all eight cubes so that their colored sides are up and the center cell is vacant as before. This is to be done in a minimum number of moves. I should appreciate receiving all solutions of fewer than 40 moves. Next month's answer will give Harris' best solution (it is fewer than 40), but in a later column I shall report on readers who do as well or better. For uniformity in recording solutions please use *U, D, L, R* for roll up, down, left and right, and start all solutions with *URD*. (Any other way of starting is symmetrically the same.)

The combinatorial playing-card problems given in this department last November produced memorable letters from readers who analyzed and extended the problems. Alan Hadsell and Stoddard Vandersteel together used a minicomputer to generalize David L. Silverman's problem. When the highest card value does not exceed 13, solutions exist only for $n = 3$, 5, 8, 9, 10, 12 and 13, and each solution is unique. From 14 through 31 all values of n have multiple solutions. They report that the number of solutions, beginning with $n = 14$, are 2, 4, 3, 2, 5, 15, 21, 66, 37, 51, 144, 263, 601, 1,333, 2,119, 2,154, 2,189, 3,280, ...

Angeloo Papaikonomou, a bioengineer at the Free University of Amsterdam, after exploring the general problem for squares turned to the general problem when cubes are substituted for squares. He was surprised to find that for every solvable n the solution is unique. The series of solvable n's begins 7, 19, 26, 37, 44, 56, 63, ... Papaikonomou found a simple recursive procedure that both gives this series and constructs the solutions.

Roland Silver of San Cristobal, N.M., exploring botdrops for cycles, discovered why the game usually terminates in a



Map-coloring problem with poker chips

king-queen loop. There are no other 2-cycles and none of 3 and 4. Although 5-cycles exist, the probability of entering one is low. For example, if the face-up packet, reading up from the bottom, starts with 10, *J*, 2, 3 and has an ace on top, the packet is in a botdrop 5-loop.

Herbert S. Wilf of the University of Pennsylvania reported a delightful discovery about topdrops that provides a proof of the game's finiteness. A card is in "natural position" if its value is the same as its position in the packet. For example, if the face-up packet, reading down from the top, is

$$7, 2, J, 8, 5, K, 6, A, 9, 10, 3, Q, 4,$$

there are five cards (2, 5, 9, 10, *Q*) in natural position. If we take these values as powers of 2, we can create what I shall call the Wilf number: $2^2 + 2^5 + 2^9 + 2^{10} + 2^{12} = 5{,}668$. After any move in topdrops the Wilf number must increase.

"The reason that the number increases," Wilf writes, "is that the cards which were in natural position and which were too far down in the deck to be reached by the reversal operation will still be in natural position afterward. The fate of the cards which are involved in the reversal is less clear, except for one thing: the card which was on top before the move will be in natural position after the move, and its power of 2 is large enough to drown out any changes from cards above it. (A power of 2 is larger than the sum of all earlier powers of 2 by exactly one unit, a fact which is the basis of binary counting.)

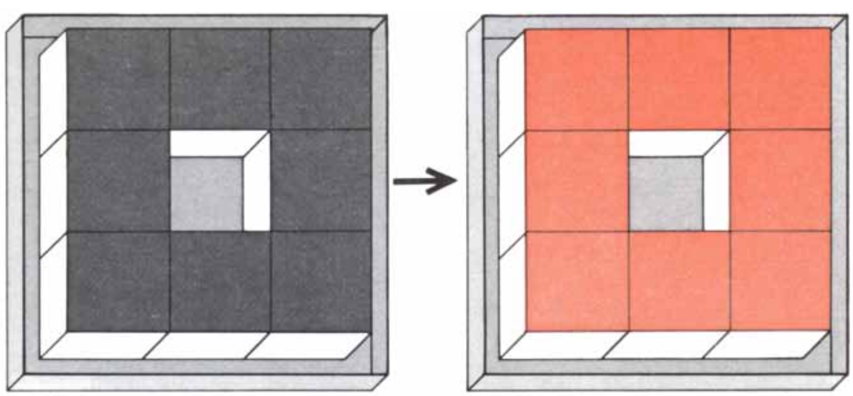
"Since the numbers increase steadily but cannot exceed 16,382, it follows that the game must halt after at most that many moves. A slightly more careful study shows, in fact, that for a game with n cards, no more than 2^{n-1} moves can take place."

This raises an interesting unsolved question: What arrangement of the 13 cards provides the longest possible game of topdrops?

Lance Rips, Henry B. Schuyler, Mel Stover and Julian Weitzenfeld pointed out that a simpler version of Tom Ransom's card-turning puzzle, without Ransom's joke of using what magicians call a double-back card, has been the object of research by psychologists. It is considered, for example, in *The Psychology of Reasoning*, by Peter Wason and Philip Johnson-Laird (Harvard University Press, 1972), and Wason's article, "The Psychology of Deceptive Problems," in *New Scientist* (August 15, 1974, pages 382–385).

Reaction to the disclosure that a card could be double-backed varied enormously. Jay Snyder began a letter: "Wait a minute! Hold it! Point of order! Not so fast! One moment please! Pull over, buddy." Stover sent a set of five taped-down cards to Wason at University College London for analysis. (Wason is planning to write an entire book on the problem.) When Wason was later informed of what he would have found if he had pulled off the cards to check their undersides, he opened his next letter, "Mea culpa!"

Several readers found the problem's wording ambiguous as to the meaning of "minimum number of cards" to be turned. The problem is to specify a minimum set of cards that, when they are reversed, will in all possible cases guarantee a correct answer to the question: Are all red-backed cards jokers? (The problem is best presented with cards that may have either red or blue backs.) If we do not care whether or not our answer is correct, obviously we can "answer" the question by turning no cards. In some cases turning one card will guarantee a correct answer. What we want is the smallest set that will provide a correct answer for all cases.



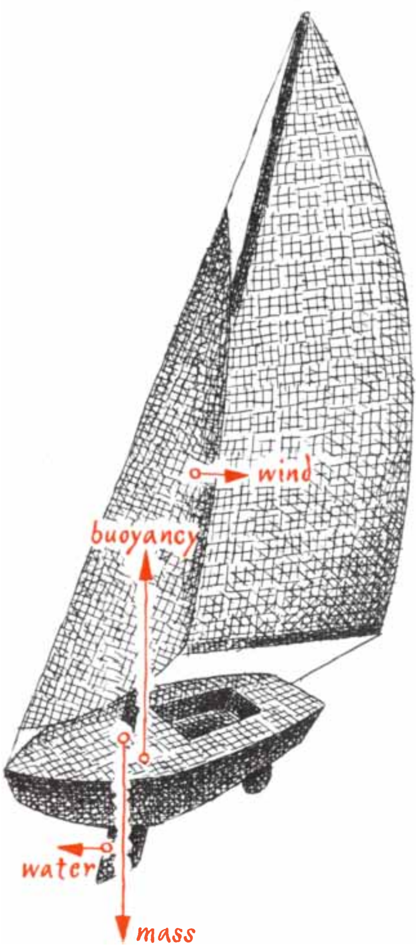
Rolling-cube puzzle

THE AMATEUR SCIENTIST

The ultimate in sailing is a rig without a hull

Conducted by C. L. Stong

Sailors have long recognized the hull as the chief impediment to speed in a racing yacht. That being the case, why not omit the hull? A conventional yacht is propelled by the net force of wind and water acting on the sails and the keel. In contrast, the forces that act on the hull merely keep the yacht afloat, meanwhile abstracting some of the energy provided by the wind. Even for highly streamlined yachts, such as contenders in the America's Cup races, most of the power of the wind is spent by the hull to stir the water. The hull functions merely as a rigid link that ties the sail to the keel (and also carries the crew).



The pattern of forces acting on a sailboat

Such chores might better be assigned to a more efficient device. A few years ago J. G. Hagedoorn (Wijttenbachweg 43, 2407 Oegstgeest, the Netherlands) set out to contrive the ultimate sailing machine: a yacht that requires no hull. Hagedoorn, who is professor of geophysics at the University of Leiden, describes the development of the system as follows:

"The most important criterion for the quality of any sailboat is its performance in sailing close-hauled into the wind. A vertical plane contains the four main resultant forces of the wind and water [*see illustration at left*]. The force of the wind (*A*) on the sail can be regarded as acting at a single point. An opposing force (*W*) in the same plane effectively acts at a single point on the keel. The two forces *A* and *W* are displaced, forming a couple not unlike a seesaw.

"This turning moment is opposed by the forces of buoyancy (*B*) and weight, or mass (*M*), that are generated by the hull in response to gravity. These two forces also lie in the plane of *A* and *W*. The centers at which *B* and *M* act are displaced to form a couple that opposes the moment of the couple formed by *A* and *W*.

"In normal operation the boat heels (tilts to leeward) to the point at which the vector sum of the moments generated by the couples of *A* and *W* and *B* and *M* equals zero. The sail makes an angle with respect to the direction of the wind. The keel makes a different angle with respect to the motion of the boat through the water. A force that arises from the deflection of air and water by the sail and the keel drives the boat to the speed at which the opposing force of drag induced by the hull equals the driving force. Drag thus limits the speed of the craft.

"The simple picture of the two opposing couples was chosen to illustrate their essential difference. The buoyancy and the weight of the craft are constant forces. As the boat heels the center of buoyancy moves to leeward because of the width of the beam and the shape of the hull. The center of weight is determined in part by the inclination of the keel, which contains some of the mass. The center of weight can also be altered by shifting the position of the crew. The perpendicular distance between the centers of buoyancy and weight (*B* and *M*) is severely limited.

