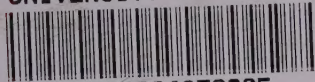


UNIVERSITY OF ARIZONA



39001004872605













*INTERNATIONAL TRACTS IN*  
**COMPUTER SCIENCE AND TECHNOLOGY  
AND THEIR APPLICATION**

GENERAL EDITOR

N. METROPOLIS—*Chicago*    E. PIORE—*New York*    S. ULAM—*Los Alamos*

assisted by an International Honorary Editorial Advisory Board

---

VOLUME 9  
PRINCIPLES OF  
SELF-ORGANIZATION

I\*

1 48735 U/A MAIN: C-BKS  
88 U/A 07/01/02 5063

*Already published in this series:*

A PRACTICAL MANUAL ON THE MONTE CARLO METHOD FOR  
RANDOM WALK PROBLEMS. *By* E. D. CASHWELL and C. J. EVERETT, 1959.

SELF ORGANIZING SYSTEMS. *Edited by* M. C. YOVITS and S. CAMERON,  
1960.

ANNUAL REVIEW IN AUTOMATIC PROGRAMMING. Vol. 1. *Edited*  
*by* R. GOODMAN, 1960.

COMPUTING METHODS AND THE PHASE PROBLEM IN X-RAY  
CRYSTAL ANALYSIS. *Edited by* R. PEPINSKY, J. M. ROBERTSON and  
J. C. SPEAKMAN, 1961.

EXPERIMENTAL CORRELOGRAMS AND FOURIER TRANSFORMS.  
*By* N. F. BARBER, 1961.

ANNUAL REVIEW IN AUTOMATIC PROGRAMMING. Vol. 2. *Edited*  
*by* R. GOODMAN, 1961.

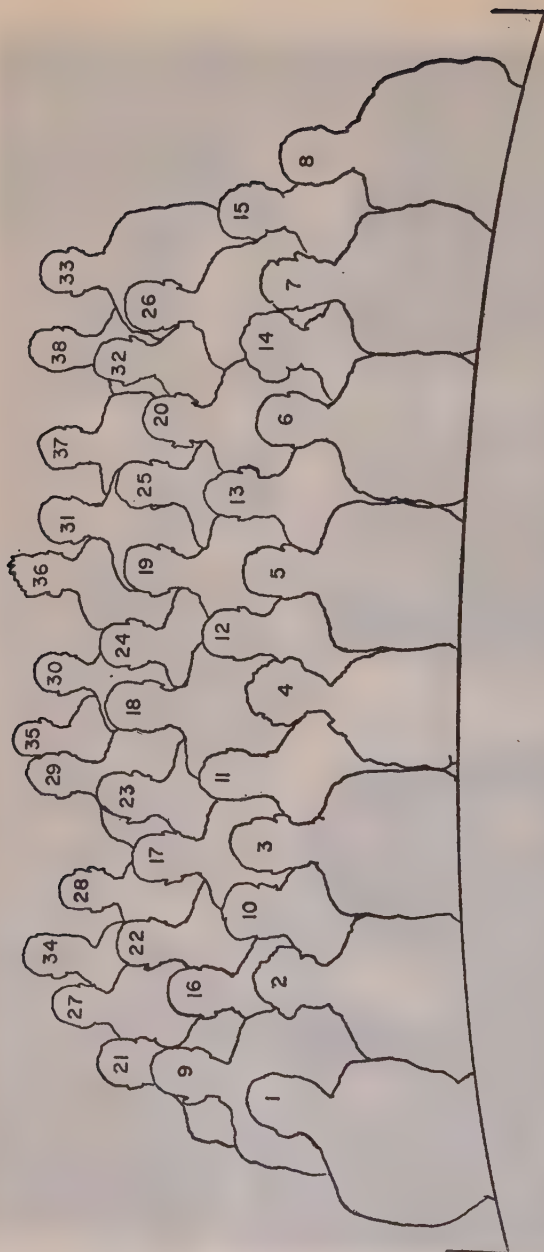
ELECTRONIC DIGITAL COMPUTERS. *By* G. D. SMIRNOV, 1961.











THE PARTICIPANTS AT ALLERTON HOUSE

- 1 Saul Amarel, 2 Gordon Pask, 3 Manuel Blum, 4 Kathy Forbes, 5 Peter Greene, 6 Ross Ashby, 7 Jack Cowan, 8 Heinz Von Foerster, 9 Alfred Inselberg, 10 Ludwig von Bertalanffy, 11 Scott Cameron, 12 Murray Babcock, 13 John Tooley, 14 Cornelia Schaeffer, 15 Stephen Sherwood, 16 George Jacobi, 17 Hans Oestreicher, 18 John Bowman, 19 Jack Steele, 20 Friedrich Hayek, 21 Hewitt Crane, 22 Anatol Rapaport, 23 Raymond Beurle, 24 Jerome Elkind, 25 John Platt, 26 Charles Rosen, 27 Roger Sperry, 28 Frank Rosenblatt, 29 Joseph Hawkins, 30 Albert Novikoff, 31 Stafford Beer, 32 Paul Weston, 33 David Willis, 34 George Zopf, Jr., 35 Albert Mullin, 36 Warren McCulloch, 37 Marshall Yovits, 38 Leo Verbeek

# PRINCIPLES OF SELF-ORGANIZATION

*Transactions of the University of Illinois*

*Symposium on Self-Organization*

*Robert Allerton Park, 8 and 9 June, 1961*

*Sponsored by*

INFORMATION SYSTEMS BRANCH  
U.S. OFFICE OF NAVAL RESEARCH

*Editors*

HEINZ VON FOERSTER

*University of Illinois*

SYMPOSIUM CHAIRMAN

GEORGE W. ZOPF, JR.

*University of Illinois*

SYMPOSIUM SECRETARY

SYMPOSIUM PUBLICATIONS DIVISION

PERGAMON PRESS

NEW YORK · OXFORD · LONDON · PARIS

1962

Q  
325  
455  
1961

PERGAMON PRESS INC.  
122 East 55th Street, New York 22, N.Y.  
1404 New York Avenue N.W., Washington 5 D.C.

PERGAMON PRESS LTD.  
Headington Hill Hall, Oxford  
4 & 5 Fitzroy Square, London, W.1

PERGAMON PRESS S.A.R.L.  
24 Rue des Écoles, Paris V<sup>e</sup>

PERGAMON PRESS G.m.b.H.  
Kaiserstrasse 75, Frankfurt am Main

Copyright © 1962  
PERGAMON PRESS LTD.



LIBRARY OF CONGRESS CARD NUMBER 61-16895

*Printed in Great Britain by J. W. Arrowsmith Ltd., Bristol*



006  
U-8  
1961

## FOREWORD

IN early June, 1960, the Biological Computer Laboratory of the Department of Electrical Engineering at the University of Illinois was privileged to bring together many distinguished persons representing diverse disciplines in a Symposium on the Principles of Self-Organization. It seems particularly appropriate that this interdisciplinary gathering should have been organized under the auspices of electrical engineering, which itself is interdisciplinary in character. For electrical engineering, once concerned chiefly with electric power and electrical communication, has progressed through radio and electronics to develop into the "new electronics" which, in its broadest sense, is concerned with extending man's senses in space, in acuity, in range, and in speed.\* But when the electrical engineer observes that it presently requires a 50 million bit per second channel to transmit a picture which is comprehended adequately by the human viewer at a modest 50 bit per second rate, he becomes acutely conscious of the limitations of existing electronic systems. This awareness of the large disparity between present accomplishment and possible future achievement by more man-like systems, should give the electrical engineer an urgent concern for the subject matter of this Symposium.

EDWARD C. JORDAN  
Head, Dept. of Electrical Engineering,  
University of Illinois,  
Urbana, Illinois.

---

\* W. L. EVERITT, Let us redefine electronics. *Proc. IRE* 40, No. 8, 899 (1952).



## PREFACE

"I am convinced a time will come when the physiologist, the poet, and the philosopher will all speak the same language and mutually understand each other."

CLAUDE BERNARD.

"MUCH has been said and written about specialized topics" Phaidros complained about two thousand five hundred years ago during a social gathering of Agathon's friends who came to celebrate his recent triumph as a poet. "Monographs on minute details in the lives of Hercules or Homer are swamping the market, but who discusses an interdisciplinary problem as, for instance, love?" Thus commenced in Agathon's house the immortal first interdisciplinary symposium which was attended by philosophers, statesmen, playwrights, poets, social scientists, linguists, medical doctors and students of various disciplines.

Plato's account leaves no doubt as to the success of this meeting. Its essential ingredients were the friendship amongst the participants, a stimulating atmosphere and, of course, everybody's interest in the topic.

The problem which was—and is—close to the hearts of the people of the Biological Computer Laboratory of the University of Illinois is a clarification of the problems associated with "self-organization". Is this term perhaps a misnomer for an otherwise trivial process and our concern for self-organization nothing but an expression for fuzzy thinking which constantly overlooks a hole in an apparently closed adiabatic surface? Or is self-organization a useful concept which may help the elucidation of a multitude of problems closely connected with artificial intelligence, mechanization of thought, automation of perception, intelligence amplification, inductive inference machines, cellular organization, growth, evolution, etc?

---

\* This point is discussed in H. VON FOERSTER, *Self-organizing systems and their environments, Self Organizing Systems* (Ed. Yovits, M. C. and Cameron, S.), Pergamon, New York, 1960.

In order to find out about these questions we followed the famous example of antiquity. Allerton's Mansion, 25 miles from the University Campus, replaced Agathon's house, and—under the patronage of the Information Systems Branch of the Office of Naval Research\*—played host to thirty-one guests from all over the world who were friends before they knew each other—because they knew each others' work—who created a stimulating atmosphere and who were at least as interested in our topic as were the men in Agathon's house in theirs.

One of the tangible results of this symposium is the volume you hold in your hands, the collection of twenty-three papers which were presented and discussed during tightly scheduled meetings on the two days of 8 and 9 June 1960.

The other results, tangible as well, cannot be presented between hard covers at the present moment. These results will materialize sooner or later as the offspring of a delightful spirit the adequate description of which we must leave to a Plato. This spirit took hold of all participants whose co-operation, productive criticism, respect for the integrity of the ideas of others and whose enthusiasm extended the informal sessions into the small hours of the next day. Did one of us at four o'clock in the morning finally prove "Bowman's First Theorem": *The number of people shaking hands an odd number of times is even?* Or did somebody come up with the proper extension of the sequence  $a_i < 100$ : 8, 18, 80, 88, 85, . . . which happens to be 84, 89, 81, 87, . . .? Yes, perhaps, if he was prone to think lexicographically. However, the puzzled reader may ask what has this to do with principles of self-organization? The following papers may give an answer to this question. There seems to be almost no field left which is not drawn into the circle of problems associated with self-organization. Some problems appear at first insoluble—as for example the silly sequence above—until they are approached from an entirely different angle.

There exist  $23!$  different ways in which the collected papers of this volume can be arranged; and there are many possible choices of arrangement which could claim to be the best. We finally chose to present them in the sequence in which they appeared during the meeting. Since the symposium organized itself while

---

\* Contract Nonr 1834(21).

under way, this presentation seems to maintain at least some of the natural flow of ideas. However, a guide who is to lead the reader through the jungle of papers may stake out five areas of interest and may point out the speakers who illuminated these areas with their contributions.

This guided tour would go as follows:

#### I. Theoretical and Experimental Foundations of Self-Organization

- |             |           |
|-------------|-----------|
| 1. Ashby    | 5. Bowman |
| 2. Sperry   | 6. Pask   |
| 3. Beurle   | 7. Willis |
| 4. Rapoport | 8. Rosen  |

#### II. Theories of the Behavior of Complex Systems

- |           |               |
|-----------|---------------|
| 1. Beer   | 3. Rosenblatt |
| 2. Amarel | 4. Zopf       |

#### III. Immunology of Self Organizing Systems

- |              |            |
|--------------|------------|
| 1. McCulloch | 4. Cowan   |
| 2. Blum      | 5. Löfgren |
| 3. Verbeek   |            |

#### IV. Preorganization in Cognitive Systems

- |            |             |
|------------|-------------|
| 1. Platt   | 3. Greene   |
| 2. Shimbel | 4. Novikoff |

#### V. Componentry of Self Organizing Systems

- |          |           |
|----------|-----------|
| 1. Crane | 2. Tooley |
|----------|-----------|

The eight papers in the first group may be regarded as a concise introduction into the epistemology of self-organization. The speakers develop the fundamental concepts dealing with such processes and explicitly describe systems in which we recognize self-organization.

In the second group, Stafford Beer opens the discussion on the analysis of complex systems in general by appropriately presenting an electroencephalogram of one of Britain's largest steel mills. A variety of strategies in the approach to the study of complex systems are presented in the four papers of this group.

Probably the most tightly correlated papers are collected in Group III. These five papers could easily go as a monograph on



the fascinating question: what are the structures which make complex systems immune against errors of all kinds? As an extra bonus the reader may find a lucid exposé of multivalued logics (Cowan). That this topic should come up in a meeting on principles of self-organization will be clearly seen, if one follows the arguments steadily developed in the preceding papers.

Evolution and experience are ordering processes on different levels. Evolution provides the structures on which experience can grow. The four papers in Group IV cover the problems of pre-organization in self-organizing and cognitive systems from highly specialized examples to the most generalized theorems. If time during the meeting, or space in this volume had been more abundant, some of our work\* in differential geometry for extracting invariants in a set of stimuli would have served as complement to Novikoff's work in integral geometry.

Although only two papers deal explicitly with componentry and technology of self-organizing systems (Group V), we believe that the two papers fully justify their existence in a group by themselves. Their contents and promise in the art of artificial neural nets are far-reaching.

It is now upon us to express our thanks to all who came and so generously gave their time, their energy and their good spirits to this symposium, particularly to Stafford Beer, John Bowman, F. A. Hayek, Warren McCulloch and Anatol Rapoport who served as chairmen to the various sessions, a task which seemed hopeless in prospect and appears incredible in retrospect, if one considers the sheer quantity of information processed in those 36 hours.