"On the other hand, the couple formed by the wind and the water acting on the keel (*A* and *W*) depends on the strength of the wind and the resultant movement of the yacht through the water. For every sailboat of this type there is an optimal wind in which use can be made of an optimal couple between buoyancy and weight to achieve maximal performance. Crews of small sailboats demonstrate the optimal couple when they stretch as far as possible from the windward gunwale.

"The essential difference between these two couples explains why large boats are faster than small ones. For example, increasing the linear dimensions of a sailboat 10 times might be supposed to result in a craft that would sail just about the same as it did before the modification. The area of the sails and the area of the drag-generating hull and keel would each have been increased by a factor of 100. Both the weight and the buoyancy, however, would have been increased by a factor of 1,000. As a result the proportion of the speed-determining area of the sail and the keel would be increased by a factor of only 10.

"Innumerable other factors influence the relative performance of sailboats: cross section, aspect ratio (length to width) of sail and keel, shape of the hull, smoothness of the hull and so on. The frustrating fact, however, is that mass and money must always win. The small, lively sailboat dashing along in a smother of spray, with the expert sailor balancing to windward on an outboard trapeze, is inexorably overtaken by the comfortably large and sedately moving yacht.

"The proa, the catamaran and the trimaran are all successful attempts to increase the arm of the couple formed by the buoyancy and the weight. In principle the catamaran and the trimaran have been fitted out respectively with one boat and two boats too many. In all but relatively calm seas destructive couples arising from wave motion can develop in multiple-hull craft.

"In contrast, the proa represents an almost ideal method of increasing in a small boat the length of the arm of the mass-buoyancy couple. This remarkable little vessel, which resembles a long but narrow canoe fitted with a triangular sail and a slender outrigger to windward, was developed hundreds (perhaps thousands) of years ago by natives of the Mariana Islands. The length of the arm of the mass-buoyancy couple is increased when a crewman, preferably a heavy one, moves out along the outrigger. The vessels were first described to the Western world in 1521 by a member of Magellan's crew. The little boats flitted along at speeds on the order of 30 kilometers per hour, approaching the velocity of the wind.

"Nonetheless, European sailors were not impressed. Evidently they preferred their slow, unwieldy boxes to the heathen contraptions. It is astonishing to me that no one, including the Polynesians south of the Marianas, realized that the secret of the ultimate sailing craft can be derived from the proa.

"The hull of the proa requires no inherent lateral stability. Stability is provided by the outrigger and its living weight. The hull can be as long and slender as wished. What is not generally realized is that a body of this shape requires an amazing amount of correction by a fin or a rudder to keep it going straight. No hint of the difficulty of the art of steering such a craft is suggested by a Papuan happily paddling along. When an amateur tries to operate such a craft, however, the hollowed log inexorably turns sideways. No amount of frantic paddling can prevent it from ramming a mudbank. The beginner must acquire through diligent practice an instinctive knack of flicking the paddle to create rudder action that anticipates the sideways motion and counteracts it before it starts. To keep a long, slender body aligned in a desired direction requires either an active rudder or some stabilizing couple.

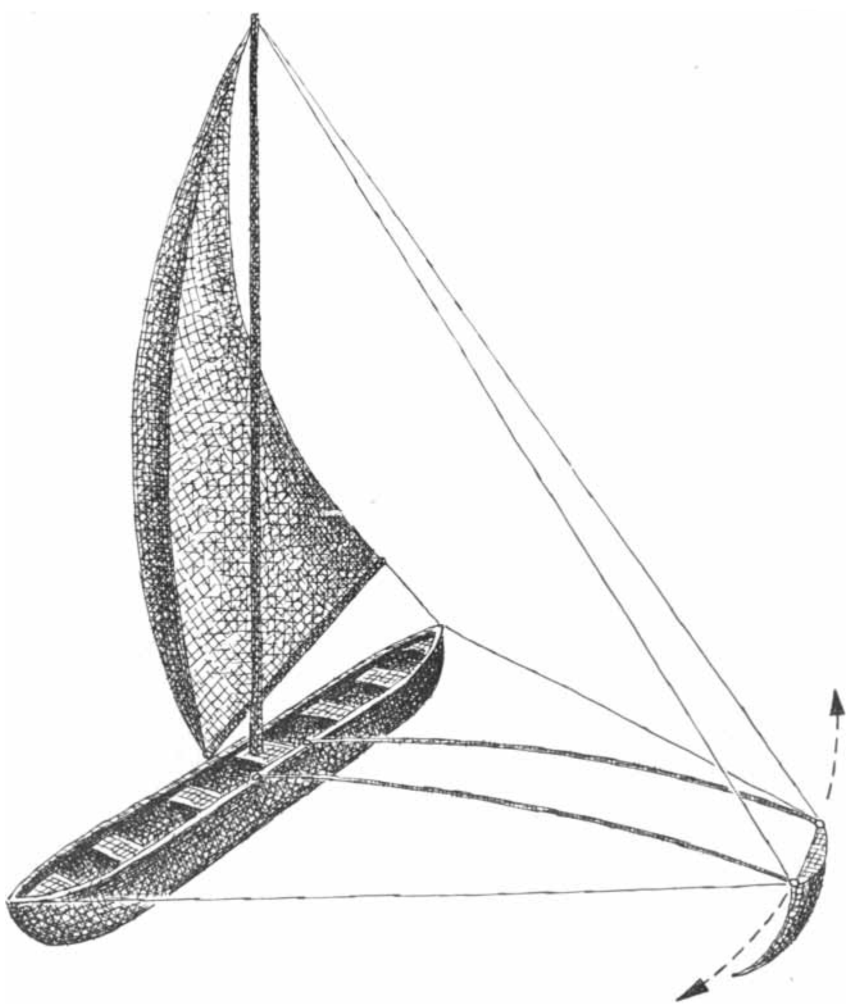
"The stabilizing couple in a sailboat can be provided by forces that act on the sail and the keel. Some boats can be adjusted to sail stably on a given course with the rudder in a fixed position. A boat with a centered keel is easier to adjust to this mode than one with leeboards. A 'luffboard' could be expected to work even better.

"The logical conclusion is that the best position for the keel is at the end of a windward outrigger. This takes the trickiness out of the proa. Moreover, the live ballast can be dispensed with because the keel can be shaped to generate a force that opposes the part of the wind force that tends to tip the boat and lift the keel from the water.

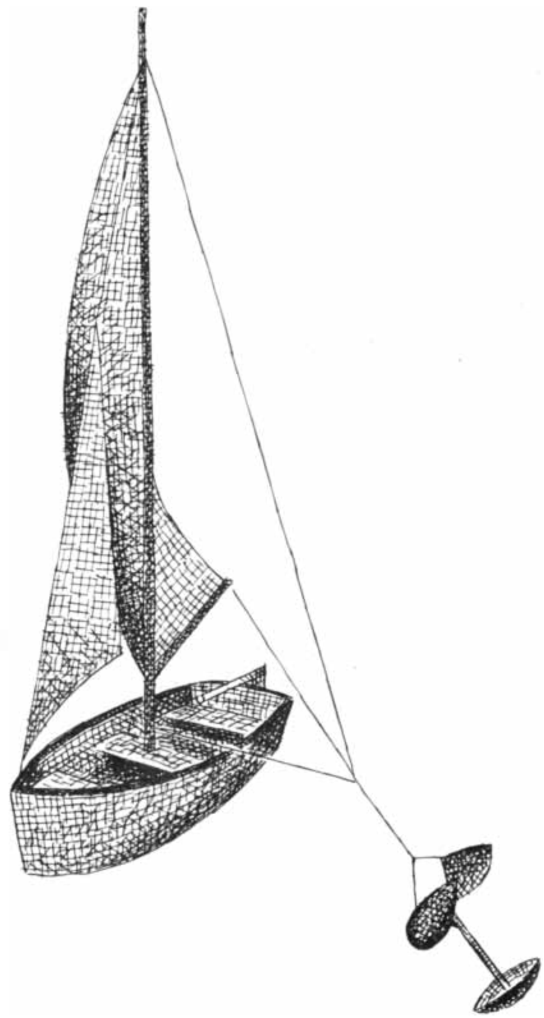
"It turns out that the required keel has the shape of a curiously curved symmetrical dagger, for which I have coined the name 'hapa,' primarily because it sounds Polynesian. The hapa is supported by two outrigger arms that pivot on vertical axes at both ends so that the orientation of the hapa with respect to the hull can be varied by changing the length of the lines connected to both ends of the boat [*see illustration below*]. This linkage system enables the crew to alter the hapa's angle of attack as required by the strength of the wind.

"As in the proa, the hull and the sail are symmetrical fore and aft. The hapa is similarly symmetrical. The craft can sail equally well in either direction, with the outrigger always to windward. The Mariana Islands, where the proa was developed, lie roughly at right angles to the prevailing westerly wind. Interisland traffic moves essentially north and south. To return from a voyage native crews simply reverse their sitting position instead of turning the proa around.

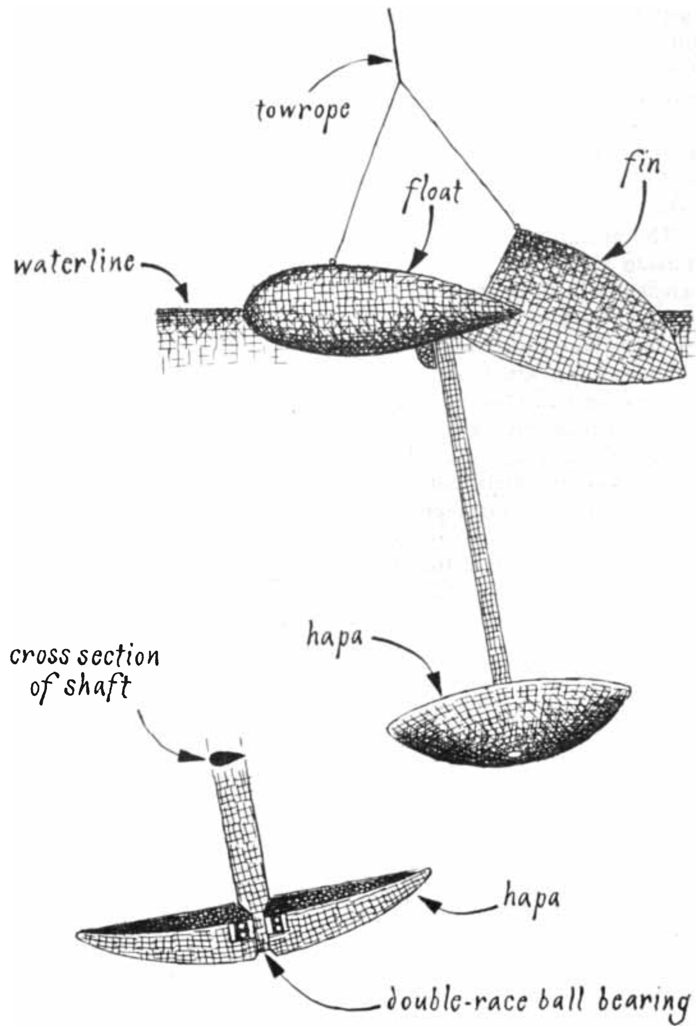
"Although natives of the Marianas were content with a boat designed to sail north and south, most of the rest of us insist on sailing in any direction we



A proa fitted with J. G. Hagedoorn's "hapa," or curved keel



A sailboat with a hapa



Details of the hapa

please. Changing course with a boat employing a hapa as an outrigger might be exciting in a strong wind. The outrigger assembly is a restraining device, but the actual villain is the boat. It cannot move sideways easily.

"Instead of 'going through the wind' in tacking to windward it would doubtless be best to pull the hapa to its new position and simultaneously let the sheet go at the mast so that the triangular sail would turn broadside and stop the boat. The other sheet would then be tightened near the foot of the mast, after which the sheet that is now aft would be pulled to set the new course. It is quite probable that the first time this maneuver was attempted in a strong wind the whole thing would capsize at the precise moment the hapa reversed direction.

"As I have mentioned, the connection between the hapa and the boat through the pair of rigid arms constitutes an unsatisfactory restraint. Why not replace the outrigger by a flexible system of lines? One problem is that to make the hapa behave as one wished would require so many guy ropes that their induced drag would be prohibitive.

"The idea of a hapa on a flexible leash was so attractive, however, that I could not resist the temptation to redesign it. The new hapa would have to resemble a paravane, which is the underwater kite employed in minesweeping. It turned out to be frustratingly difficult to design a hydrodynamically efficient hapa that can be pulled on a line. The contraption must follow a stable course just below the surface of the water, a problem that is more difficult than stabilizing a kite because the hapa must operate within the zone of radical transition from water to air.

"My first experimental models either jumped out of the water or dived to the bottom, even at walking velocities. It became obvious that a solution might lie in the elimination of cramping restraints. (Recognition of this principle made practical machines of helicopters.)

"Pursuit of this concept eventually resulted in a well-behaved hapa, at least through the anticipated range of running speeds. It includes a dishlike hydrofoil resembling a shallow meniscus lens supported by a rigid shaft on which the lens is free to turn [*see illustrations above*]. The hapa can never exert a couple, even if it strikes an obstruction.

"Although this arrangement is attractive, it imposes drastic limits on the shape of the hapa, which must function effectively as a hydrofoil. For example, the aspect ratio of a disk cannot exceed unity, whereas the hapa developed for hulls and rigs of the proa type may have an aspect ratio of 4 : 1 or even 5 : 1. The complete hapa includes the lenslike hydrofoil and the haft, or shaft, attached to a torpedo-shaped float fitted with a tail fin. The assembly is towed by a bridle.

"The principal parts are made of epoxy reinforced by fiberglass. The hollow lens fills with water. Its buoyancy is essentially neutral in both fresh water and salt water. The haft is mounted in a double-race ball bearing and can turn freely around the axis of symmetry of the lens. The fulcrum point at the other end of the haft (in the float) was placed about 5 percent forward of the axis of the lens for maximum pull-to-drag ratio.

"The haft takes the place of the three lines that customarily serve to maintain a hydrofoil in the water at the correct angle of attack for the optimum lift-to-drag ratio. It would be possible to support the disklike member by a bridle of three lines to achieve an essentially unrestrained system, but the scheme would be vulnerable to submerged obstacles. The apparatus would also be somewhat less free to turn.

"The chief difficulty of the central haft stems from its required thickness. It must be rigid and strong. The haft of one experimental model measured 1.5 by five centimeters. Its streamlined cross section introduces negligible drag, but it achieves this performance at the cost of symmetry. The haft has the same pernicious preference for traveling broadside to the direction of motion as the Papuan's dugout.

"A compromise must be made between the unacceptable high resistance of a symmetrical (cylindrical) haft and the instability that is introduced by the hydrodynamically ideal cross section. The unavoidable instability is counteracted by the float with its fixed fin. The float also performs the indispensable function of causing the system to operate near the surface.

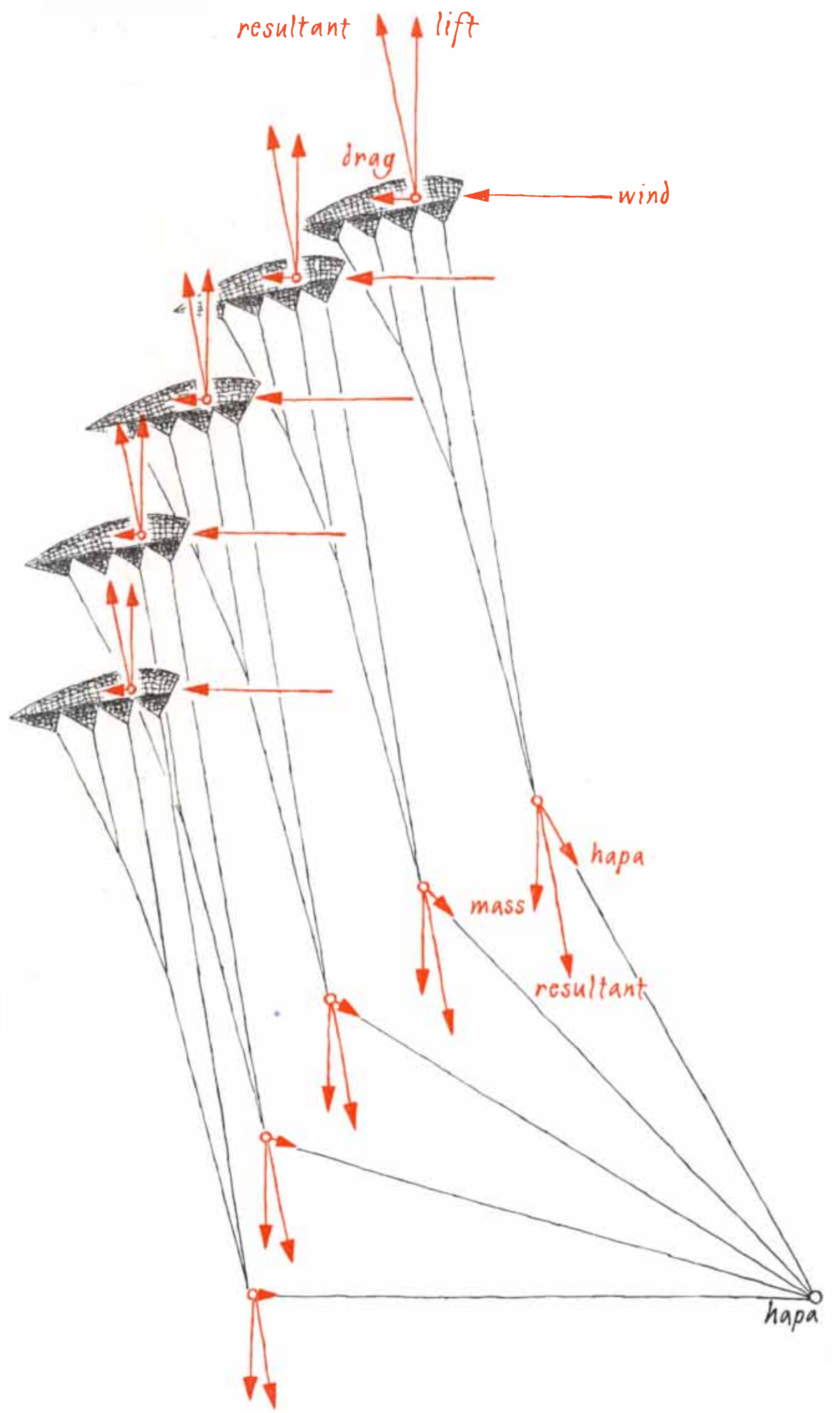
"The weight of the system is equal to half of its totally immersed buoyancy, and the centers of mass and buoyancy are well forward of the point of attachment of the haft to the float. This arrangement results in equal turning couples toward the correct, half-submerged condition, when some random force causes the float to plane up out of the water or to submerge. The towrope is attached to the bridle at a point that causes the axis of the haft to align itself with the rope.