We are obliged to the personnel of Allerton House and to the numerous helpers of the Department of Electrical Engineering who moved behind the scene to provide food, transportation, boards, records, programs, books, slides, reprints, drawings, refreshments, etc. In her secretarial powers Kathy Forbes is to be admired for having successfully arranged within two hours more

---

\* M. L. BABCOCK, A. INSELBERG, L. LÖFGREN, H. VON FOERSTER, P. WESTON, and G. W. ZOPF, Jr.: Some principles of preorganization in self organizing systems, *Tech. Rep.* No. 2, Contract Nonr 1834(21). Electrical Engineering Research Laboratory, Engineering Experiment Station, University of Illinois, Urbana, Illinois, 1960.

than twenty last minute flight reservations on a dozen or so national and international airlines.

Our thanks go to Miss Cornelia Schaffer whose presence during the meeting was catalysis and whose advice in preparing these transactions was synthesis. Gratefully acknowledged is the co-operation of the publishers of this volume in all matters of organization, who endured with patience the slow trickle of parts of the manuscript and gave this volume its pleasing appearance.

It is impossible to conclude this preface without giving credit to the man who made this meeting possible: Dr. Marshall C. Yovits of the Information Systems Branch of Office of Naval Research. Furthermore, he not only mobilized the never tiring wings of MATS to fly our transatlantic friends to Illinois, but his benevolent spirit was present throughout this symposium.

H. V. F.



## CONTENTS

	PAGE
List of Participants	xv
Some Self-Organizing Parameters in Three-Person Groups <i>A. Rapoport</i>	1-
Toward the Cybernetic Factory <i>S. Beer</i>	25
Symbolic Representation of the Neuron as an Unreliable Logical Function <i>W. S. McCulloch</i>	91
Properties of a Neuron with Many Inputs <i>M. Blum</i>	95
On Error Minimizing Neural Nets <i>L. Verbeek</i>	121-
Many Valued Logics and Reliable Automata <i>J. Cowan</i>	135
Limits for Automatic Error Correction <i>L. Löfgren</i>	181
A Proposed Evolutionary Model <i>G. Pask</i>	229-
Principles of the Self-Organizing System <i>W. R. Ashby</i>	255
Orderly Function with Disorderly Structure <i>R. W. Sperry</i>	279
Functional Organization in Random Networks <i>R. L. Beurle</i>	291-

How a Random Array of Cells can Learn to Tell Whether a Straight Line is Straight <i>J. R. Platt</i>	315
Attitude and Context <i>G. W. Zopf, Jr.</i>	325
Integral Geometry. An Approach to the Problem of Abstraction <i>A. Novikoff</i>	347
The Functional Domain of Complex Systems <i>D. G. Willis</i>	369
Strategic Approaches to the Study of Brain Models <i>F. Rosenblatt</i>	385-
The Neuristor <i>H. D. Crane</i>	403
A Transmission Line Leading to Self-Stabilizing Systems <i>J. R. Bowman</i>	417
An Approach to a Distributed Memory <i>C. A. Rosen</i>	425
An Approach to Automatic Theory Formation <i>S. Amarel</i>	443
Networks which Realize a Model for Information Repre- sentation <i>P. H. Greene</i>	485-
Thresholding and Micro-Miniaturization with Semicon- ductors <i>J. Tooley</i>	511
A Logical Program for the Stimulation of Visual Pattern Recognition <i>A. Shimbel</i>	521
Index	527



## LIST OF PARTICIPANTS

- \*Dr. SAUL AMAREL,  
David Sarnoff Research Center, Radio Corporation of  
America, Princeton, New Jersey.
- \*Dr. W. ROSS ASHBY,  
300c Electrical Engineering Research Laboratory, Univer-  
sity of Illinois, Urbana, Illinois.
- Dr. MURRAY L. BABCOCK,  
111 Electrical Engineering Research Laboratory, Univer-  
sity of Illinois, Urbana, Illinois.
- \*Mr. STAFFORD BEER,  
United Steel Companies, Ltd., Cybor House, 1 Tapton  
House Road, Sheffield 10, England.
- \*Dr. R. L. BEURLE,  
English Electric Valve Co., Ltd., Waterhouse Lane, Chelms-  
ford, Essex, England.
- \*Mr. MANUEL BLUM,  
c/o W. S. McCulloch, Rm 26-027, Massachusetts Insti-  
tute of Technology, Cambridge 39, Massachusetts.
- \*Dr. J. R. BOWMAN,  
Technological Institute, Northwestern University, Evans-  
ton, Illinois.
- Mr. SCOTT CAMERON,  
Armour Research Foundation, 10 W. 35th Street, Chicago  
16, Illinois.
- \*Mr. JACK COWAN,  
c/o W. S. McCulloch, Rm. 26-027, Massachusetts Insti-  
tute of Technology, Cambridge 39, Massachusetts.

- \*Dr. H. D. CRANE,  
Division of Engineering Research, Stanford Research  
Institute, Menlo Park, California.
- Dr. J. I. ELKIND,  
Bolt, Beranek and Newman, Inc., 50 Moulton Street,  
Cambridge 39, Massachusetts.
- \*Dr. PETER H. GREENE,  
Committee on Mathematical Biology, University of  
Chicago, 5741 Drexel Avenue, Chicago 37, Illinois.
- \*Mr. J. K. HAWKINS,  
Aeronutronic Division, Ford Motor Company, Newport  
Beach, California.
- \*Dr. F. A. HAYEK,  
Committee on Social Thought, University of Chicago,  
Chicago, Illinois.
- Mr. ALFRED INSELBERG,  
107 Electrical Engineering Research Laboratory, Uni-  
versity of Illinois, Urbana, Illinois.
- Dr. GEORGE T. JACOBI,  
Armour Research Foundation, 10 W. 35th Street, Chicago  
16, Illinois.
- \*Dr. LARS LÖFGREN,  
107 Electrical Engineering Research Laboratory, Uni-  
versity of Illinois, Urbana, Illinois.
- \*Dr. WARREN S. McCULLOCH,  
Room 26-027, Massachusetts Institute of Technology,  
Cambridge 39, Massachusetts.
- \*Mr. A. A. MULLIN,  
108d Electrical Engineering Research Laboratory, Univer-  
sity of Illinois, Urbana, Illinois.
- \*Dr. ALBERT B. J. NOVIKOFF,  
Division of Engineering Research, Stanford Research  
Institute, Menlo Park, California.

- DR. HANS OESTREICHER,  
Bioacoustics Laboratory, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio.
- \*Mr. GORDON A. PASK,  
Systems Research Ltd., 20 Hill Rise, Richmond, Surrey, England.
- \*Dr. JOHN R. PLATT,  
Department of Physics, University of Chicago, Chicago 37, Illinois.
- \*Dr. ANATOL RAPOPORT,  
Mental Health Research Institute, University of Michigan, Ann Arbor, Michigan.
- \*Dr. CHARLES A. ROSEN,  
Division of Engineering Research, Stanford Research Institute, Menlo Park, California.
- \*Dr. FRANK ROSENBLATT,  
Hollister Hall, Cornell University, Ithaca, New York.
- Miss CORNELIA SCHAEFFER,  
ATHENEUM Publishers, 162 E. 38th Street, New York 16, New York.
- \*Dr. STEPHEN SHERWOOD,  
Illinois State Psychiatric Institute, 1601 W. Taylor Street, Chicago 12, Illinois.
- †Dr. A. SHIMBEL,  
Illinois State Psychiatric Institute, 1601 W. Taylor Street, Chicago 12, Illinois.
- \*Dr. R. W. SPERRY,  
Division of Biology, California Institute of Technology, Pasadena, California.
- Dr. J. E. STEELE,  
Bioacoustics Laboratory, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio.

- \*Mr. JOHN R. TOOLEY,  
Central Research Laboratory, Texas Instruments, Inc.,  
13500 N. Central Expressway, Dallas, Texas.
- \*Dr. L. A. M. VERBEEK,  
c/o W. S. McCulloch, Rm. 26-027, Massachusetts Insti-  
tute of Technology, Cambridge 39, Massachusetts.
- Dr. LUDWIG VON BERTALANFFY,  
3211 Sena Drive, Topeka, Kansas.
- \*Dr. HEINZ VON FOERSTER,  
215 Electrical Engineering Research Laboratory, Univer-  
sity of Illinois, Urbana, Illinois.
- Mr. PAUL WESTON,  
108h Electrical Engineering Research Laboratory, Uni-  
versity of Illinois, Urbana, Illinois.
- \*Dr. DAVID G. WILLIS,  
Missile and Space Division, Lockheed Aircraft Corpora-  
tion, Sunnydale, California.
- \*Dr. MARSHALL YOVITS,  
Code 427, Physical Sciences Division, Office of Naval  
Research, Washington 25, D.C.
- \*Mr. GEORGE W. ZOPF, JR.,  
300c Electrical Engineering Research Laboratory, Uni-  
versity of Illinois, Urbana, Illinois.

---

\*Gave paper or participated in discussion.

†Did not attend Symposium.

## **ANATOL RAPOPORT**

*The University of Michigan*

# **SOME SELF-ORGANIZATION PARAMETERS IN THREE-PERSON GROUPS\***

### **INTRODUCTION**

We shall describe two experiments, in each of which a three-person group is the subject. In each experiment the performance of the group is measurable in terms of certain parameters. Some of the parameters can be taken as indices of the extent to which the group has organized itself, so as to perform the task more efficiently (in the first experiment) or so as to assure the largest gains to all the members (in the second experiment). The singling out of such "parameters of self-organization" was one of the goals of the experiments. Once they have been singled out, other experiments suggest themselves designed for studying the dependence of these parameters on experimental conditions, on the composition of the groups and on other controllable or observable variables.

### **A ROTE LEARNING TASK**

In our first experiment the three-person group was faced with the task of learning how to respond correctly in a coordinated manner to each of a repertoire of stimuli given by the experimenter in fairly rapid succession (every few seconds). The task was so designed that it was possible for the members of the group to divide the memory load among them, provided they gave each other certain information in the process of learning. If such

---

\* This research was supported by the United States Air Force under Contract No. AF33 (616)-6096, monitored by Aero Medical Laboratory, Directorate of Research, Wright Air Development Center, Wright-Patterson Air Force Base, Ohio.

information were not given at all, it would still be possible for the group to learn the task but only at the cost of each member's storing the entire information. That is to say, in this case, the overlap in the information stored by each member would be maximal. If, on the other hand, all the necessary information were freely flowing among the members, the overlap in the memory loads could be zero: each member could store one-third of the total information in the task. The actual overlap or rather the degree of non-overlap in the information stored becomes then an index of efficacy of intra-group communication for reducing the individual memory loads or, as we may put it, a measure of the self-organizing potential of the group.

The apparatus used in the rote learning experiment is shown in Fig. 1, below.

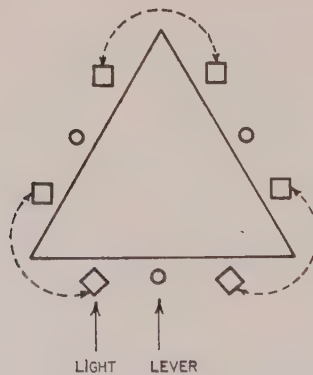


FIG. 1

The three group members are seated around a triangular table, each facing a panel, on which there are two lights (the squares in the diagram) and a lever between them (the circle in the diagram). Initially all the six lights are on. The group members must extinguish them by moving the levers, each of which can be moved right or left from the central position, to which it returns when released.

The dotted lines connecting pairs of lights indicate that the connected pairs are always extinguished together. Actually, to extinguish each pair, it is necessary to move the two adjoining



levers towards the pair. For example, if subject *A* moves his lever to the left and if, while this lever is still inclined to the left, subject *B* turns his lever to his right, the pair of lights between *A* and *B* will be extinguished.

It follows that depending on the order in which the various levers are moved, the three pairs of lights will be extinguished in a given order. There are six possible orders of extinguishing the three pairs of lights. Call these six orders *A, B, ..., F*. A particular assignment of a permutation of the six digits, 1, 2, ..., 6, to these six orders constitutes a problem. Each digit is called a "target number". The group has learned a problem if in response to each digit called out by the experimenter, the group extinguished the lights in the order associated with that target number. The feedback in the learning process consists only of verdicts "Right" and "Wrong" given by the experimenter after each response.

Since there are 6! permutations of six digits, there are 720 possible problems. A selection of a certain number of such problems (in our experiment eight) constitute a task. A task, therefore, is learned when the group has learned to respond correctly (i.e. extinguish the lights in the prescribed order) to every target number in every problem included in the task.

In our experiment, the mechanics of the apparatus were explained to the subjects in a practice session, so that the task to be learned was simply the correct association of properly coordinated movements of the levers to the corresponding target numbers.

Our data are recorded as a cumulated error curve, i.e. the cumulated erroneous responses are plotted as ordinates against stimulus presentations as abscissae. Obviously all such curves will be monotone increasing and, since errors practically disappear as learning goes to completion, the curves tend to horizontal asymptotes.\*

A theoretical curve, derived from a mathematical model to be presently described, was fitted to the data by adjusting two free parameters. One of these could be interpreted as an index of the rate of storing information while the other could be interpreted as an index of efficiency of intra-group communication (efficiency

---

\* In practice, a small residual accidental error rate was observed in most cases and was corrected for.

in the sense of reducing the individual memory loads). The following definitions and discussion should elucidate the meaning of these interpretations.

In a particular problem assume that any permutation of the six digits 1, 2, ..., 6 is associated with *a priori* equal probability with the six orders of extinguishing the lights, *A, B, ..., F*. Then the learning of the problem involves the selection of one of the 720 permutations. The *a priori* probability of such selection being 1/720, the initial uncertainty associated with a problem is  $\log_2 720 = 9.5$  bits. The problem being independently selected, the total uncertainty of the task is  $8 \times 9.5 = 76$  bits. Call this theoretical uncertainty  $H^*$ . When the task has been completely learned, the response to each target number is certain. Hence, the final uncertainty is zero, and so  $H^* = 76$  bits of information has been presumably "stored" in the group.