"The clumsy-looking device actually behaves better than one might expect. Its pull-drag ratio is about 5 : 1, and the system is stable up to about 10 kilometers per hour.

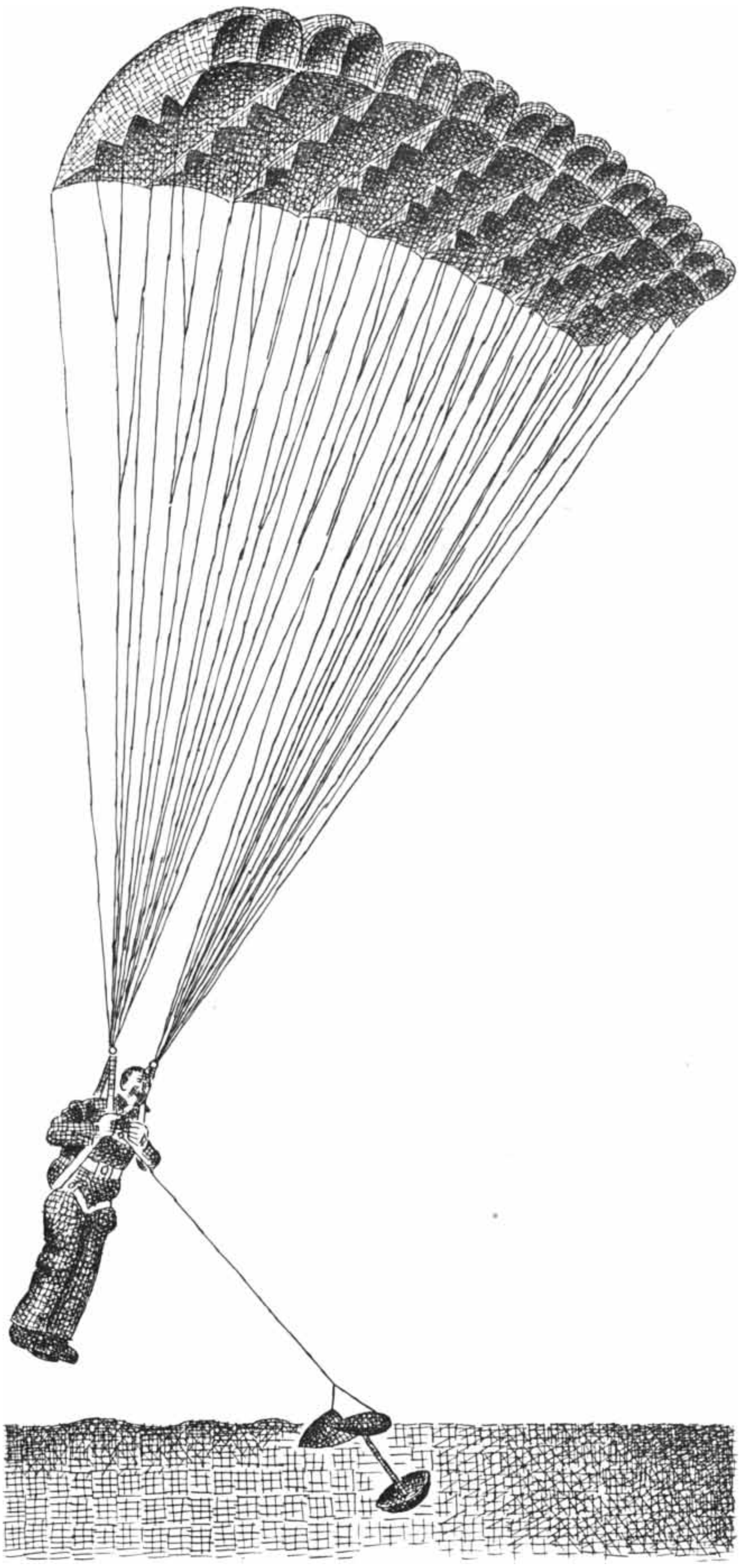
"This particular hapa can travel in only one direction, but its central plane of symmetry through the axis of the float enables it to be flipped over and so to travel equally well in the reverse direction. This capability may not always be desirable, because an accidental flip-over might result in an awkward situation. For ocean sailing left-hand and right-hand hapas might be safer.

"In the accompanying illustration of the hapa [*at left on opposite page*] the boat and the sail are intended merely to depict one of many possible ways of sailing with a hapa that is towed like a kite. The mast does not have the usual stays because the mast-sail system is fully supported by the three lines connecting the three corners to the hapa. A rudder would be useless.

"A change of course is effected by paying out or pulling in the line at the foot of the mast. In tacking, which involves changing course in such a way as to change the side of the sail the wind hits, the skipper can come about (tack while moving upwind) simply by pulling the hapa in with the front line and at the proper instant launching a second hapa to leeward. Although I have not tried to jibe (tack while heading downwind), the maneuver could probably be execut-



Pattern of forces acting between a Para-Foil and a hapa



Hagedoorn's ultimate sailing machine. (The pickup boat is not shown.)

ed by flipping the hapa over and paying out the front line. Both maneuvers could be exciting in a strong wind.

"Consider the consequences of introducing the hapa. The boat, always the dominant element in the conventional system of sailing, has been stripped of its keel, rudder, ballast and stays. It has been degraded to the status of a nuisance, contributing to the system little more than instability and undesirable resistance. The inevitable next step toward developing the ultimate (no hull) system of sailing involves the use of two kites, one in the water and one in the air. The boat is eliminated.

"A large inventory of designs for airborne kites is available for consideration by the experimenter. Only a few of them are efficient and safe enough for carrying a man. It seems incredible that aircraft designers should have been employing sophisticated airfoil sections for more than half a century while specialists in the design of kites and sails have for the most part blindly stuck with thin sections of sheet material. Perhaps the explanation lies in the convenience of stowing and transporting kites that are collapsible.

"A major breakthrough was the invention of an inflatable multicell airfoil of fabric by D. C. Jalbert of Boca Raton, Fla. The device was subsequently made operationally safe by J. D. Nicolaides, professor of aerospace at the University of Notre Dame. It is called the Para-Foil [*see illustration on preceding page*]. Its shape, in the form of a rigid, low-speed wing, is maintained entirely by air that enters openings at the leading edge to build up internal pressure. On landing the Para-Foil, like a conventional parachute, can be collapsed into a manageable bundle of lines and cloth, even in a fairly strong wind. One can also jump with it from an airplane. Moreover, with the Para-Foil a military flier who has had to abandon his airplane in combat can perform breathtaking evasive maneuvers that are not possible with a conventional parachute. He can also fly horizontally for distances some five times the ejection altitude and can land in stronger winds than is possible with a parachute because he can fly directly into the wind at the moment of landing.

"A kite is relatively easy to stabilize for free flight, because the attached weight tends to revert automatically to its correct vertical position below the lifting surface. It is much more difficult to achieve stability when the kite is tethered or towed. The Para-Foil is quite stable, even in the tethered condition.

When it is launched correctly, it gains height rapidly and flies stably.

"In the ultimate sailing system the Para-Foil is tethered to the hapa. The skipper is suspended by a harness, much like the one worn by the pilot of a hang glider [*see illustration on opposite page*]. The hapa extends to windward and passively takes the direction of the Para-Foil, which is controlled by the yachtsman (who might better be called the 'aquaviator'). To turn a Para-Foil and, even more important, to stop it from turning, the yachtsman pulls down on the appropriate webbing of the two webbings attached to the harness. Each webbing has connected to it half of the rear forked shroud lines. I am always impressed by the sensitivity of this control.

"This, then, is ultimate sailing: a system that requires the absolute minimum of force-generating surfaces in both air and water. It is clear that the surfaces of the hapa and the kite must be proportioned so that they are compatible for optimal efficiency to sail within a practical range of wind velocities during tacking, reaching (sailing more or less broadside to the wind) and running before the wind. The proportions are best determined experimentally. For wind velocities of about 20 kilometers per hour the ND2(242) Para-Foil, with which I have been experimenting, easily lifts 80 kilograms (177 pounds). It has an aspect ratio of 4 : 1 and an area of 22.48 square meters (242 square feet).

"Beginners should learn to sail with a relatively slow hapa, similar to the one illustrated here. This hapa's towrope is attached to the bridle at the point where the angle of attack results in the maximum lift-to-drag ratio. It is obviously possible to fix lines to the hapa for altering its angle of attack and thus its velocity.

"The advanced student learns to control the hapa so that it tends to occupy a following position on a close-hauled course and a forward position on a reaching or running course. To attain the forward position the sailor lets the hapa gain on him and then tries to make it run faster by applying more force to the towline at a low angle. In this maneuver the control lines of the Para-Foil must be manipulated in such a way that the sailor almost gets his feet wet.

"Both the sail and the hapa in particular are still very much in the first stage of development. A tremendous amount of imaginative thinking, experimenting and designing will have to be completed before the first aquaviator wins a transatlantic sailing race."


The Lives of a Cell
Lewis Thomas
THE CHEMIST'S COMPANION
Theory of Computation
THE ENCYCLOPEDIA OF PHYSICS

BOOKS

World arms in a period when the tactical balance is shifting in favor of the defense

by Philip Morrison

ARMS AND STRATEGY: THE WORLD POWER STRUCTURE TODAY, by Laurence Martin. David McKay Company, Inc. ($15.95). The arms industry, largest in the world, has perforce acquired its professional critics, a body of theory and a literature. Nowadays the intense interaction of means and ends symbolized by the mushroom cloud has stimulated a changing discipline in which new missile types and the intricate logic of nuclear force and counterforce hold the stage alongside the antique military verities of conflict, alliances, fear and gain.

This book by the Professor of War Studies at King's College, London, is a serious reader's introduction to the new strategic world. The author says, without immodesty, that his book is "unusually comprehensive" and can prepare a reader to "make his own judgment of strategic happenings."

The well-made book opens as it must, with the central nuclear balance, the U.S.-U.S.S.R. pivot on which the world anxiously spins. Photographs bring the texture of reality to these remote and disagreeable topics; many specialized maps and graphs, lightened perhaps a bit self-consciously with rows of little planes and missiles, add valuable specificity to the text. There follows an account of the lesser nuclear powers and the issues of proliferation. (One photograph shows the new mushroom cloud rising over the salt flats near Lop Nor in old Sinkiang.) Then the technology of limited war is examined in detail: the nature of battle by land, sea and air.

Perhaps the most novel point arises here. The very light guided missile, made cheap enough and small enough to be a ubiquitous weapon of infantry patrols or small naval patrol boats, may soon come to threaten all but fatally the helicopters, the heavy tanks, the costly and ferocious attack-fighter planes and even the largest ships of the great fleets. Integrated circuits (consider your pocket calculator) might endow with guidance and some discrimination the 50 pounds of rocket fuel and shaped explosive charge that form the basis of such weapons, which can be guided by wire, radar, laser or infrared radiation. Maybe those three British commandos in blackface shown lying in the salt grass alongside their Rapier missile system—a few boxes on a heavy tripod and a set of four rockets on a stand rather like a vacuum cleaner—really can bring down a multimillion-dollar fighter once in a while. The rocket would be easily the victor even at a rate of one hit in hundreds of rounds. If the cost is low enough and the chips are well designed for convenient change to frustrate countermeasures, then "the resulting environment could be nerve-racking for support aircraft." Tanks too, even with spaced armor, may become vulnerable to volleys of cheap missiles flying in from every infantry patrol on the scene. (The gloomy prospect of today's world suggests that we may have a lesson in this problem of technology assessment all too soon in the Middle East.)

Professor Martin devotes a chapter to guerrilla warfare worldwide, a social technique rather than a technology, and one to tactical nuclear weapons, a NATO and European problem with more than a little air of paradox. Then he treats the "arenas of conflict": Europe, the Middle East, Asia from the Indus to the Amur, southern Africa. Finally he considers, somewhat sketchily, the massive economics of arms, the flourishing arms trade with the third world and the spindly measures of arms control. The volume closes with a valuable summary of arms-control agreements and with the "military balance," a telegraphic listing of all forces country by country.

No other book like it has come to the reviewer's attention. Its closest counterparts are the well-known reference annuals of the Stockholm International Peace Research Institute (SIPRI) and the Institute for Strategic Studies in London. Indeed, those organizations created the international discipline, but their excellent publications are in no way introductory; they are generally more topical and hence not so comprehensive. (The London group supplied the military balance sheet given here.) The comprehensive annual weapons-system catalogues known as *Jane's*, on the other hand, are much more detailed, without the background and clear textual arguments of this book. Many readers with new personal interest or public responsibility in this complex field will want an overall guide to it.

The book is cool, too cool. Catastrophic events here take on a neatness that is far from real. Professor Martin is, one hopes, aware of this; it is a quality akin to the style of making quantitative statements that are easy but irrelevant because the qualitative ones are hard and distasteful. The U.S. war in Vietnam demonstrated that characteristic failing of the analyst very sharply. On Vietnam Martin comments: "It seems improbable that the bombing need have been used so freely.... The impact of television reporting...has perhaps been the most revolutionary innovation of all. [The] management of the mass media will constitute a new and difficult aspect of military operations by the democracies." Viewer, be warned.

Even these cool experts are not truly expert; they lack direct experience and detailed insight. They must report consensus, even opinion, and enthusiasts sometimes induce widespread follies. This kind of error is not often visible in this useful book (although the Weathermen are patently less active guerrillas than the PAIGC in Portuguese Guinea, in spite of the listing here), but it is not to be certified as being absent. One wonders a little why targetable missiles and the MIRV counterforce they imply were able to so greatly disturb the apparent stability of thermonuclear deterrence. Professor Martin sees this as proving that "there is no natural technological plateau on which strategy may rest." The point is not empty, but the reader is uncertain whether it is the security of the U.S. or the future of the U.S. Air Force that is the truer aim of this new "strategy." It is a pity that no such healthy cynicism crosses even for a moment the knowing but accepting pages of this insider's introduction to strategy today. (The work was completed in the middle of 1972; the October War, the oil embargo, the Indian atomic test and SALT II are of course absent.)

LEPROSY: DIAGNOSIS AND MANAGEMENT, by Harry L. Arnold, Jr., and Paul Fasal. Charles C Thomas, Publisher ($14.75). A PLAGUE OF CORN: THE SOCIAL HISTORY OF PELLAGRA, by Daphne A. Roe. Cornell University Press ($11.50). Diseases are not properties of organisms alone but of human society as well. By this fact medicine transcends biology to enter the arts. These two distinct books put the case well. The first is a book of good news. Between the 1953 first edition and the 1973 second one the ancient scourge leprosy was subdued. This small volume is not meant for the specialist but for the physician who has a patient suffering from leprosy. No more need he isolate or exile the patient; on the contrary, he can treat him as an outpatient living normally with his family and at work.

Leprosy has its mysteries still. It is today a disease of warm climates, between the 30th parallels except for Japan and Korea. From the eighth century to the 15th it was endemic in Europe, even in Britain and Scandinavia; in these northern lands it has steadily declined. It marks perhaps 10 million patients worldwide, at least several thousand in the U.S. Much suffering has come to its patients from translations of the Old Testament that used the term to render the Hebrew word for "any scaly disease"; with the name went the dread of disfiguring contagion and a real persecution of the unhappy patients. ("His clothes shall be rent, and his head bare, and he shall put a covering upon his upper lip, and shall cry, Unclean, unclean.")

We have suspected for a century, since the work of Armauer Hansen in 1874, that the disease is the result of infection by a specific bacillus. True, the full critical test according to the famous postulates of Koch has never been made: the agent will not grow on any artificial medium yet known, and scores of attempts to inoculate human volunteers have failed. (Two young marines from Michigan did develop the lesions simultaneously in tattoo marks they received in the same Australian shop on the same day in 1943.) The disease has, however, routinely taken in the footpads of inoculated mice and even systemically in the armadillo!

The key to the mystery seems to be at hand, at least at a pragmatic level. Most people are deeply immune to the disease; they cannot acquire it while their immune system remains intact. In the few highly susceptible individuals brief contact is enough to induce the disease, and prolonged contact makes it sure. There are two polar types of infection. Some patients have no cellular immunity at all; they develop the disease without resistance. In their skin lesions and even within internal organs the bacilli are abundant; these patients are infectious, and live bacilli can be demonstrated by the mouse-foot test. The other form is found in some people who can contract the disease but still hold strong cellular immunity. It does not spread within the body but slowly does damage mainly to nerve tissue. Its victims harbor few or no live bacilli and are not infectious. Sometimes this form of the disease regresses and heals spontaneously; its effect is almost always self-limited and relatively benign.

Such a pattern explains most of the fears and paradoxes surrounding the condition. Now we have two effective drugs. The sulfone dapsone, known since before World War II, slowly ends the growth of the bacillar rods. The fast-acting antibiotic rifampicin, found in a soil organism by Lepetit in Milan in 1957, was proved against leprosy in the 1960's. Before these drugs there was mainly fear and exile; when only the sulfone was known, the patient still had to be isolated as long as any bacilli at all could be detected microscopically in smears, often a matter of some five years. Now the presence of the live bacilli can be monitored by the mouse-foot test (or even by skillful microscopy); the antibiotic eliminates infective rods in a week, as against a couple of months for dapsone. Protracted (maybe lifelong) management with these or related drugs seems to offer full control of this old torture. Prevention of infection by chemotherapy is equally effective for those—such as the children of patients—who have had close contact with the infectious disease. Restoring the use of damaged and nerveless hands and feet is a slow, demanding and well-studied healing art.

Pellagra is truly a social disease. Its victims are the oppressed: sharecroppers, landless peasants, orphans, psychotic patients, cotton-mill workers tied to the company store. It is "the destiny of those who are poor because they are landless and are maize eaters because that is all they can get to eat."

The disease has three classic symptoms: the early rosy discoloration and peeling of the skin (the name is probably a form of the Italian for "rough skin"), an uncontrolled diarrhea and a psychosis (not entirely unjustified!) of depression and feelings of persecution. Death may follow eventually after a wasting of the muscles of the heart. A week's intake of the vitamin niacin suffices to reverse even a severe and chronic case. The body can synthesize its niacin from the amino acid tryptophan (no stranger to neurochemistry), and so the deficiency implies not only an inadequate intake of niacin but also an insufficiency of tryptophan. Ordinary maize lacks both substances; pellagra is "a plague of corn."