Now at some intermediate stage of the learning process, there is still some residual uncertainty in the responses, which is less than  $H^*$  (since correct responses are now made with frequency greater than chance) but greater than zero (since errors are still occurring). Call this residual uncertainty  $U(t)$ , where  $t$  is the number of stimulus presentations that have occurred at that stage. In general  $U(t)$  is a monotone decreasing function of  $t$ . Hence

$$h(t) = H^* - U(t) \quad (1)$$

must be the amount of information stored by the group at "time"  $t$ . (Time is here measured by the number of stimulus presentations).

Our mathematical model now consists of two assumptions:

1. The rate of gain of information per error is constant.
2. At all times, residual uncertainty is equi-distributed among all the target numbers.

Let  $w(t)$  be the total number of errors cumulated at time  $t$ . Our first assumption, then, says:

$$dh/dw = k \quad (\text{a constant}). \quad (2)$$

Now if  $w(t)$  is the total number of errors, then  $dw/dt$  can be taken as the probability of error at time  $t$ . Therefore  $1 - dw/dt$  is the probability of a correct response at time  $t$ . Our second

assumption, then, says

$$-U(t) = M \log_e(1 - dw/dt), \quad (3)$$

where  $M$  is the number of different stimuli in the task. We have changed our base of logarithms to the natural exponential base for convenience of performing the operations of the calculus. Our units of information will then become "nits" ("natural bits": 1 bit =  $\log_e 2$  nits).

*Remark.* The justification for equation (3) is our assumption that residual uncertainty remains equally distributed among all the  $M$  stimuli. When the responses are examined in detail, this assumption is seen to be clearly false. Some correct responses are fixed early in the learning process. Hence these responses are certain, and we cannot maintain that the residual uncertainty is equally distributed. The assumption becomes increasingly more justified, however, when performances of several groups are averaged, because different groups will generally fixate the learned responses in different orders. Moreover (3) is at any rate a fair approximation to the actual residual uncertainty of response, within considerable fluctuations of probabilities of correct response among the individual stimuli. In our data performances of several groups will be averaged, and we hope that this gives sufficient justification for our assumption 2.

Integrating (2) and imposing the initial condition  $w(0) = 0$ , we have

$$h = kw. \quad (4)$$

Combining (1), (3) and (4), we obtain

$$kw = H^* + M \log_e(1 - dw/dt). \quad (5)$$

Taking exponentials and rearranging, we have

$$\exp\{(kw - H^*)/M\} = 1 - dw/dt. \quad (6)$$

This differential equation can be easily integrated. When the initial condition  $w(0) = 0$  is imposed, we finally obtain

$$w = \frac{H}{k} - \frac{M}{k} \log_e[(e^{H^*/M} - 1)e^{-kt/M} + 1] \quad (7)$$

which represents our theoretical cumulated error curve.

Of the parameters in equation (7),  $M$ , the total number of distinct stimuli, is under experimental control (in our case  $6 \times 8 = 48$ );  $H^*$ , the total uncertainty, as we have seen, is theoretically computed. Therefore only  $k$  appears as a free parameter, to be adjusted in fitting the theoretical curve to the data. We have said, however, that we have two free parameters at our disposal. We get the second when we treat the total uncertainty as a free parameter to be adjusted. This is done in view of the following considerations.

Suppose we try to fit the curve determined by equation (7) to a set of data, i.e. a cumulated error curve. We have fixed the initial point at  $w(0) = 0$ , which, of course, corresponds to the data. Observe now that from (7) we get

$$w(\infty) \equiv W = H^*/k. \quad (8)$$

Hence the simplest estimate of  $k$  is from  $H^*/W$ , where  $H^*$  has been theoretically calculated and  $W$  has been observed as the final number of errors. Hence the asymptote of the theoretical curve will coincide with that of the experimental one. It remains to see how the remaining points fit the theoretical curve. Disregarding for the moment the case where the intermediate experimental points may fall both above and below the theoretical curve, we will derive the consequences from the following hypothetical cases.

Case 1. The experimental points all fall below the theoretical curve.

Case 2. The experimental points all fall above the theoretical curve.

Assume now that the total uncertainty of the task is not the fixed, theoretically calculated  $H^*$ , but a parameter  $H$  to be adjusted. It is easily shown from equation (7) that

$$\left( \frac{\partial w}{\partial HW} \right) \geq 0. \quad (9)$$

That is to say, if the asymptotic value  $W$  is held fixed and a greater value of  $H$  is assumed, the entire theoretical curve will be shifted upward and, of course, vice versa. Therefore by assuming for the total uncertainty values of  $H$  larger or smaller than the theoretically calculated value  $H^*$ , we could "tune" the theoretical curve so

as to bring it into a closer fit to the data, provided the experimental points all lie either above or below the theoretical curve.

But let us now interpret the meaning of taking a value of  $H$  smaller or larger than  $H^*$ . Clearly taking a smaller value of  $H$  than that calculated theoretically cannot be permitted, since the theoretically calculated value is the *uncertainty inherent in the task*, i.e. the very minimum quantity of uncertainty that can be assumed. It follows, therefore, that if the experimental points fall below the theoretical curve determined by this minimal  $H^*$ , our model is definitely refuted.

It is otherwise if the experimental points lie *above* the theoretical curve. Taking a larger value of  $H$  does not contradict our model. We may interpret this larger value as indicating that the subjects were not aware of all the redundancies in the task and so had to store more information than the inherent uncertainty  $H^*$ . Or we may interpret the larger value of  $H$  as follows: *Each member has stored on the average more than one-third of the total information; hence the total information stored exceeds the total uncertainty inherent in the task.*

We shall assume the latter interpretation. As we shall see, this excess stored information represents a failure to utilize completely the intra-group communication process and so to reduce the individual memory load to a minimum.

To get a measure which can be interpreted as an efficiency, we define

$$E \equiv \frac{3H^* - H}{2H^*}. \quad (10)$$

Thus when  $H$  has its minimal value,  $H^*$ ,  $E = 1$ , and when  $H$  is  $3H^*$ , i.e. when every group member has stored the entire uncertainty  $H^*$ ,  $E = 0$ . For all intermediate values of  $H$ ,  $E$  is a fraction which can be regarded as an efficiency index.

We shall first show that each individual member, in responding to the target numbers correctly, transmits (to an observer) as much information as the entire group. That is, from the movement of levers by a single individual producing a correct response the movement levers of the others can be uniquely deduced. Consider the schematic representation of the sequence of lever



movements which extinguish the lights in a certain order, associated with a given target number in a given problem (Fig. 2).

The three columns in this matrix represent three time periods. The three group members are represented by rows. The entries *L* and *R* represent right and left movements of the levers. The entry *X* means that in the corresponding time period the position of the

	1	2	3
<i>A</i>	<i>L</i>	<i>R</i>	<i>X</i>
<i>B</i>	<i>R</i>	<i>X</i>	<i>L</i>
<i>C</i>	<i>X</i>	<i>L</i>	<i>R</i>

FIG. 2.

corresponding lever is immaterial (since both of the other members are engaged in extinguishing the light between them). In the example given, the light between *A* and *B* is extinguished first, then the light between *A* and *C*, finally the light between *B* and *C*.

Note, however, that each player's pattern of movements is unique for each target number. There are, in fact, six such patterns, namely (*L*, *R*, *X*); (*L*, *X*, *R*); (*X*, *L*, *R*); (*R*, *L*, *X*); (*R*, *X*, *L*); (*X*, *R*, *L*).

It follows that given a problem, the motion pattern of each single player represents uniquely the target number called for and that therefore each member transmits the entire information.

It is by no means true, however, that each group member must store the entire information in his own brain. The amount he must store will depend on the scheme the group uses to memorize the motions. Suppose the following scheme is used. Each member remembers the target numbers out of the six in each problem on which he must move the lever both in the first and in the second time periods. By symmetry, each group member has two such target numbers to remember in each problem. Two things can be chosen out of six in fifteen ways. Therefore  $\log_2 15 = 3.91$  bits per problem must be stored by each group member using this scheme.



In addition he must remember on which of these two target numbers he moves to the left (or to the right) first. This is one more bit of 4.91 bits in all. Storing these 4.91 bits per problem is sufficient, since this information enables the member in question to signal to the partner with whom he is to move first, conveying also the direction of motion to a second member and also the fact that the third member has nothing to do in the first time period. Thus the moves of all the members follow *logically* and need not be memorized.

By this system, then, 4.91 bits must be stored in each problem by each member. This is considerably less than 9.5 bits inherent in the problem but considerably more than  $(9.5)/3 = 3.2$  bits which would need to be stored if there were no overlap between the amounts stored by the group members.

Let us now examine another scheme, leading to a somewhat more efficient division of the memory load. Suppose each member simply remembers the three target numbers out of the six on which his first move is to the left (or right). This is all that he needs to remember, since the shift of the lever is made to the other side when the corresponding light has been extinguished. Thus the shift of the lever does not depend on stored information. The signal when to shift is transmitted by the members via the apparatus.

Since three things can be selected from six in twenty ways, the amount stored will be  $\log_2 20 = 4.32$  bits. This is less than 4.91 required by the previous scheme but still more than 3.32. Note that in the scheme just described the direction of the first move requires no information stored, since the three target numbers to be remembered are those on which the first move is *always* left (or right).

We see, then, that the amount of information stored by each member may range from  $H^*/3$  to  $H^*$ , even if all the redundancies in the task as a whole are utilized. The excess of total information stored (in case each individual stores more than  $H^*/3$ ) is a consequence of failing to divide the memory load in the most efficient manner.

*Remark.* In our theory no provision is made for a leakage of stored information, which would be analogous to noise in a channel. Therefore we can equate efficiency with maximum compression. Obviously if leakage does exist, overlaps in the division

of memory load may be quite useful, for in that case, the group members can to some extent give each other assistance when memory lapses occur.

We pass to the results of the experiment. The schedule of stimulus presentations was as follows. In each problem, each of the six target numbers was presented three times in succession, and the six numbers in the natural order, thus: 1, 1, 1, 2, 2, 2, ..., 6, 6, 6, etc. through each of the eight problems. Following each response, the experimenter (who saw the target number actually produced by the group recorded on his own panel, not visible to the subjects) announced the verdict "Right" or "Wrong". After the target numbers of the eighth problem have been presented, those of the first problem were again in order. Thus the process went in cycles. Each cycle contained  $6 \times 3 \times 8 = 144$  stimulus presentations and took about 30–60 min. The performance of a typical group lasted 6–10 cycles. Our experimental points were the numbers of errors cumulated at the end of each cycle.

We will call the three responses to each of the three repeated presentations of a target number respectively the *a*-responses, the *b*-responses and the *c*-responses. We can plot the cumulated errors in each of these types of responses alone or in combination. For the purposes of our analysis, we shall examine the curves consisting only of errors in the *a*-responses (the *a*-curves) and the curves consisting of the errors in all the responses (the gross curves).

Now the theory described above does not take into consideration any peculiarities of the stimulus presentation schedule (it is an extremely gross theory). We would expect, then, if the theory has general validity, that each of the different cumulated error curves will be fitted by an equation of the type given by (7). Of course, the values of the parameters adjusted to fit the theoretical curve to the data, will be different in each case. The value of the parameter *k*, for example, representing the rate of information storage, should depend on the feedback received in the learning process. In the *a*-responses, the feedback reflected in the record is the verdict "Right" or "Wrong" received on *three* trials, because each *a*-response follows three different "trial and error" responses to the same target number. But in the gross curves, where all the responses are recorded, the feedback reflected is the verdict "Right" or "Wrong" received on only *one* trial. Therefore the value of *k*

associated with an  $a$ -curve of a given group or a population of groups ought to be larger than the corresponding value associated with the gross curves.

With respect to the parameter  $E$ , given by equation (10) we would expect that its value associated with the  $a$ -curve would be smaller than that associated with the combined curve because the  $a$ -curve represents the responses on only the first of three identical stimulus presentations, i.e. responses in which intra-group communication has not been utilized for, say, correcting wrong responses. Thus we expect a relative excess of  $k$  in the  $a$ -curve of a group or in the average  $a$ -curve of a population of groups performing under given conditions, and a relative excess of  $E$  in the gross curve.

Denoting the parameters associated with the  $a$ -curves by appropriate subscripts we define

$$k^* \equiv \frac{k_a - k}{k_a} \quad (11)$$

and

$$E^* \equiv \frac{E - E_a}{E}. \quad (12)$$

The first of these indices can be regarded as a measure of the extent to which the group as a whole utilizes the feedback in the learning process (the experimenters' verdicts "Right" and "Wrong") to correct errors. The second of these indices can be regarded as a measure of the extent to which the group utilizes intra-group communication to divide the total memory load efficiently.

Table 1 and Fig. 3 show sample data from some experiments conducted at the Mental Health Research Institute, University of Michigan.

### THREE PERSON NON-ZERO-SUM GAMES

In our second experiment, the members of three-person groups were engaged in a sequence of three-person non-zero-sum games of a type in which all three members win if their choices of strategy are properly coordinated. Built into the pay-off structure of each

of the games (with one exception to be noted) was a "temptation" to defect from the coordinated strategy choice. The situation is best illustrated in the so-called Prisoner's Dilemma type of two-person non-zero-sum game. In this class of games each player has a choice of two strategies, which can be called cooperative

TABLE 1. COMPARISON OF THEORY WITH EXPERIMENT  
Average performance of five groups memory load appears divided with maximum efficiency in the gross curve; with  $E_a = 0.83$  in the *a*-curve

Cycle	1	2	3	4	5	6	7
Observed errors	76	138	180	199	211	220	223
Theory: $k = 0.23$ $E = 1.00$	84	142	179	200	211	216	218

Gross curve

Cycle	1	2	3	4	5	6	7
Observed errors	35	62	84	94	101	107	110
Theory: $k_a = 0.64$ $E_a = 0.83$	34	61	80	93	100	106	108

*a*-curve

and competitive respectively. The characteristic feature of these games is that *one* defector from the cooperative solution is rewarded (compared to his cooperative gains), but if *both* defect, both are punished.

Generalizing this idea to three-person games, we have the following variants:

Type 1. One defector is rewarded; two are rewarded.

Type 2. One defector is rewarded; two are punished.

Type 3. One defector is punished; two are rewarded.

In all cases at least one case of defection must be rewarded, and

all three defectors must, of course, be punished, to preserve the chief feature of the "dilemma".

There is, however, another interesting class of games in which unanimous "defection", like unanimous cooperation, is rewarded.

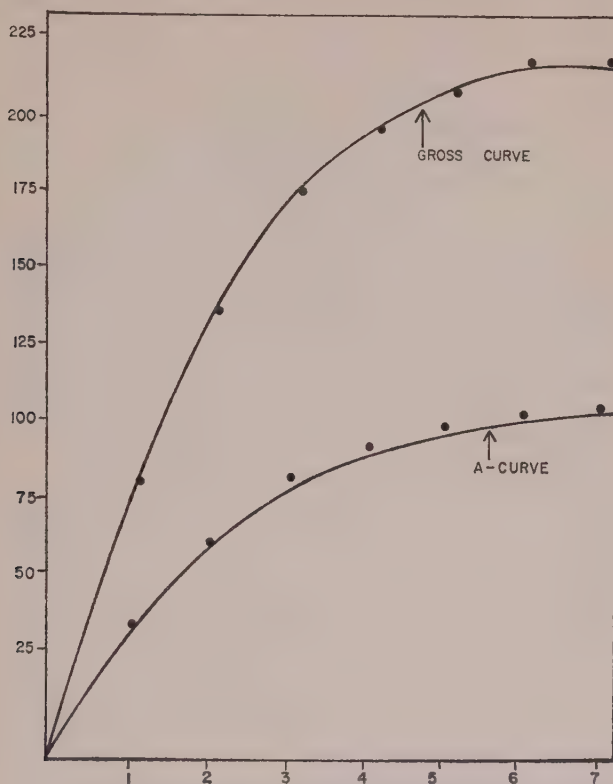


FIG. 3. Graphical representation of Table 1.

The results give the following values of the parameters:

$$k^* = 0.33 \quad E^* = 0.17.$$

Clearly, we do not speak of defection in this case. Instead we have a situation in which any unanimous choice is rewarded; the problem for the group is to decide (in the absence of communication) which choice to agree upon.\*

\* Schelling, T. C., "Bargaining, Communication, and Limited War". *Conflict Resolution*, 3, 114-19 (1959).



In our experiment, we have included one example of such a game, namely Game IV. We add, therefore,

Type 4. One defector is punished; two defectors are punished; three defectors are rewarded.

As in the Prisoner's Dilemma, the only way a group member can expect to win positive gains in the long run is by joining with the others to play the cooperative strategy in each game. (In a game of type 4, either strategy is cooperative, if it is unanimously chosen.) Therefore the problem before the group is to achieve self-organization. If communication is allowed, this can presumably be done by explicit agreement (collusion). But if communication is not allowed, such collusion can only be tacit. In this case, it is necessary for each group member to trust the other two implicitly, to believe that they will also trust him and each other and will resist the temptation to cash in on the trust of the others by defecting.

The subjects were partitioned from each other and forbidden to communicate during the run. In a preliminary orientation session, they were told how their collective choices to press one or the other of two buttons (right or left) affected all three scores in each of the eight games. Each had the entire score chart in front of him, which he could study for a half hour during the orientation session and the "dry runs".

In the experimental run, the number of the game to be played appeared on a display board visible to all. The subjects were given six seconds to make their choice. When all three buttons were pushed, the score to all the players appeared on the display board and their winnings or losses were added to the respective cumulative scores, also visible to all three. The scores were convertible to money at a mill per point.

In each experiment, each of the eight games was played 150 times, 1200 plays in all (about six hours). The games were presented in "shuffled" order, according to a random number schedule, somewhat modified to equalize the total frequencies of each of the eight games.

The score matrices are shown in Table 2.

Sixteen three-person groups in all participated in the experiment. In one group (No. 18), no communication was allowed even during the breaks. In the other fifteen the members were allowed



TABLE 2

If players A, B and C each choose the right or the left button as shown in the first column, the respective scores will be as shown in the remaining columns.

A B C	Game I Type 1			Game II Type 2			Game III Type 3			Game IV Non- compet.			Game V Type 1			Game VI Type 2			Game VII Type 3			Game VIII Type 2					
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
R R R	-1	-1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	1	-2	-2	-2	1	1	1	1	1	1	-2	-2	3
R R L	2	2	-2	-2	-2	6	2	2	-2	0	0	-1	-1	-1	3	-1	-1	1	1	1	-1	1	1	-1	1	1	1
R L R	2	-2	2	-2	6	-2	2	-2	2	0	-1	0	-1	3	-1	-1	1	-1	1	-1	1	1	-1	1	1	-2	-2
R L L	3	-2	-2	1	-2	-2	-2	1	1	-1	0	0	-1	2	2	6	-1	-1	-3	3	3	-2	3	-2	-2	-2	1
L L L	1	1	1	-1	-1	-1	1	1	1	1	1	1	-2	-2	-2	1	1	1	-2	-2	-2	-2	-2	-2	-2	-2	1
L L R	-2	-2	3	-2	-2	1	1	1	-2	0	0	-1	2	2	-1	-1	-1	6	3	3	-3	3	3	-3	-1	-1	-1
L R L	-2	3	-2	-2	1	-2	1	-2	1	0	-1	0	2	-1	2	-1	6	-1	3	-3	3	3	-3	3	3	-2	-2
L R R	-2	2	2	6	-2	-2	-2	2	2	-1	0	0	3	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-2	1	-2

to talk to each other only during the breaks. No suggestions were made to the subjects about the possibility of negotiating an agreement about how to play the games subsequent to the break. A comparison of the over-all relative frequencies of cooperative choices (counting all individual choices) before and after the first break, which occurred after 300-600 responses, is shown in Table 3.

TABLE 3

Group No.	No. of responses before first break	% Co-operative choices		Self-organization achieved?
		Before break	After break	
02	500	42.8	84.7	After communication
03	600	31.9	69.8	After communication
04	400	22.4	74.6	After communication
05	300	52.6	77.2	After communication
06	400	84.4	86.5	Tacitly
08	550	42.8	41.5	No
09	500	37.6	93.7	After communication
10	400	51.7	93.0	After communication
11	400	44.6	89.4	After communication
12	500	74.2	68.1	Tacitly
13	500	51.5	46.3	No
14	400	39.9	42.6	No
15	400	44.3	59.0	After communication
16	600	68.6	94.1	Tacitly
17	500	93.9	90.5	Tacitly
18	(Communication disallowed throughout)	42.1 (total)		No

The criterion for "achieving self-organization" is taken arbitrarily as 67 per cent individual cooperative responses. From the data we see that where this criterion is reached, it is usually substantially exceeded. In Group 15 it was not exceeded, but examination of the run shows that virtually complete cooperation was actually achieved after the break, but that it fell apart toward the end of the run, bringing the frequency of cooperative choices below criterion.

Taking grand averages to compare frequencies of cooperative response before and after communication, we get 51 per cent cooperative response before communication and 74 per cent after.

Morton Deutsch reports 36 and 70 per cent in runs of a two-person non-zero-sum game with communication disallowed and allowed respectively and with the subjects oriented "individualistically", i.e. instructed to try to maximize their individual gains.\*

Our grand average is over 8100 responses before communication and over almost twice that number after communication. Deutsch's sample, being from a single experiment, is much smaller. Hence comparison is not trustworthy. If, however, the difference is real, it appears that a greater propensity for cooperative responses was observed in our experiments than in Deutsch's.

We may attribute the observed greater tendency to form collusions in three-person game experiments either to some feature of the three-person game (as compared with the two-person game) or to the protracted length of our experimental runs. Both hypotheses seem plausible. It may be that the three-sided game is not seen quite as sharply as a "competition" compared with the two-sided one; or it may be that a protracted experience of being punished for non-cooperation induces the players to seek ways out of the impasse, an effect which is not observed in comparatively short runs.

We will now seek criteria which may be relevant to greater or lesser propensity to cooperate in games of this sort. Examination of the relative frequencies of cooperative choices shows that these frequencies vary considerably from game to game and that the rank order of the games with respect to their cooperative choice frequencies is approximately preserved in all the sixteen groups. It may be possible therefore to find a numerical index, such that the frequency of cooperative choice is strongly correlated with that index, as we range over the seven competitive games.†

We have postulated in turn the following criteria:

Criterion 1 (expected gains). If each player views each game as a choice between two simple gambles, each having four equiprobable outcomes (associated with the four possibilities of what the other two players may do), then each of the two choices has associated with it an expected gain. If now the player compares

---

\* Deutsch, M., "Trust and Suspicion". *Conflict Resolution*, 2, 265-79 (1958).

† Game IV, as has been pointed out, is non-competitive. In this game a unanimous choice was fixated early and remained throughout. This game, therefore, presents no interest in the present study.

the advantage of defection in terms of the excess of the expected result of defection over that of cooperation, we have the results shown in Table 4.

TABLE 4

Game	Advantage of defection over cooperation in expected pay-off
I	11
V	7
II, VII	3
VI	2
III, VIII	0

Criterion 2 (advantage over the other). In defecting from the cooperative choice, the player does not know whether other players will also defect. Suppose he computes the advantage (the difference of scores) over the *non-defecting* player(s) averaged over the two possibilities, namely that either no other player defects or that one other player defects. (If both other players defect, there is no advantage, since in that case all three scores are equal.) The results of rank-ordering the games according to this criterion are shown in Table 5.

TABLE 5

Game	Average advantage over non-defecting players
I	4.5
V	3.5
II, VI	2.5
VII	2.0
VIII	1.0
III	0.5

Clearly, many other criteria can be tried and compared. We shall confine ourselves to comparing the results with the indices derived from just these two criteria.

Plotting the indices for each of these criteria as abscissae against the observed frequencies of cooperative choices averaged over the sixteen groups, we obtain the graphs shown in Figs. 4 and 5.

Since we view the amount of cooperation achieved by the group *in the absence of explicit communication* as a measure of its self-organizing potential, we have thus obtained a measure of the

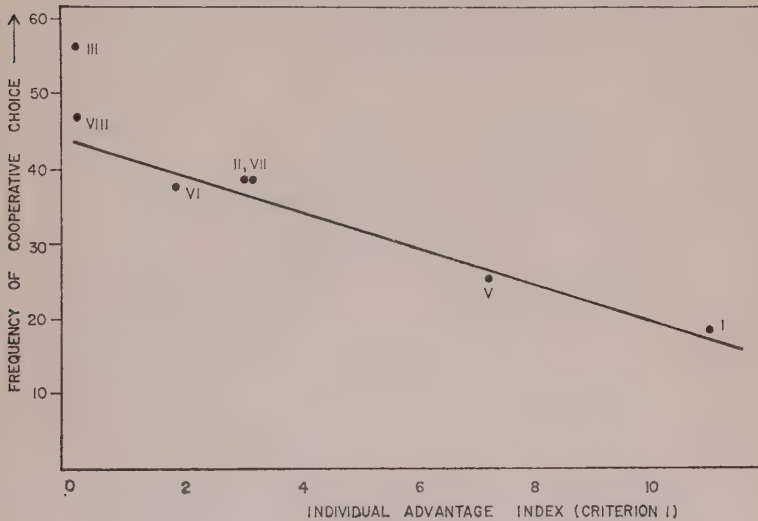


FIG. 4.

degree to which the competitive element in the situation adversely affects the self-organizing potential. This measure is reflected in the negative slope of the regression lines in Fig. 5, if Criterion 2 is accepted.

However, since the abscissae are not dimensionless (being in score units) while the ordinates, being relative frequencies, are, this measure can be used only relatively, e.g. in comparing one such regression line with another obtained under other experimental conditions or from other populations. If the correlation remains strong and only the slope of the line changes, we can draw conclusions about the relative strengths of competitive factors in various situations or in various populations as inhibitors of the self-organizing potential of the group.

In Group No. 18, communication among the members was not allowed even during the breaks. Here we could follow the time course of the cooperative choice frequency curve through the entire session in search of secular trends. These time courses are shown in Figs. 6, 7 and 8. The abscissae mark time periods of 200 responses each. We note the following.

1. Nearly perfect collusion is attained even in the absence of

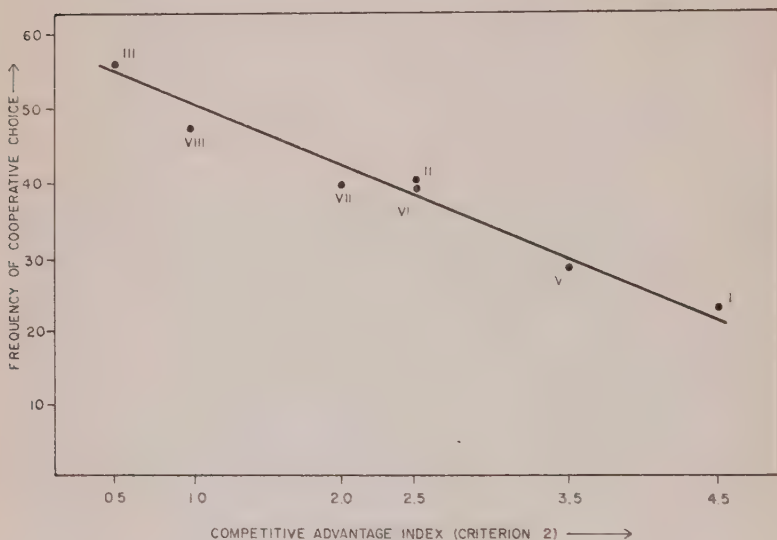


FIG. 5.

communication only in Game III (Fig. 6). Game III has the smallest index by both criteria and hence induces the greatest propensity for cooperation.

The other games (except VII) all show a characteristic U-shaped time course pattern. That is to say, typically cooperation frequency decreases during the greater part of the process, but increases toward the end of the process. Game VII, however, fails to "recover".