Yet the cultures of the Americas, which domesticated maize into the great staple of this hemisphere, did not know pellagra. Before Columbus no one from Massachusetts to Patagonia fed exclusively on a diet of cornmeal, turnips and greens; in the Americas corn and beans have gone happily together for six millenniums at least. (The very grubs in the green corn supplied some animal protein, along with the game of the hunter.)

It was left to Europe and its alienation of the tiller from his land to invent the condition. Maize culture came to the Old World certainly by Columbus' third voyage, if not before. By about 1700 maize was becoming an important crop on poor land in the warmer regions of Europe: in Spain, southern France and Lombardy. Its high yield of calories made it a cheap food. The peasant could subsist on a small share of the crop; the landlord sold the plentiful residue in town. The tax collector was happy, and the disease appeared. It was first recognized by Gaspar Casal, a physician of Oviedo in northwestern Spain. His poor laborer patients told him they suffered from *mal de la rosa,* a new ailment. Goethe saw the eaters of the grain mush called polenta when he came to the Italian side of the Brenner Pass in 1786: "I believe that their unhealthy condition is due to their constant diet of maize and buckwheat..., yellow polenta and black polenta." By 1820 the poorest farmers of the south of France showed the disease. By 1890 it was severe in lower Egypt, to which maize had been brought before 1840. There was some account of it in the U.S. in the 19th century, but the track is hard to read. Union prisoners at Andersonville died of hookworm and pellagra, a physician fellow prisoner judged, but we cannot be sure. The insane asylums of the Southern states first reported the disease in epidemic proportions about 1905, but there is evidence that it was the diagnosis that was the chief new element.

The theories were numerous: an infection from sheep, a toxin in corn or perhaps only in fungus-infected corn, a contagion carried by the buffalo gnat or a congenital defect in these low families. The great Cesare Lombroso thought it was partly fungus-borne and partly hereditary; family trees of pellagrins were still cited—with significantly many sibling sufferers—in an American medical journal of 1916!

It was the work of Joseph Goldberger in the South between 1914 and 1916 that established, by both heroic volunteer methods and epidemiological insight, that pellagra was a specific deficiency disease. In the village where the cotton-mill workers had the most pellagra they depended entirely on the company store. In healthier villages a market sold fresh foods and a vender came by with fruits and vegetables. Goldberger even connected the condition with the black-tongue disease in hound dogs fed on corn bread (as Eijkman and Grijns before him had seen beriberi in poultry fed on polished rice).

The vitamin was isolated and identified, and enriched cornmeal and flour were introduced in the U.S. by World War II. The worst pellagra years were 1929 and 1930; as the Depression deepened, the disease declined because when the cash crop King Cotton would not sell, the sharecroppers began to grow and eat their own peas, beans, peanuts and vegetables.

Pellagra has not vanished. It first left France—where it was aptly called *mal de la misère*—because the government early required the use of potatoes and bread for human food and because by 1910 the landed estates had largely been divided among the peasants. It left northern Italy—as sharecropping disappeared and rural wages rose—more slowly than France, not because the French knew more than the Italian physicians but because the government of Italy did not act to change the food of the polenta-eaters. It dwindled in Spain in the same way. By World War II it had been eliminated in the U.S., at least as a public-health concern.

There is a rare congenital failure to metabolize tryptophan whose victims are endangered by a diet that is not abundant in niacin. Millet-eaters can show pellagra too even though millet (in the Deccan Plateau of India) seems to have as much niacin as rice—on which no one gets pellagra. There are still uncertainties; it may be that the leucine content of the millet interferes with tryptophan metabolism in people who eat marginal amounts of protein.

The endemic pellagra areas on the map are smaller now but still plain. One is among village laborers of upper Egypt, fed on cheap corn brought from the lower delta and lacking the subsidized beans and bread that are sold to poor townsfolk. The second area comprises the Bantu reserves (and an independent enclave) within the Republic of South Africa. The third is among the agricultural laborers living on millet on the central plateau of India, and others not far away who subsist on maize.

"Anyone can cure pellagra with a handful of pills," but the social disease still appears in "the same conditions of misery" first seen among the Asturian farm workers when George Washington was a boy—before Malthus, before the population explosion but not before landlords, high rents and low wages.

These two books are both well illustrated; the leprosy text has striking color photographs, not to be recommended casually. Dr. Roe, a nutritionist and dermatologist at Cornell University, includes in her book strong images of pellagra patients from the time of Casal up to Margaret Bourke-White.

ON ANCIENT CENTRAL-ASIAN TRACKS, by Sir Aurel Stein. The University of Chicago Press ($4.50). THE GREAT CHINESE TRAVELERS, an anthology edited and introduced by Jeannette Mirsky. The University of Chicago Press ($4.50). THE EXHIBITION OF ARCHAEOLOGICAL FINDS OF THE PEOPLE'S REPUBLIC OF CHINA, an illustrated handlist. National Gallery of Art ($2.75). THE EXHIBITION OF ARCHAEOLOGICAL FINDS OF THE PEOPLE'S REPUBLIC OF CHINA, text by the Organization Committee of the Exhibition of Archaeological Finds of the People's Republic of China. National Gallery of Art ($1.50). In 1906 the little column started out for the lonely ruins in the waterless desert. Every available camel was loaded with 400 or 500 handy pounds of the chief necessity: ice (to provide water). Surprises such as ice-cold deserts abound in this charming, modest, somehow serene account of three remarkable expeditions to "innermost Asia" by a scholar-traveler, Sir Aurel Stein of the Indian Educational Service, during the Edwardian years. His book, first published in 1933, complements with a personal narrative his dozen heavy folio reports. He was a tireless "little gnome of a man," a superb Orientalist who knew Sanskrit, colloquial Turki, Early Iranian and much more. He walked and rode 25,000 insistent miles with a few Indian companions (a surveyor, a cook and a handyman), a couple of Turki camel-caravan specialists, a Chinese scholar-secretary ("for a serious study of Chinese ... I regret I never had had needful leisure") and shifting bands of diggers he recruited locally when he needed them.

The main theater of these travels was the Tarim basin tucked north of the mountains fringing the high Tibetan plateau, south of the Tien Shan mountains and west of the snowy Pamirs, where the passes lie at 15,000 feet. Eastward is the outermost section of the Great Wall of China, a flanking portion built in Han times and first excavated on Stein's third trip. The central desert of this basin, the size and shape of California, is still uninhabited; it is perhaps the most barren region of such size in the world, with only one narrow transverse band holding a scant oasis population. Flanking the desert north and south, however, living wherever the snow-fed rivers come down from the mountain frontiers to sink into the rainless plain, a few million oasis dwellers grow grain and fruit under irrigation, and even cotton and grapevines in the lower altitudes, which fall to 500 feet below sea level in hot, arid, fertile Turfan.

Along this chain of oases Chinese travelers to the West have passed for two millenniums. The diplomat Chang Ch'ien, who visited the Huns in the second century B.C. and brought back news of their fine horses (and perhaps the first wine grapes), and the monk Hsüan-tsang, who in the seventh century of our era traveled to India to secure Sanskrit texts then known only in part to Chinese Buddhists, are two of the most famous. Stein studied the accounts of Hsüan-tsang and proclaimed his devotion to that "prince of pilgrims," whose footsteps he followed from India across mountain and desert. Marco Polo came too, and his account is also verifiable: the split track of the Old Silk Road, passing by the desert in the north and the south alike, is rich with the evidence of use.

Stein reclaimed from the dry, preserving sands (the glacier-fed streams have gradually dwindled over the centuries, so that not far beyond the present limits of irrigation there lie many abandoned and drifted-over settlements) many fascinating documents and a rich pictorial treasure of scroll painting and fresco. Here he found the oldest samples of paper known, made from hemp textiles reduced to pulp exactly as the old Chinese texts described paper on its invention in A.D. 105. Eight folded letters in a watch post along the oldest segment of the Great Wall were what he recovered. They were written in a novel script in Early Sogdian, an Iranian language, by Central Asian traders. All around lay Chinese military documents, wood tablets and thin slips of wood "stationery" to which "Chinese conservatism clung."

This region is a high crossroads of the Old World. A single shrine has painted panels with the clear "impress of Hellenistic style" depicting the Chinese princess who first took contraband silkworm eggs out of China. A single document bears two seals side by side, one in Chinese lapidary characters and the other a "portrait head unmistakably cut after Western models."

The old seabed around Lop Nor at the eastern end of the basin, a terrain terrible with hard salt crust in big slanted cakes and pressure ridges, showed Stein a "straight wide track" worn by centuries of the passage of pack animals and probably even carts. Square-holed Han coins and government-issue bronze infantry arrowheads still lay on the salt surface. The find records the sizable traffic maintained for centuries by the Chinese across 120 miles of utterly barren ground for logistical and civil purposes, "an achievement fraught with momentous results for the exchange of civilizations."