Since there were three other groups (8, 13 and 14) in which collusion was not achieved throughout the entire run, we could examine the time course of the responses in these runs to see whether the features observed in Group 18 were reproduced. We



found that the characteristic initial decline and subsequent rise of cooperative choice frequency were not reproduced. The fluctuations of cooperative response frequencies over the time periods were apparently random in the other three groups.

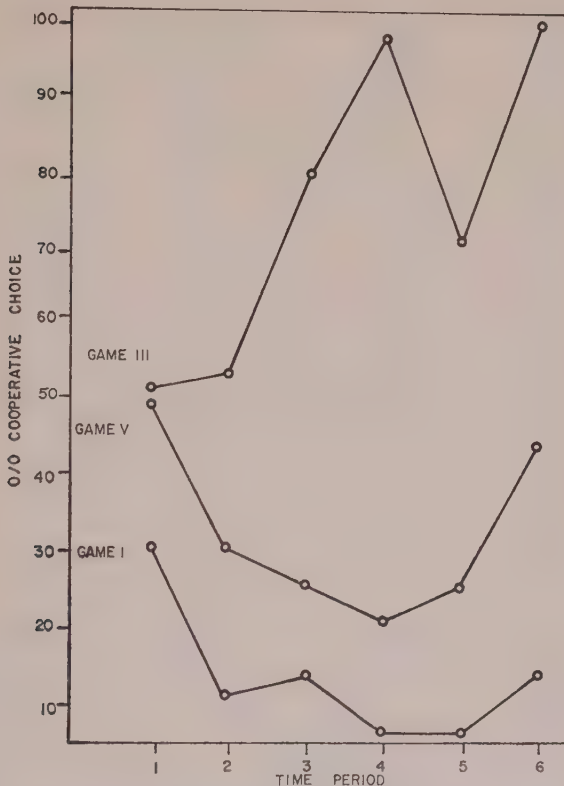


FIG. 6. Time course of frequency of cooperative choices in Games I, III and V.

The other interesting result pertaining to Game III was corroborated. In all the other three groups, where general collusion was not achieved, collusion was achieved in Game III, to the extent of 98, 90 and 67 per cent cooperative choices respectively, following the first break. Examining all early runs we see that collusion

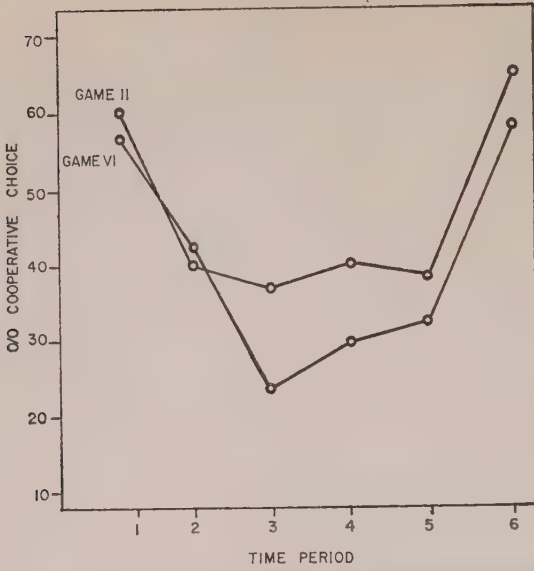


FIG. 7. Time course of frequency of cooperative choices in Games II and VI.

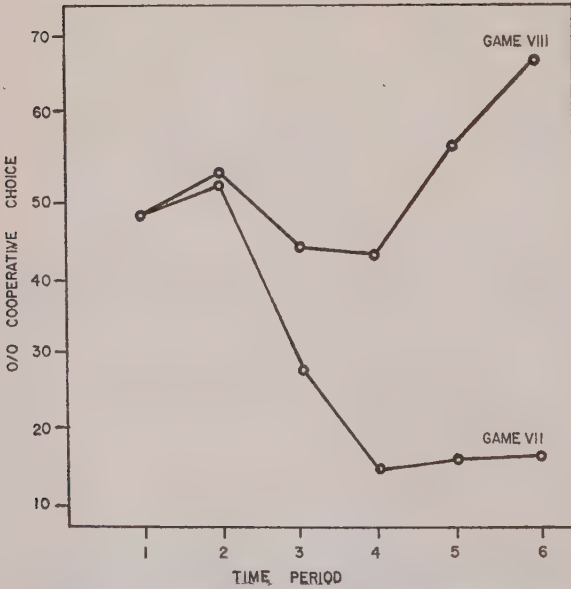


FIG. 8. Time choice of frequency of cooperative choices in Games VII and VIII.

in Game III is often already reached even in that stage, as shown in Table 6.

TABLE 6

Group	No. of runs to first break	% of cooperative choices in Game III	General collusion achieved?
02	500	57	No
03	600	35	No
04	400	16	No
05	300	91*	No
06	400	95	Yes
08	550	52	No
09	500	39	No
10	400	67*	No
11	400	61	No
12	500	89	Yes
13	500	81*	No
14	400	54	No
15	400	85*	No
16	600	74	Yes
17	500	96	Yes
18	(No communication allowed)	75*	No

\* Cases where collusion was reached in Game III, but not generally.

We have noted further that no other game except Game III has attracted tacit collusion where general collusion has not been reached.

To summarize, we can view the ability of a group to effect a collusion, i.e. to induce a willingness in its members to resist temptations to defect in pursuit of personal gain, as a self-organizing potential. When communication is disallowed, collusion usually fails to develop at least in the first few hundred runs. (*N.B.* Our runs consisted not of a repetition of the same game, but of several types of games, sharing the principal features described but varying in detail.) However, even in the absence of communication, self-organization was observed in four of the sixteen groups. On the other hand, in three of the sixteen groups, no self-organization was observed even after communication was allowed.

In runs where a general collusion has not taken place, the frequency of individual cooperative choices varies considerably and fairly consistently from game to game. The most consistent ranking arranges four of the games in the order of increasing frequency of cooperative choice as follows: I, V, VIII, III. Games II, VI and VII are usually found between V and VIII. Their rank order is not consistent, but usually II ranks above VI in frequency of cooperative choices.

Two indices have been suggested to explain the variation. The resulting variation lines both fit the data fairly well. The root-mean-square deviation between observed frequencies of cooperative choice and the theoretical regression line is 4.1 per cent if the comparative expected value index (Criterion 1) is used, and 1.9 per cent if the comparative competitive advantage index is used.

### CONCLUSION

Indices of group performance which can be taken to reflect the group's propensity for self-organization can be established in laboratory experiments where quantitative relations among performance variables can be observed. If these quantitative relations are expressed as equations, then the *constants* of these equations are the indices in question. The expectation is that as experimental conditions or the population from which the groups are recruited vary, the values of the parameters will change but the general form of the equations will remain the same.

If this expectation is realized, we can study the quantitative relations between these parameters and some indices related to the conditions or to the population. These relations should then be in turn expressed as equations which, in turn, will have parameters. One can then proceed as before with these parameters of the second order as the new variables, and so on.

## STAFFORD BEER

*Head of Department of Operational Research and Cybernetics,  
The United Steel Companies Limited, Sheffield*

# TOWARDS THE CYBERNETIC FACTORY

## PART 1

### *Informal Introduction*

#### 1. THE CONCEPT OF A CYBERNETIC FACTORY

A cybernetic system is recognizable by three outstanding characteristics. It is exceedingly complex: to the point where its interconnectivity is indefinable in detail. It is exceedingly probabilistic: to the point where its structure though complex becomes undifferentiated, and every trajectory is equiprobable. It is unreal to suppose that any such system can be controlled by the imposition of rules from outside; because the system by definition defies analysis, and therefore no test can be applied by which the adequacy of the rules could be judged. The third characteristic of a cybernetic system is, therefore, that the fundamental organization it displays is generated from within: it is self-organizing.

Any industrial company is, by these preliminary criteria, a cybernetic system. We may go further, and regard it as an integrated organism operating within an environment. This organism has a physical manifestation in a works (its body), and a set of interacting systems which nourish, energize, and regulate it (its digestive, cardiovascular and endocrine systems). It has a rate of working (its metabolism); it can grow by reproduction at the cellular level (mitosis), and can duplicate itself in subsidiary companies by reproduction entire.

Above all, it must interact intimately with its environment: there is a whole ecology of industry which has been little explored.

For there are only a few formal links which are generally recognized as standing for this connection between the company and the world outside, and these by no means reflect the abundant interaction that actually exists. A file of official correspondence ostensibly records the passage of information; but this does not reveal the personal contact behind the letters, nor the innuendoes they may hold for the recipient, nor the conventions by which they conceal what is meant, nor the information about the supplier contained in the product itself and its mode of delivery, nor the information about the customer contained in his acceptance of a slightly off-standard product. People may understand the intimate relationship between the company and its environment, but they have no formal model adequate to its expression.

Analogical thinking about the company as an organism can be continued in this literary vein indefinitely; and I believe that (if necessary) the model of the company, operating within its context of supply, demand, labour market, business confidence, and so on, could be mapped with scientific precision onto an organism living in its environment. But we will assume that no serious objection exists to the drawing of this analogy, since no detailed use will be made of its morphology, and pass to the question of the higher levels of control by which a system that is already viable in a passive sense achieves purposeful activity.

This involves a consideration of the nervous system. The flow of information between an industrial concern and its environment, and within the concern as affecting its own behaviour, is to be considered as an analogue of the flow of information into the sensory receptors, through the neural network, and out of motor activities in the organism. Again, a fairly loose analogy is adequate for our purposes at the lower levels, because no detailed use will be made of the comparison. Consider an automatic factory. This is very like a spinal dog; it has a certain internal cohesion, and reflex facilities at the least. When automation has finished its work, the analogy may be pursued in the pathology of the organism. For machines with over-sensitive feedback begin to "hunt"—or develop ataxia; and the whole organism may be so specialized towards a particular environment that it ceases to be adaptive: a radical change in the market will lead to its extinction. This is why "automation" in the generally accepted engineering usage of



the term, is applicable only to industries (such as the automobile industry) with a guaranteed effective demand, a fairly rigid design policy, and a merely superficial response to a relatively slow cycle of change (e.g. fashions in colour and size of "tail fin"). The cybernetic factory seeks a more adaptive, more readily responsive, automatic system, which would be capable of controlling any kind of company.

When men have been almost eliminated from the factory, and it runs smoothly and efficiently by automatic regulation, error-controlled feedback, and programmed response to a specified and limited variety of situations, we have the living organism of the company as the analogue of (say) an animal whose nervous system stops at the cerebellum. It is sentient: "aware" in the sense of a capacity for autonomic response to states of its own body and a limited range of stimuli from the outside world; but it is suffering seriously and indeed fatally from sensory deprivation. At present, such an automatic factory must rely on the few men left at the top to supply the functions of the cerebrum. And we may note that, even if they constitute a board with high intellectual ability, the whole organism is a strange one—for its brain is connected to the rest of its central nervous system at discrete intervals of time by the most tenuous connections. The survival-value of such a creature does not appear to be high.

The concept of a cybernetic factory is that the industrial company *as a whole* is a living organism; and it must display the features by which a living system operates if it is to remain viable. Now in the past industrial companies have been like this: they were self-organizing systems, and they were viable. But as economic and social pressure increases, they are increasingly driven to become more efficient, more specialized, more automatic, and so on, all of which tendencies threaten their long-term existence. Competition in our modern society forces management to aim at high and rapid profits: in ecological terms, at *dominance*; and with ecological dominance is linked (as a fact of natural science) decreasing adaptability. It is only with man, with the emergence of intelligence and the ability to predict, that this trend in evolution has been checked (we hope: but the state of the world today suggests that nature may yet impose her veto on the dominant species). For, we say, man is a highly specialized creature who

has nevertheless retained his adaptability by growing a large brain and obtaining volitional control of himself in relation to his environment. Thus he loses, by specialization, his prehensile toes and tail; but he invents machines called elevators which still enable him to climb the Eiffel Tower in safety.

When we turn to contemporary management theory, then, it is not surprising that techniques are aimed at short-term dominance. Pay-off is paramount, and the five-year plan is regarded as a triumph of perspicacity. Accountants are expected to organize institutional information to the maximization of profit; linear programmers are asked to solve scheduling and transportation problems under the minimizing functional of cost; publicity men are awarded large appropriations to make less mutable an environment to which the organism cannot adapt. If the dinosaur can no longer live in the world, the world must be turned into a dinosaur sanctuary.

This will not do. The spinal dog is short of a built-in cerebrum; and the automatic factory is short of a built-in brain. The research discussed in this paper is directed towards the creation of a brain artefact capable of running the company under the evolutionary criterion of survival. If this could be achieved, management would be freed for tasks of eugenics: for hastening or retarding the natural processes of growth and change, and for determining the deliberated creation or extinction of whole species.

## 2. THE HEURISTIC APPROACH USED

It is possible to contemplate the problems of creating a cybernetic factory from an armchair, but very much more valuable to study them in an actual works. We have been most fortunate in having the opportunity to do so.

The company concerned is a relatively small, self-contained firm whose product is steel rod manufactured in a variety of sizes and qualities. At the time when the work began the company had just acquired a new senior management group, who had every intention of modernizing and improving an already prosperous firm. My department joined with them in this task, and a three-pronged attack was launched using the methods of work study, operational research, and cybernetics. Stock control, stores control, financial

control, cost control and other functions of management were investigated; proposals were made, accepted and implemented. The plant itself was intensively studied: some of its equipment was changed, and (with the co-operation of an engineering consultant) a revolutionary kind of mill was developed and is now undergoing trials. Automation was greatly extended by the management, and several radically new kinds of electronic control equipment were invented by the study team. This is background information only; but it will be seen that under its new executives the company rapidly attained the status of a first-rate up-to-date concern, with consequent benefits to its economy and stability.

The cybernetic study, beginning with the concept already explained, went on to construct a model of the company organism in its environment, and to detect the brain-like aspects of its control. A generalized model for a brain artefact is being evolved as the theoretical basis for a machine to undertake the functions of an automatic brain. Simultaneously, a specialized model appropriate to controlling an industrial company is being constructed as a means of linking the developing theory with real life. This special theory derives from a detailed operational research analysis using actual data, in which an arbitrarily chosen set of the intended operations of the brain artefact is used to direct calculations carried out by hand and eye. The results are offered to management as at least "partly-cerebrated" information, and have an immediate practical significance and value. Thus the theoretical work directs the growth and exploitation of its own evolving exemplification in real life, while the real-life exemplification helps to generate and validate the theory. All this work is virtually complete. The next and final stage is to construct machinery capable of interpreting the full theory in practice.

In other words, this heuristic approach has been to study what actually happens, using the language of organic cybernetics to structure it; and to develop both a special and a general theory to account for it. The special theory is *ad hoc*; it is put into effect by empirical methods: recording, analysis and synthesis of actual data; statistical, mathematical and logical investigations of these data; the use of graphical methods, calculators and computers; the construction of perspex models; the cinephotography of the model of the adapting organism in its environment. The general

theory claims, on the other hand, to be general in two ways. It does not make arbitrary decisions about the nature of reality, as any exemplification of the special theory has to do; and it is not uniquely appropriate to an industrial company, as the whole special theory has to be.

### 3. THE PRESENTATION OF THE THEORIES

There is no particular virtue in placing the general before the special theory, as has been done. It is perhaps more natural for cyberneticians to consider the generalization first, as offering a more complete insight, and to regard the special theory as a limiting case. The man more versed in industry, however, might prefer to master the special theory first, as offering a tangible example of the thinking. In the latter case, he must be careful to recognize that questions settled there empirically are dealt with by quite other means when considered by the general theory.

And this leads to the elucidation of a difficult point. The special theory deals with brain artefacts for cybernetic factories, and the general theory with brains and their artefacts. It is not possible to present here either theory in full (they are both incomplete, in any case). Thus what is really put forward in Part II is that part of the general theory which is relevant to this industrial project; and what appears in Part III is a particular exemplification of the *special* theory (which in its turn exemplifies the general theory). This fine distinction must be understood in order to comprehend the status of the cybernetic machine it is intended to build. This is a manifestation of the special theory, and an exemplification of the general theory.

These considerations have led to an important difference between the following two parts of this presentation. Insofar as Part II is incomplete, it is because whole areas of the general theory are not mentioned. For example, the theory includes accounts of reproductive logical machines and of the amplification of intelligence that are not immediately relevant. They are simply omitted. But Part III is incomplete in the different sense that it is impossible to enumerate all exemplifications of the special theory which is embodied in physical apparatus. A convention is therefore adopted in Part III which attempts to *infer* the special theory from its quoted exemplification.



There is this to say in conclusion about the *special* theory and its physical embodiment. A considerable knowledge of Ross Ashby's work<sup>(1,2)</sup> is assumed. In particular, I am drawing on the fundamental ideas that a real system proliferates variety; that to control this it must be "absorbed" by an equal or greater ("requisite") variety; that this can be done by generating variety to match it; and that the generation does not have to be artificial (such as injecting random white noise), but can be achieved by defining the system itself so that the parts are in homeostatic balance and the whole ultrastable. Secondly, a considerable knowledge of Gordon Pask's work<sup>(3)</sup> is assumed. I am drawing especially on the colloidal thread-structured machines he has built (which provide a suitable self-organizing high-variety fabric), together with my own<sup>(4)</sup> conclusions on the same theme. We are continuing our collaboration in designing the physical apparatus for this project.

Finally, there is this to say in conclusion about the *general* theory and its relevance to neurophysiology. At the start of this work, I made serious attempts to distinguish between my understanding of the living brain on the one hand, and the invention of the brain artefact on the other. A word such as *medulla* refers to something in the head; a "medulla" in an artefact is not a genuine medulla. But the term had to be used, because there is no counterpart in engineering, and so it went into inverted commas. But I am not a neurosurgeon; doubtless the aspects of the medulla I had in mind were insignificant properties of the real medulla. Before long, the word medulla was appearing in two kinds of print and two sets of commas. Meanwhile, I had become simultaneously so involved in the brain and the company and the artefact and the logic, that it was not any longer possible to be sure of which I was thinking. The conventions adopted disintegrated. So they were discarded. The result is that Part II reads as if there is no difference between any of the manifestations of brain—in reality, in analogy, in machinery, and so on.

In a sense, this does not matter. For the object is to obtain a mapping of a brain onto a machine, and insofar as this is done they *are* the same things. The critical word is "insofar". They are formal homomorphisms (albeit of a many-many-one transformation), and they are operational equivalents (albeit at a trival level of isomorphism). In speaking of "the brain", I speak only of a

*useful slant on a useful aspect* of the brain. What ignorance this causes me to betray about the living jelly in all our heads I dread to think; but I trust my fellow symposiasts will sympathize, for they have met the same expository difficulties themselves.

## PART II

### *From the General Theory : Set-Theoretic Formulation of the Brain Model*

#### 1. PASSIVE SENSIBILITY

1.00 *General*. The first aspect of the brain artefact to be considered is its passive sensibility. This is defined as a collection of inputs which inform the brain about the state of the world, which is in turn defined as the state of the whole organism in relation to its environment. The word "passive" is intended to show that these inputs are purely sensory: they register in a sensory cortex, stop short of any possible motor reflex, and have no conative implications.

1.10 *The Sensory Cortex*. In any brain or brain artefact, because it is a physical manifestation, there must be a limited total channel capacity for sensory input, of variety  $W$ . Consider a time interval  $\delta t$ , called a quantum of time,  $t_0$ . At time  $t_0$ , each of the sensory input channels delivers to the cortex a sensation, including the case where the sensation registered is "no sensation". In the living brain, each sensory input is represented by a large number of afferent nerves, each of which delivers one bit of information at time  $t_0$ . In the brain artefact, we shall consider an analogue input, which will reduce the number of afferent channels and preserve the total variety  $W$  by permitting each input to take up a value  $x$  on a continuum.

It is now possible to give a set-theoretic description of the sensory cortex.

1.11  $s$  is a sensory input, an element of the set of sensory inputs,  $\mathcal{S}$ .



1.12 To every element  $s$  of  $\mathfrak{S}$  there corresponds a dense, denumerably infinite set  $\langle x \rangle$  equivalent to the set of rational numbers in natural order of closed interval  $\langle 0, 1 \rangle$ , whose elements  $x$  are values ascribed to  $s$  on the continuum.

1.121  $x \in \langle x \rangle$  is uniquely specified for every  $t_0$ .

1.13 The set  $\mathfrak{S}$  of sensory inputs  $s$  side numerable and finite. It can therefore be ordered by correspondence with the set of natural numbers, and can be written down as the well-ordered set:

$$\mathfrak{S} = \{s_1, s_2, s_3, \dots, s_n\}$$

of cardinal numbers  $|\mathfrak{S}|$ .

1.131 A rule of the ordering shall be that sensory inputs deriving from the organism itself (in this case the company) belong to the segment  $\mathfrak{S}_{s_v}$ , determined by  $s_v$ .

1.132 Sensory inputs deriving from the environment belong to the complement of  $\mathfrak{S}_{s_v}$  in  $\mathfrak{S}$ , namely the remainder subset.

1.133 Thus for example:

$$s_t \in \mathfrak{S}_{s_v}$$

$$s_n \in \mathfrak{S} - \mathfrak{S}_{s_v}$$

1.20 *Sensory Configurations.* The question now arises whether the separate elements  $s$  of the input set  $\mathfrak{S}$  can be treated as independent of each other or not. For there is clearly a sense in which a particular set of inputs forms a sensory gestalt which behaves as an entity.

1.201 For example, the sensory gestalt recognized as "a rose" is a configuration of visual, tactile and olfactory sensations. If the scent of the rose is missing from the configuration, the recognition probabilities may swing from "rose" to "artificial rose".

1.21 It can be asserted that for the brain artefact, as also for the living brain, no element  $s$  of  $\mathfrak{S}$  is independent of every other element  $s$  of  $\mathfrak{S}$ . The world picture is structured, and with it the object-language by which the brain models the world, by axioms of configuration of various kinds. For example:

1.211 as given in logic:

$$(s_1 \cdot s_2 \cdot \sim s_3) \vee (s_1 \cdot s_2 \cdot s_3 \cdot \sim s_4) \cdot \supset \sim s_5,$$

where the activation ( $s$ ) or not ( $\sim s$ ) of a sensory input refers

simply to the binary possibility that there is present or absent  $s_1 =$  a rolling mill,  $s_2 =$  a labour force,  $s_3 =$  some raw material,  $s_4 =$  some product,  $s_5 =$  some income.

1.212 or as given in mathematics:

$$x(s_4) = f[x(s_1), x(s_6)],$$

where  $x$  is the variable denoting  $s_4 =$  some product,  $s_1 =$  the speed of the rolling mill,  $s_6 =$  the gauge of rod.

1.213 or as given in statistics:

$$r = \frac{\text{cov}(x(s_7), x(s_4))}{\sqrt{(\text{var } x(s_7) \text{ var } x(s_4))}} \rightarrow 1,$$

where  $x$  is a variate measuring the level of  $s_7 =$  wages,  $s_4 =$  product, for a system of monetary incentives.

1.22 Such axioms of configuration are given in the language that models the world. They are given, that is to say, in the object-language itself and not in any syntax-language (taking Carnap's usage of these terms), despite the foregoing attempts to *allude* to them in what are evidently syntactical languages. The object-language is the language *of the brain itself* (for the specification of which von Neumann used to appeal) which the cybernetic artefact must speak.

1.221 The brain, then, adopts these configurations as gestalten: as axioms of the object-language they must be developed as the brain "learns to speak". This implies a relatively short-term learning programme for the "infant" brain, which will establish these structures as permanent features of the cognitive landscape.

1.2212 For example, the multiplication table may be looked upon as a sensory configuration. Given a brain-with-a-language (that is an "adult" brain) its mathematical judgments will be analytic of the axiomatic structure of its own language; and (since we *started* with an adult brain, or (also) since all analytic judgments are by definition *a priori*) mathematical judgments are always *a priori analytics*. But if what is given is a pristine brain, whose language is yet to evolve, we shall have to consider its behaviour *operationally*. In this case, mathematical judgments will be learned *a posteriori* from instances of the naming of collections of numbers; and (since we *started* with a pristine brain, or (also) because

all experiential judgments are by definition synthetic) mathematical judgments are always *a posteriori synthetics*. (This cybernetic and essentially operational examination of the problem throws possible light on the controversy over Kant's doctrine that mathematical judgments are always *a priori synthetics*.)

1.222 There are two main reasons why this structuring of the object-language into axioms of configuration, with which correspond the cognitive structures of gestalten, is essential to a viable brain or artefact:

1.2221 They greatly constrain the freedom with which a particular point representing an important world event can range over a multidimensional phase-space, and thereby enhance the speed and accuracy of its recognition and identification.

1.2222 Without this mechanism, the brain would presumably waste most of its energy discovering and learning (precariously) every new exemplification of an axiomatic relationship that came its way. The brain would thereby rediscover in endlessly new guises structures implicit in the world picture object-language from the beginning, and doubtless fail to discover vital operational relationships about the way these effectively aprioristic structures happen (fortuitously) to work.

1.23 It is therefore supposed that the brain, in order to make speedy, accurate and non-trivial judgments about the world picture, must include machinery for modelling axioms of configuration from the input set  $\mathfrak{S}$ .

1.231 These configurations may well be time-dependent in some cases (for an adequate object-language is steadily modified and enriched); so our discussion of configuration will be concerned initially with time  $t_0$ .

1.24 The transmission of quantitative analogue information deriving from the set complex  $\mathfrak{S}(X)$  need not be considered in this problem of configuration. We consider instead the formal cortical networks generated by  $\mathfrak{S}$ , for which the  $i$ th elemental sensory input is either activated ( $s_i$ ) or not ( $\sim s_i$ ).

1.25 A sensory configuration is a subset of  $\mathfrak{S}$ , namely  $S$ , of elements  $s$ . Since every subset of a well-ordered set is itself well-ordered, we have:

$$S = \{s_\nu, s_\delta, \dots, s_\theta\}$$

for all subsets of  $\mathfrak{S}$ .

In particular, a sensory configuration may be:

1.251 any proper subset:  $S_a \subset \mathfrak{S}$ .

1.252 the improper subset:  $S_n \subseteq \mathfrak{S}$ .

1.253 a unit set:  $S_b = \{s_\mu\}$ .

1.254 the empty set:  $S_0 = \mathfrak{S} - S_n$ .

1.26 The totality of possible sensory configurations is therefore the set of the subsets of  $\mathfrak{S}$ , namely  $\mathcal{U}(\mathfrak{S})$ ,

1.261 of which there are  $2^n$  (because  $|\mathfrak{S}| = n$ : *vide* 1.13).

1.27 Thus the total possible sensory configurations are given by the well-ordered set of all subsets of  $\mathfrak{S}$ :

$$\mathcal{U}(\mathfrak{S}) = \{S_1, S_2, S_3, \dots, S_{2^n}\}$$

of cardinal numbers  $2^{|\mathfrak{S}|} = 2^n$ .

1.28 The brain or brain artefact may seize, then, on any sensory configuration as a convenient instrument for removing effectively axiomatic structures from its sensory input (*vide* 1.23).

1.281 An illustration would be to consider *jointly* the tonnage output in a given time interval, together with the gauge of rod being rolled. Assuming a constant mill speed, the use of this configuration  $S_a = \{s_4, s_6\}$  gives a network which (given a suitable transformation for the appropriate  $X$  sets of this ordered pair of inputs) allows the brain to measure relative *efficiency*, which is certainly not itself a raw sensation.

1.282 The brain may seize on as many such sensory configurations as are convenient at any time  $t_0$ , subject to an upper limit of the number of such configurations available, namely  $2^n$ .