When Sir Aurel traveled, the Chinese homeland was disunited and weak. His most famous disclosure to the world, 500 cubic feet of scrolls packed in the treasure recess of a chapel at the cave shrine of the Thousand Buddhas, was a scholar's library find more than it was an archaeologist's: the place was a living cult center, staffed by monks and thronged by pilgrims even though it was far from—and forgotten by—the officials and scholars of distant, preoccupied Peking. No longer, for China is now united and notably conscious of such things. "Let the past serve the present" is the watchword of Chinese archaeologists, who are busy now in the Sinkiang-Uigur Autonomous Region as they are over all China.

The catalogue of the Chinese Exhibition displays, among the yield of a hundred far-flung sites, a few recent finds from along the Old Silk Road. Here are damasks and twills of silk and wool excavated in Turfan and in Min-Feng, a site (Niya) first dug by Stein. There are even two dumplings, or *chiaotzu;* they are still available at your nearest northern-style Chinese restaurant, but these were prepared in the eighth or ninth century. (Stein reported similar grave goods, elaborate pastries preserved by the dryness of the climate.) The wonderfully lively little bronze horse on the front cover of the catalogue of this fine exhibition—the Flying Horse of Kansu—celebrates the "celestial horse," the tall Western strain introduced along the Silk Road about 100 B.C. to supplant the stocky, woolly horses traditional in China. It was unearthed at Wuwei, an interior town on the Silk Road.

These are three complementary books. The Stein work is an accessible paperbound presentation to a new generation of a masterwork, a fascinating volume whether read as travel, as archaeology or as autobiography. The companion anthology supplies an indispensable background to Stein and to Sinkiang (not to overlook the revealing pieces on 15th-century sea voyages to Africa or by the 19th-century Chinese travelers to our Western lands), but it is harder to read by a good deal. The variety of styles, and the antique quality of some of the reports and the stiltedness of others, do not ease the reader's path. The maps in both of these books are adequate but the recent atlas of the People's Republic of China (reviewed here in January, 1973) gives welcome aid. The exhibition catalogue provides a good pictorial introduction to the abundant legacy of ancient China, although it is only a start. It is most heartening, because these objects represent finds of the 25 years of the People's Republic, only a little of all to come: the "rich store" of the hard work of "the forefathers of the Chinese people."

HEAT TRANSFER IN FIRES: THERMOPHYSICS, SOCIAL ASPECTS, ECONOMIC IMPACT, edited by Perry L. Blackshear. Halsted Press Division, John Wiley & Sons ($28.50). At an "idyllic site" in Yugoslavia a year or two ago an international school was convened to review the current scientific basis for the control of fire, both in man-made structures and in forest and field. Nine lecturers from five countries contributed, and this thick typewriter-composed volume reports what they said. They discussed the costs of fire in life and in property, the incidence of fires, the complex geometry of fires, the transfer of heat and of matter that lies at the heart of the event, along with its modeling and its physicochemical basis, and the special problem of the transfer of heat by radiation.

A few simple pages are here for general readers, but the impact of the work is the complexity of the problem, whether viewed socially or technically. It is plain that most of the loss of life is in small fires, confined within the rooms of residences, where the gases and smoke of oxygen-limited combustion bring death. The issue is no minor one: there are a million building fires in the U.S. every year, and one person dies for every 100 fires. (In the United Kingdom the fire death rate is lower by a factor of 10 although the population is only four times smaller.) Remedies are not easy. For example, the British authorities took steps against injury resulting from the ignition of clothing (the U.S. still lags here), and the death rate showed results.

Nearly all "wild fires" involve the combustion of polymers: wood and other plant materials, textiles, plastics and paints. The reaction we see as flame is in the gaseous phase; the first step in burning must be depolymerization of the solid into the flammable radicals of the gas. For cellulose this step is becoming understood: the modified monomer levoglucosan is the main gaseous combustible; flame follows the "tar path," not the char path. Chemically chosen retardants might someday inhibit this step. But wood is not simply cellulose, and there is a lot more to learn.

It is easy to see in some fires a change in phase, a point at which slow burning suddenly speeds up, the temperature goes sharply higher, the rate of spread increases. One explanation is the feedback loop: once the fire is hot enough to cause—say by radiation—the emission of flammable gas from every fuel surface, the fire is increasingly fed with fuel. Flames shoot out of windows as unburned gases escape and ignite when they mix with the oxygen they need. Mass motion of air, turbulence and convection plumes are evidently important: simpler energy-balance equations with plausible scaling laws appear to work only in rather restricted circumstances. The rate at which a single stick burns has been observed to vary with about the inverse three-halves power of its diameter. This empirical result is rationalized by the recognition that it is heat diffusion into the solid interior, where it produces the needed gas release, that is the controlling factor. Yet it is not simply surface temperature that determines the process but a much more complicated surface condition. "The fire itself determines the rate of fuel combustion."

It is characteristic of nature that science can predict fully the motion of a planet or a comet but not as yet the burning of a match. There is a note of progress about this volume but no very loud one. Strategic insights are likely to give practical results long before thorough analysis can succeed.

INDEX OF ADVERTISERS

MARCH 1975

BIBLIOGRAPHY

Readers interested in further reading on the subjects covered by articles in this issue may find the lists below helpful.

AN EXPERIMENT IN WORK SATISFACTION

RESEARCH ON SELF-DEVELOPING FORMS OF ORGANIZATION. Karl-Olof Faxén in *International Economic Association Ravello Conference on Labor Productivity,* August, 1971.

TOWARDS A SOCIAL ECOLOGY: CONTEXTUAL APPRECIATIONS OF THE FUTURE IN THE PRESENT. Fred E. Emery and Eric L. Trist. Plenum Press, 1973.

GREATER INFLUENCE OVER YOUR JOB. Lars E. Björk, Reine Hansson and Peter Hellberg. Swedish Council for Personnel Administration, 1974.

X-RAY-EMITTING DOUBLE STARS

EVOLUTIONARY PROCESSES IN CLOSE BINARY SYSTEMS. B. Paczyński in *Annual Review of Astronomy and Astrophysics,* Vol. 9, pages 183–208; 1971.

OBSERVATIONS OF GALACTIC X-RAY SOURCES. H. Gursky in *Black Holes,* edited by C. DeWitt and B. S. DeWitt. Gordon and Breach, 1973.

PROGRESS IN X-RAY ASTRONOMY. Riccardo Giacconi in *Physics Today,* Vol. 26, No. 5, pages 38–47; May, 1973.

THE BINARY X-RAY STARS—THE OBSERVATIONAL PICTURE. H. Gursky and E. Schreier in *Proceedings of the I.A.U. Symposium No. 67.* D. Reidel Publishing Company, in press.

THE EVOLUTIONARY ORIGIN OF MASSIVE X-RAY BINARIES AND THE TOTAL NUMBER OF MASSIVE STARS WITH A COLLAPSED CLOSE COMPANION IN THE GALAXY. E. P. J. van den Heuvel in *Proceedings of the 16th International Solvay Conference on Physics.* University of Brussels, in press.

INTERACTIVE HUMAN COMMUNICATION

PRELUDE TO 2001: EXPLORATIONS IN HUMAN COMMUNICATION. Alphonse Chapanis in *American Psychologist,* Vol. 26, No. 11, pages 949–961; November, 1971.

STUDIES IN INTERACTIVE COMMUNICATION, I: THE EFFECTS OF FOUR COMMUNICATION MODES ON THE BEHAVIOR OF TEAMS DURING COOPERATIVE

PROBLEM-SOLVING. Alphonse Chapanis, Robert B. Ochsman, Robert N. Parrish and Gerald D. Weeks in *Human Factors,* Vol. 14, No. 6, pages 487-509; December, 1972.

THE COMMUNICATION OF FACTUAL INFORMATION THROUGH VARIOUS CHANNELS. Alphonse Chapanis in *Information Storage and Retrieval,* Vol. 9, No. 4, pages 215–231; April, 1973.

THE EARTH'S MANTLE

ULTRAMAFIC AND RELATED ROCKS. Edited by Peter J. Wyllie. John Wiley & Sons, Inc., 1967.

THE DYNAMIC EARTH: TEXTBOOK IN GEOSCIENCES. Peter J. Wyllie. John Wiley & Sons, Inc., 1971.

CONTINENTS ADRIFT: READINGS FROM *SCIENTIFIC AMERICAN.* Introductions by J. Tuzo Wilson. W. H. Freeman and Company, 1972.

EARTH. Frank Press and Raymond Siever. W. H. Freeman and Company, 1974.

EARTH, MECHANICAL PROPERTIES OF; EARTH, STRUCTURE AND COMPOSITION OF; EARTH AS A PLANET; EARTHQUAKES in *The New Encyclopaedia Britannica: Macropaedia, Vol. 6.* Encyclopaedia Britannica, Inc., 1974.

PLANET EARTH: READINGS FROM *SCIENTIFIC AMERICAN.* Edited by Frank Press and Raymond Siever. W. H. Freeman and Company, 1974.

VISUAL PIGMENTS AND COLOR BLINDNESS

DIAGNOSIS AND GENETICS OF DEFECTIVE COLOR VISION. Hans Kalmus. Pergamon Press, 1965.

VISION AND THE EYE. M. H. Pirenne. Halstead Press, 1967.

LIGHT, COLOUR AND VISION. Yves Le Grand. Chapman and Hall Ltd, London, 1968.