1.29 At any particular time  $t_0$ , therefore, there will be  $c \leq 2^n$  sensory configurations in the brain artefact, which will be visualized, for obvious neurophysiological reasons, to exist as a cortical network in the fourth layer of the sensory cortex. Any recognition the brain might take of a raw sensation itself will occur elsewhere: perhaps in the first (molecular) layer, to which some afferent fibres pass directly, for instance.

1.30 *The Fourth Layer of the Sensory Cortex* is at least partly defined by a set of many-one mappings of  $\mathfrak{S}$  into  $\mathcal{U}(\mathfrak{S})$ . One such mapping is  $S_b$ , which is a subset of  $\mathfrak{S}$  of elements  $s$  (*vide* 1.25), called a sensory configuration.

1.31 At time  $t_0$ , there exists in the fourth layer a particular set

$[S]$  of subsets  $S$  of the set  $\mathfrak{S}$ , which is a subset of  $\mathcal{U}(\mathfrak{S})$ :

$$[S] = \{S_b \dots S_\gamma\} \subset \mathcal{U}(\mathfrak{S}).$$

1.32 Regardless of time, therefore, there may occur in the second layer *any* such set  $[S]$  of elements  $S$ ; and all possible states of the fourth layer are given by the sets of the sets of the subsets of  $\mathfrak{S}$ :

$$\mathcal{U}(\mathfrak{S}) = \{[S_1], [S_2], [S_3], \dots, [S_{2^n}]\}$$

of cardinal number

$$2^{2^{|\mathfrak{S}|}} = 2^{2^n}.$$

1.33 Notwithstanding the high variety made available at this stage (which seems no more than realistic in a brain artefact) it is to be expected that the full flexibility it provides is only strictly valuable during the “infancy” of the organism. There is reason to think that coarse structures, “macronetworks”, are soon formed by the living brain: they should be formed early in the artefact too. The machinery for categorical associations (e.g.  $2+2=4$ ), and paths of facilitation for fundamental modes of behaviour (e.g. certain conditioned motor reflexes), can be decided upon quite soon and may never have to be altered.

1.331 But the high level of redundancy this implies for the “adult” brain in this facet of its behaviour is a reserve available should subtle modifications of the macronetworks be demanded. (Gross traumata might of course result in a demand for all this redundancy in the task of re-adaptation.)

1.34 A mechanism has to be envisaged to manifest these proposed operations of the fourth cortical layer of the artefact, which do indeed appear to imitate closely the corresponding activities of the living cortex (*vide* 1.35).

1.341 The afferent inputs must arrive, and register in the fourth layer in ways which will permit of scansion, grouping and pattern recognition necessary to the establishment and operation of the postulated system of sensory configuration (*vide* 1.355).

1.342 They must be registered in different modes of connectivity, to permit flexible interactions between themselves and outgoing axons (*vide* 1.351).

1.343 Summations of activity against time will inevitably lead



to rhythmic activities which map periodicities in the real world, and these ought to be detectable. If the artefact continues to imitate its living model by a spreading horizontal network of afferent fibres and the dendrites of pyramidal neurons in the first layer, then the artefact must display these rhythms superficially (*vide* 1.354).

1.3431 A sufficiently strong surface stimulus will therefore generate superficial responses, as happens with living tissue, and the artefact mechanism must account for this (*vide* 1.352).

1.344 The mechanism as so far discussed is primarily concerned with identifying and *separating* sensory configurations, information about which will be sent back to the mesencephalon via the axons of pyramidal cells. But there are cross-correlations and comparisons to be drawn between such outputs. An archetype for such activity might be found at deeper cortical levels (the fifth layer?), but it would presuppose preliminary "cross-talk" between fourth layer outputs. What mechanism can there be for spreading this information across the layer of pyramidal cells? Surely this is the artefact analogue of *deep* response to a surface stimulus, which is propagated at the pyramidal level, and which is subsequently detected superficially elsewhere (*vide* 1.353).

1.35 This schematic diagram attempts to describe the machinery of the fourth layer of the cortex in a form suitable for handling information as required in the brain artefact for the cybernetic factory, as required by 1.34. It is a composite diagram, based mainly on Eccles,<sup>(5)</sup> but is also indebted to the work on intracortical chains of neurons by Lorente de No,<sup>(6)</sup> and to the group scanning process ascribed to McCulloch by Wiener.<sup>(7)</sup>

1.351 The picture comprises eleven basically similar arrangements: a pair of afferent input fibres reach the cortex and ascend to the fourth layer. Here they branch, and end in synaptic knobs on dendrites of pyramidal cells or on interneurons (indicated by circles). The required connectivity by which the afferent fibres and neighbouring dendrites can communicate in every mode is indicated in the channel marked *B* (cf. 1.342). The channels 0-8 indicate a simplified convention for the same network. Only apical dendrites are shown from the deep pyramidal cells.

1.352 On the extreme left a pair of stimulating electrodes ( $S_1$ ) is applied to the cortex, and a wave of superficial responses is seen



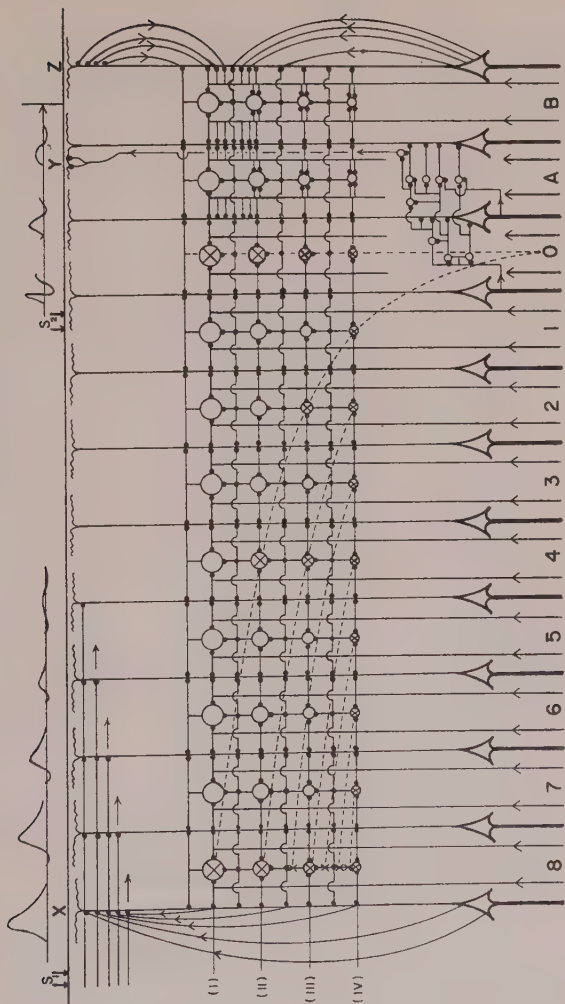


FIG. 1.

moving across the cortex from  $X$  to the right. Eccles's postulated mechanism for generating this response is that the axons of horizontal cells (or afferent fibres) transmit impulses in the direction of the arrows, and contribute synaptic knobs to the dendrite on which each terminates, as well as to each which it encounters on the way. Lines of extrinsic current flow are shown for  $X$  alone, rising from deep sources to sinks on the apical dendrite. Similar lines will be visualized rising to successive dendrites, of decreasing intensity of current flow corresponding to a decrease in the superficial propagation of impulses as the horizontal axons terminate (cf. 1.3431).

1.353 In channel  $A$  an ascending axon (broken on the diagram) appears at  $Y$ , by which Eccles supposes that deep responses spread to the surface. His postulated mechanism for deep response itself begins with a strong stimulus at  $S_2$  which passes down the pyramidal cell dendrite between 1 and 0 to the soma, and along its axon and axon collateral. Special short-axon neurons are drawn by which impulses are propagated to the two neighbouring pyramidal cells—a response which would continue to spread to the right. Synapses excited at these deep levels provide sinks for extrinsic current flow downwards from sources on the apical dendrite (cf. 1.344).

1.354 On the extreme right, in channel  $B$ , an afferent volley generates an initial positive wave on the cortical surface by Eccles's third postulated mechanism. Here there is extrinsic current flow from sources on both the superficial level of the apical dendrite  $Z$  and the soma to sinks on the deep dendrite level (cf. 1.343).

1.355 Across channels 0–8 is shown a new connectivity to provide a group scanning and selection mechanism. Channel 0 is to be regarded as an arbitrary central focus, and connections are seen to develop on the left which would be mirrored on the right. These effect a logarithmic transformation for the afferent inputs such as is needed in a configuration selection and pattern recognition process. The layers of interneurons (i)–(iv) are scanned cyclically. Each successive layer contains a higher cell density, made up of smaller cells, than its predecessor: the cells affected by the particular mechanism shown (i.e. for centre 0) are marked to display these facts (cf. 1.341).

1.36 Returning now to the output of this fourth cortical layer artefact, a particular set of outputs will be discussed. This is:

$$[S_a] = \{S_b, \dots, S_v\} \subset \mathcal{U}(\mathfrak{S}) \subset \mathcal{UU}(\mathfrak{S}).$$

1.361 For we have from 1.31 that this is a possible set of outputs, consisting of sensory configurations: a subset of the set of subsets  $\mathcal{U}(\mathfrak{S})$  of sensations  $\mathfrak{S}$ .

1.362 And we have from 1.32 that this set  $[S]$  is included in the time dependent range of possible configurations of configurations, that is, of  $\mathcal{UU}(\mathfrak{S})$ .

1.363 In other words, any particular configuration  $S$ , and any particular set of configurations  $[S]$ , are determined by a particular time  $t$ . If the denotation of  $S$  and of  $[S]$  are determined by time  $t_0$ , then the denotation may have changed at time  $t_1$ . (This is relevant to 1.40.)

1.364 On the other hand, from the argument at 1.33, for a sufficiently short span of time, an "adult" brain artefact will not actually modify a well-established configuration  $S$ , nor the configuration of such configurations  $[S]$ . (This is relevant to 1.50 et seq.)

1.40 *Computable Values of Configuration Measures.* Every element  $s$  of every configuration  $S_n$  which is a subset of the set of configurations  $[S_a]$  is a sensory input. With every  $s$ , at every time  $t$ , is associated a measure variable  $x$ : itself a member of the population of measures  $\langle x \rangle$  which define  $s$  quantitatively through time (vide 1.12).

1.41 Thus for any set of elements  $s$  which denote any configuration  $S_n$ , a one-one correspondence with a measure set  $X_n$  may be asserted by the equivalence:

$$S_n = \{s, s_1, s_2, \dots, s_n\} \sim \{\langle x \rangle, \langle x' \rangle, \langle x'' \rangle, \dots, \langle x^n \rangle\} = X_n.$$

1.411 and at time  $t_0$ , when  $x \in \langle x \rangle$  uniquely:

$$S_n(t_0) = \{s, s_1, s_2, \dots, s_n\} \sim \{x, x', x'', \dots, x^n\} = X_n(t_0).$$

1.42 The set  $X_n$  has in general a *computable value* called  $\hat{x}_n$ . This is a number which ascribes a quantity to  $S_n$ , and which varies with time.

1.421 For  $X_n(t_0)$ ,  $\hat{x}_n(t_0) = f(x, x', x'', \dots, x^n)$ .

1.422 But this statement (1.421) *cannot be generalized*. For in general, and even under appropriate constraints,

$$\hat{x}_n \neq f(x, x', x'', \dots, x^{\max})$$

for all  $t$  with respect to  $S_n$ .

1.4221 *Proof*. Suppose  $\hat{x}_n = f(x, x', x'', \dots, x^{\max})$  for all  $t$ . Then  $\hat{x}_n \supset X_n(t) = \{x, x', x'', \dots, x^{\max}\}$ . Then if  $\hat{x}_n(t_1)$  is computable, the residual set is a null set:

$$1 - X_n(t_1) = \Lambda$$

by which device

$$X_n(t) \sim S_n(t_1)$$

while

$$X_n(t_0) \sim S_n(t_0). \quad (\text{vide 1.411})$$

But (cf. 1.363) suppose that

$$S_n(t_1) = \{s_2 \dots s_n\}$$

then

$$S_n(t_1) \subset S_n(t_0) \text{ per def.}$$

while

$$X_n(t_0) \subset X_n(t) \text{ per def.}$$

Then

$$X_n(t) \sim S_n(t_1) \subset S_n(t_0)$$

and

$$S_n(t_0) \sim X_n(t_0) \subset X_n(t).$$

Therefore by Bernstein's equivalence theorem,

$$\begin{aligned} X_n(t) &\sim S_n(t_0) \\ &= \{x, x', x'' \dots x^{\max}\} \sim \{s, s_1, s_2 \dots s_n\} \end{aligned}$$

which is true if  $n \equiv \max$ —a very special case. Therefore

$$\hat{x}_n \neq f(x, x', x'', \dots, x^{\max}).$$

1.4222 This says that although the value of the measure of any configuration can be written mathematically as an algebraic function, using all variables  $x$  as arguments, for a given time  $t$  (at which the logical structure of the configuration can be examined), it cannot be generalized for that  $S$  for all  $t$ . For although there is a mathematical device (namely: put  $x_a = 0$ ) which would permit of a correct particular computation within this general model, the use of the model destroys the tenuous time-dependent *logical* structure of  $S$ . Just as "the set of all cardinal numbers" is a meaningless concept, so "the set of all variables  $x$  associated with all inputs  $s$ " becomes meaningless when predicated of a particular

*S.* (Some time has been spent on this point, because it is the main reason for the choice of a set-theoretic language in this description.)

1.43 Thus from 1.41 we have  $S_n \sim X_n$ , by which it is quantified, and from 1.42 that  $X_n$  has a computable value  $\hat{x}_n$  defined at time  $t_0$  according to 1.421 by an algebraic function. But either this function cannot be stated in a general form for all  $t$  with respect to  $S_n$ , or alternatively it cannot be stated in particular form for fixed  $S_n$  with respect to all  $t$ . That is: if the configuration  $S_n$  retains its identity through time, then it cannot be denoted by a general algebraic function; whereas if it can be so denoted it cannot retain its identity. This law is called the *Indeterminacy of Configuration Structure*.