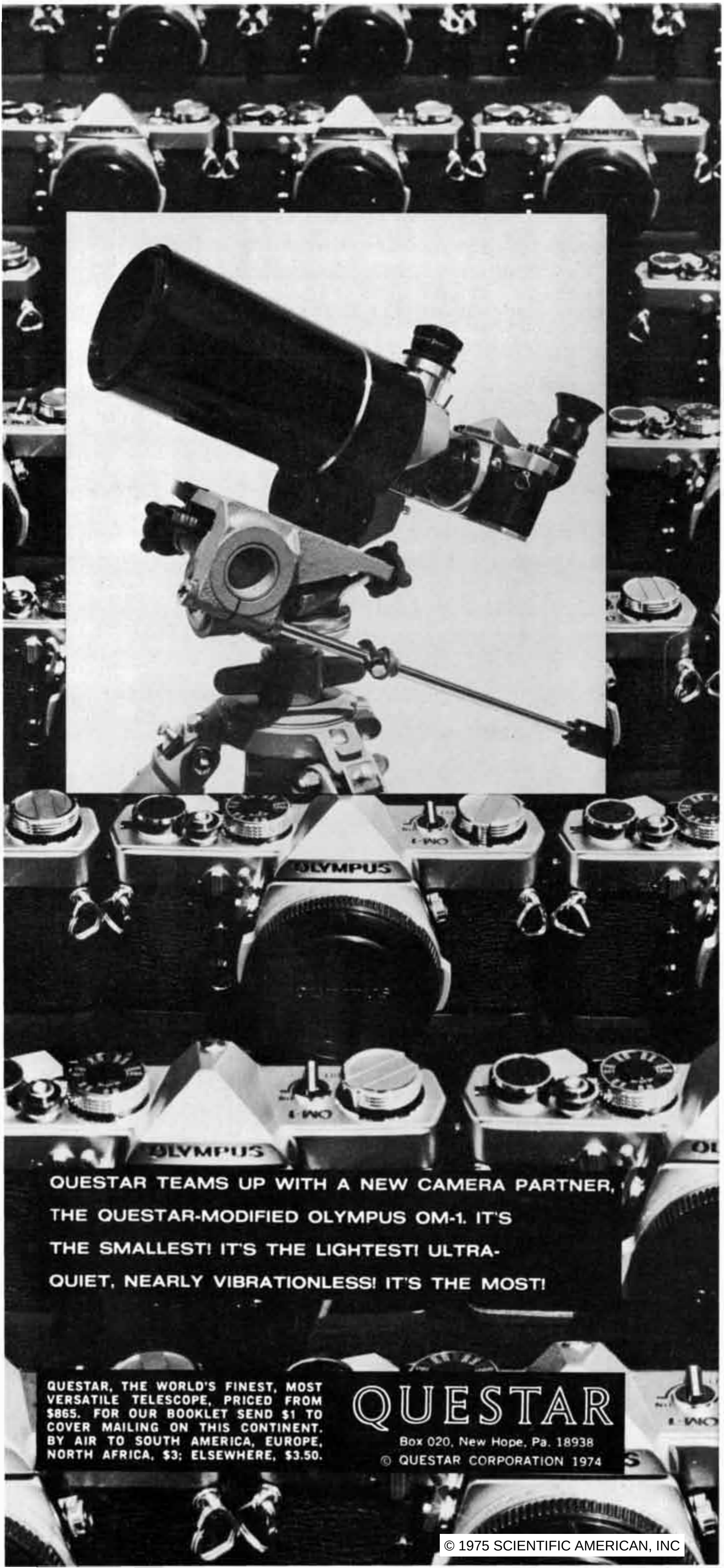
LIGHT AND COLOR. Clarence Rainwater. Western Publishing Company, Inc., 1971.

EYE AND BRAIN: THE PSYCHOLOGY OF SEEING. Richard L. Gregory. McGraw-Hill Book Company, 1973.

THE ROLE OF WAX IN OCEANIC FOOD CHAINS

DISTRIBUTION AND IMPORTANCE OF WAX ESTERS IN MARINE COPEPODS AND OTHER ZOOPLANKTON. Richard F. Lee, Jed Hirota and Arthur M. Barnett in *Deep-Sea Research,* Vol. 18, pages 1147–1165; 1971.

WAX ESTERS: MAJOR MARINE METABOLIC ENERGY SOURCES. A. A. Benson,



R. F. Lee and J. C. Nevenzel in *Current Trends in the Biochemistry of Lipids: Biochemical Society Symposia No. 35,* edited by J. Ganguly and R. M. S. Smellie. Academic Press, 1972.

LIPIDS IN THE MARINE ENVIRONMENT. Richard F. Lee, Jed Hirota, Judd C. Nevenzel, Richard Sauerheber and Allan Lewis in *California Cooperative Oceanic Fisheries Investigations Reports of the California Marine Research Commission,* Vol. 16, pages 95–102; 1972.

WAX IN CORAL MUCUS: ENERGY TRANSFER FROM CORALS TO REEF FISHES. A. A. Benson and L. Muscatine in *Limnology and Oceanography,* Vol. 19, No. 5, pages 810–814; September, 1974.

THE MOST POISONOUS MUSHROOMS

THE MUSHROOM HUNTER'S FIELD GUIDE. Alexander H. Smith. The University of Michigan Press, 1963.

POISONOUS PRINCIPLES OF MUSHROOMS OF THE GENUS AMANITA. Theodor Wieland in *Science,* Vol. 159, pages 946–952; March 1, 1968.

THE MUSHROOM TOXINS. Donald M. Simons in *Delaware Medical Journal,* Vol. 43, No. 7, pages 177–187; July, 1971.

AMANITA PHALLOIDES IN THE EASTERN UNITED STATES. Leo J. Tanghe and Donald M. Simons in *Mycologia,* Vol. 65, No. 1, pages 99–108; January–February, 1973.

GALILEO'S DISCOVERY OF THE PARABOLIC TRAJECTORY

AN EXPERIMENT IN THE HISTORY OF SCIENCE. Thomas B. Settle in *Science,* Vol. 133, No. 3445, pages 19–23; January 6, 1961.

GALILEO STUDIES: PERSONALITY, TRADITION AND REVOLUTION. Stillman Drake. University of Michigan Press, 1970.

TWO NEW SCIENCES. Galileo Galilei, translated by Stillman Drake. University of Wisconsin Press, 1974.

THE AMATEUR SCIENTIST

PRELIMINARY INVESTIGATION OF A PARAGLIDER: LANGLEY RESEARCH CENTER TECHNICAL NOTE D-443. Francis M. Rogallo, J. G. Lowry, D. R. Croom and R. T. Taylor. National Aeronautics and Space Administration, 1960.

I am you.

My name is Masakazu Fukushima, and here I am trumping some neighbors at contract bridge. I play every chance I get. It beats watching television.

Not that I've got anything against watching television, mind you. Watching is fine. What really bothers me is the bulk of the sets themselves. You get a big-screen color TV in your living room and it's like you parked a small car in there. An ugly small car.

The fact is I work for a company that makes a lot of TVs . . . Hitachi. I do electronics research, so I thought I'd try and do something to make color TVs easier to live with.

All people really want from a color TV is a good picture. So we figured the set should be *shaped* like a picture. Flat. To be hung on the wall.

We figured it out, then went into the lab and put it together. We came up with what we wanted . . . the world's first working prototype for a flat-profile color TV. It came out even better than we thought it would.

Not only is this new TV flat, the way it's designed is such that the screen can be just about however big you want it. There could be whole TV walls, say, in schools.

It opens up a lot of new, better possibilities for television, and that makes me happy. By doing myself a favor—by helping to invent something I'd enjoy—I've made something that can benefit an enormous number of people.

I am you.

HITACHI
Resourceful people

Hitachi America, Ltd.
Head Office: 437 Madison Avenue, New York, N.Y. 10022
Tel: (212) 758-5420
Branch Offices: San Francisco, Chicago, Atlanta.

● Power Generation & Transmission Equipment ● Industrial Machinery ● Transportation Equipment ● Communications Equipment ● Electronics & Home Appliances

The legend purrs on.

Few cars ever achieve the stature of a legend. But the Jaguar XJ Series was touched with glory from the beginning. One of its first reviewers found the XJ6: "Uncannily silent, gloriously swift, and safe as houses."

In other words, it was everything the world expected of a Jaguar. Now the XJ legend is advanced by the remarkable XJ6L and the XJ12L.*

Both the XJ6L and the XJ12L handle in the classic Jaguar manner. They perform like sports cars with independent suspension, power disc brakes on all four wheels and the precision of rack and pinion steering.

Naturally both XJ sedans incorporate the expected luxuries of hand-matched hides, cabinet woods, and deep carpeting, enhanced by uncanny silence in motion.

XJ6L or XJ12L? The difference is what powers them.

The XJ12L offers you the innovative short-stroke, overhead-cam, aluminum alloy V-12 engine. In its smoothness and response, it has justly been compared to a turbine. Yet the engine is far from oversized. In fact it displaces 326 cubic inches as opposed to popular American luxury V-8's which displace up to 500 cubic inches.

The Jaguar XJ6L conquers distance with the overhead-cam Six that won a permanent niche in racing history at LeMans. It is smooth and indefatigable, even at European highway cruising speeds.

Both Jaguars are extraordinary. But we suggest you drive them both before you decide.

For the name of the Jaguar Dealer nearest you, call these numbers toll-free: (800) 447-4700, or, in Illinois, (800) 322-4400.

*NOT AVAILABLE IN CALIFORNIA

Jaguar

BRITISH LEYLAND MOTORS INC., LEONIA, N.J. 07605

BRITISH LEYLAND