1.44 Despite this Indeterminacy, however,  $\hat{x}_n$  is computable at any time  $t_0$ ; and the only general remark about its value is that it is a numerical "blend" of the set of values denoted at that time by  $X_n$ .

1.50 *Statistical Homogeneity of Computable Values.* It follows from 1.4 that if we consider the quantification of  $S_n$  through a period of quantized time, a set of computable values (of which  $\hat{x}_n$  is an element) will be obtained. These values may be considered as statistically distributed about their own mean, and the distributions so formed are likely to be remarkable in at least three ways.

1.501 From the arguments of 1.2, it is evident that a configuration offers a relative, not an absolute, measure. In order to remove axiomatic structures from its sensory input (*vide* 1.28) it will produce blends which are effectively ratios. This must have the effect of distorting the distribution of  $\hat{x}_n$ . It will not be symmetrical, but skewed.

1.502 The Indeterminacy law seems to imply imprecise machinery for computing  $\hat{x}_n$ , and no adequate mechanism has been invoked which would filter the original sensory inputs of noise. It seems likely, then, that the statistical time series providing the set of values will exhibit considerable chance fluctuation.

1.503 The evaluation of computable values is occurring for each configuration in the set of configurations, and there is no basis for supposing that the variability of the raw numbers produced will be comparable across the whole set.

1.51 The assumption is now made that the brain artefact will

find some degree of statistical homogeneity convenient in its treatment of these numbers. To achieve this, a succession of statistical transformations will be necessary.

1.511 There are various transforms (for example,  $\theta = \sin^{-1}\sqrt{(x/X)}$ ) that will tend to return a skewed distribution based on ratios to normal. There are various transforms (for example, those eliminating some extreme quantile) that will smooth a time series expressing the distribution to remove chance fluctuations. And there are scale transforms (arbitrary mappings) that will change greater and less variability across the whole set into a homogeneous measure of variation.

1.512 Transformations of this kind, but of unknown number and composition, are now introduced to improve the statistical homogeneity of the output of configurations.

1.52 The basic set of such outputs it was agreed to consider (*vide* 1.36) as stable for a short time (*vide* 1.364) is:

$$[S_a] = \{S_b \dots S_v\} \subset \mathcal{U}(\mathcal{S}) \subset \mathcal{UU}(\mathcal{S}).$$

1.521 At time  $t_0$ , an equivalently ordered set of computable values has been considered, and this is written:

$$\{\hat{x}, \hat{x}', \hat{x}'', \dots \hat{x}^*\} \sim \{S_b, S_c, S_d, \dots S_v\}.$$

1.53 The following table is a first matrix  $m = 1$  of such statistical transformations,  ${}_1\mathbf{T}$ :

$[S_a]$	$S_b$	$S_c$	$S_d$	.....	$S_v$	Configuration set	TIME
$X_a$	$\hat{x}$	$\hat{x}'$	$\hat{x}''$	.....	$\hat{x}^*$	Raw values	$t_0$
${}_1\mathbf{T}_1$	$\hat{x}_{11}$	$\hat{x}_{11}'$	$\hat{x}_{11}''$	.....	$\hat{x}_{11}^*$	First version of first transformation First version of second transformation.	
${}_1\mathbf{T}_2$	$\hat{x}_{12}$	$\hat{x}_{12}'$	$\hat{x}_{12}''$	.....	$\hat{x}_{12}^*$		
.	.	.	.		.		
.	.	.	.		.		
.	.	.	.		.		
${}_1\mathbf{T}_g$	$\hat{x}_{1g}$	$\hat{x}_{1g}'$	$\hat{x}_{1g}''$	.....	$\hat{x}_{1g}^*$	First version of gth transformation	



1.531 The rows refer to successive transformations of the elemental configurations given by the columns. The words "first version" refer to the fact that, given a transformation  $T_g$  whose nature is specified as being  $g$  (for example, to correct skew), there will be many possible versions (for example, the inverse sine) of which only one can be entered in the first matrix.

1.532 Successive matrices,  $m = 2 \dots m = f$ , may now be considered which, adhering to the same kinds of transformation at each row as the matrix  $m = 1$ , vary the interpretation of the actual transform. Thus  $m = 2$  will yield the succession of transforms  ${}_2T_1 \dots {}_2T_g$ ; and finally  $m = f$  will yield  ${}_fT_1 \dots {}_fT_g$ .

1.54 The succession of matrices will form a cubic lattice, of which the following table is the "base":

${}_1T_g$	$\hat{x}_{1g}$	$\hat{x}_{1g}'$	$\hat{x}_{1g}''$	.....	$\hat{x}_{1g}^*$	First version of $g$ th transformation
${}_2T_g$	$\hat{x}_{2g}$	$\hat{x}_{2g}'$	$\hat{x}_{2g}''$	.....	$\hat{x}_{2g}^*$	Second version of $g$ th transformation
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
${}_fT_g$	$\hat{x}_{fg}$	$\hat{x}_{fg}'$	$\hat{x}_{fg}''$	.	$\hat{x}_{fg}^*$	$f$ th version of $g$ th transformation

1.541 It is evident that the original set  $X_a = \{\hat{x} \dots \hat{x}^*\}$  will undergo transformations by finding a set of \* trajectories through the cubic lattice, yielding a mapping of this order-preserving form (for example):

$$TX_a = \{\hat{x}_{2g}, \hat{x}_{1g}', \hat{x}_{fg}'', \dots, \hat{x}_{4g}^*\}.$$

1.55 There are several features of this statistical transformation machine that call for examination:

1.551 We may regard this cubic lattice as a linguistic structure; that is, a formal arrangement in which the specification of three coordinates itself specifies what process occurs at that point in the phase-space. This is a highly structured language, which has a great deal to say about the universe with which it deals without having any *content* (the lattice may be visualized as having empty cells). A computable value following a trajectory through a given cell automatically transforms in passing.

1.552 This leaves us free to regard the content of each cell as a transition probability which influences the trajectory of the computable value. Each probability changes roughly in proportion to the utilization of its cell by the appropriate column trajectory, which provides (in stochastic form) a positive feedback tendency. In short: this is a learning machine. It will acquire paths of facilitation.

1.553 The "slice" of the cubic lattice appropriate to one configuration is itself a matrix (in the third dimension, not depicted in a table). This matrix is a machine for acquiring its own internal statistical homogeneity, and for learning, by facilitation, an optimum pathway. But probabilities rise *asymptotically* to unity: the pathway is never fully determined. On receiving a new stimulus from outside (by mechanisms yet to be discussed), say analogous to pain, new pathways may be facilitated. This means that this portion of the total machine is multiplexing its transforms, and is robust against external disturbance.

1.554 The whole machine is however a cubic lattice, and a mechanism may be postulated whereby it can learn to become homogeneous as between configurations—across the whole set. For the transition probabilities attaching to a particular *version* of certain transformations can be made to interact. Homogeneity in this dimension increases the stability of the whole arrangement, but without sacrificing robustness. For now the multiplexing feature is available at another order of magnitude. What is cohesive and stability-inducing in the enhanced correlation of probabilities between "slices" can at once be turned into new (multiplexing) degrees of freedom if the external stimulus warns the machine of over-specialization and a too inflexible habituation.

1.555 It is understood that certain transformations may turn out, as the result of modifications induced by these external stimuli, to be identity transformations.

1.556 The whole of this mechanism is regarded as extremely important, and it will be invoked in other contexts. For example: we have been working on a stable set of stable configurations (*vide* 1.52), while admitting always (cf. 1.363) the possibility of change. A mechanism such as that suggested here for processing computable values of configurations could also account for these viable features of configuration structure and selection.

1.56 Summarizing the outcome of these operations, the cubic

lattice may be designated as the set  $\mathfrak{M}$  of all possible matrices  $m$  as elements, the set of whose subsets is  $\mathfrak{U}(\mathfrak{M})$ . The lattice we need is  $M$ , a subset of  $\mathfrak{U}(\mathfrak{M})$ , having  $g$  elements of  $m_1 \dots m_g$ . The set of transforms which promote statistical homogeneity will be designated by a *multiple multiplexed transformation* written thus:

$$M\mathfrak{T}_g : \{\hat{x}, \hat{x}', \hat{x}'', \dots, \hat{x}^*\} \xrightarrow{\sim} \{\hat{x}_{\alpha g}, \hat{x}_{\beta g'}, \hat{x}_{\gamma g''}, \dots, \hat{x}_{\nu g^*}\}.$$

1.561 The arrow stands for the mapping process, and is combined with the equivalence symbol to indicate that the mapping is one—one: the ordering is preserved.

1.562 The outcome is a generalization of the exemplification given at 1.541. Note that order is preserved in the  $x$  component itself. The suffixes behave differently; for the first suffix  $\alpha, \beta, \gamma \dots$  is not necessarily monotonic increasing, and the second suffix is always  $g$ —for the lattice cannot absorb a trajectory, and all trajectories emerge at the  $g$ th transformation. (It is conceivable that a value  $x$  representing a variable input should emerge as an invariant, but it must emerge. The lattice might absorb all variety, but it cannot absorb the very existence of a quantified configuration.)

1.60 *Temporal Structure of Computable Values.* The measure set  $X_n$  associated with  $S_n$  has a computable value  $\hat{x}_n$  which can be defined by an algebraic function at time  $t_0$  (*vide* 1.42). This computable value has been investigated above as a member of a set of computable values associated with  $[S_\alpha]$ , still at time  $t_0$ . But consider again a solitary configuration  $S_n$ , and the set of computable values associated with this alone over a period of quantized time. This is the denumerably infinite set nominated in 21.50, which we now define thus:

$$[X_n] = \{\hat{x}_n(t_1), \hat{x}_n(t_2) \dots\}.$$

1.61 The elements of this set will, by the provisions of 1.5, undergo the multiple statistical transformation defined at 1.56, and can be arranged as a frequency distribution using the scale of their own transformed values as variate. For a finite subset of  $[X_n]$  this is simply an isomorphic mapping of the subset onto itself. This section, however, adheres to the order in which they were produced, and considers them as a time series.

1.62 There are two components of information in the time series  $[X_n]$ . The first is the numerical data transmitted at each quantum of time; the second is the structural data transmitted *through* time.

1.621 That is, the data: 2, 8, 3, 7, 2, 9, 4, 8, 1, 6 may have direct numerical significance (their mean for example is 5) which can be abstracted by the brain artefact at this stage and used. But these data have a structural significance independently of their precise numerical value: this brief series appears to have a marked periodicity, which on a full set of data could be examined statistically. The electroencephalographer finds this structural component of information (the brain rhythm) of more importance than either its amplitude or voltage.

1.622 Structural information offers the brain or artefact a new kind of configuration or gestalt, which may be vital to it—especially in matters of prediction. A mechanism is required to discard discrete numerical information (once it has been used) in favour of structural information, which requires emphasizing.

1.63 Now the set  $[X_n]$ , which is well-ordered in respect to time, but disordered with respect to computable value, could nevertheless be ordered relative to computable value (cf. 1.61) equivalently to the set of rational numbers in natural order. It follows that the set is amenable to arithmetic procedures.

1.631 In particular: the addition of two elements of the set  $[X_n]$  defines a third element; this operation satisfies the associative law, and supports the existence of both a null element and an inverse for each element. This set is therefore a group. Furthermore, this arithmetic is certainly commutative.

1.632 The set  $[X_n]$  is therefore an *Abelian Group*.

1.6321 The combinational operation has been given in terms of addition for simplicity; a multiplicative operation could be used if the need arose, although some convention would then be required to exclude zero from the group (since it has no multiplicative inverse) and this might be inconvenient.

1.64 It can now be asserted that the structure-seeking number-discarding transformation required will be a homomorphic mapping  $\phi$  of  $[X_n]$  into the group of rational numbers  $N$ , so that the set  $\phi([X_n]) \subset N$  is a subgroup of  $N$ .

1.641 For any finite subset of  $[X_n]$ , a scale transformation enables us to consider in particular the group of whole numbers, and to direct this homomorphic mapping onto a cyclic subgroup of order  $\alpha$ . It is well known that the remainder set generated by such a mapping is homomorphic with the group of whole numbers.



1.65 Thus the homomorphic mapping required is:

$$\phi([X_n]) : \hat{x}_n \rightarrow R(\text{modulo } \alpha).$$

1.651 There is no aprioristic argument which can establish an optimum value for modulo  $\alpha$ : this will depend on the variability of the input data and the nature of the statistical transformations applied to them.

1.652 The minimum value for modulo  $\alpha$  is manifestly 2.

1.653 Thus it will be circumspect to provide the artefact with a set of possible homomorphic transforms, the optimal *use* of which it can learn by evolutionary trials. These may be depicted in a simple table, since the trajectory of each computable value will now select (not some combination of *versions* of each transform but) one or more transforms which experience reveals to provide structural information.

1.66 This is the table of homomorphic mappings of the set of computable values as already transformed statistically to the form given at 1.56. *See page 50.*

1.67 Familiarly by now, the set of all possible homomorphic transformations  $h$  will be written  $\mathfrak{H}$ , which set has  $\mathcal{U}(\mathfrak{H})$  subsets. The subset chosen at time  $t_0$  will be one of these, written as  $H$ . This set of transforms, which promotes temporal structure at the expense of numerical discrimination, will be designated by a second *multiple multiplexed transformation* written thus:

$$H\bar{T}R_H : \{\hat{x}_{\alpha g}, \hat{x}_{\beta g'}, \hat{x}_{\gamma g''}, \dots, \hat{x}_{\nu g^*}\} \rightsquigarrow \{R_{H1}, R_{H2'}, R_{H3''} \dots R_{H\mu^*}\}.$$

1.671 This is again *multiple*, because more than one transformation may be used. It is again *multiplexed*, because a single homomorphism is strictly sufficient for the structure-seeking purpose, and the alternatives are available to provide a check mechanism leading to robustness under disturbance and the need for adaptation.

1.672 Even so, this set of transformations may be distinguished from the statistical set because, although it has a variety of alternatives from which to select, they are not chosen sequentially but separately. This may be visualized by mapping the table at 1.66 onto the table at 1.54 (rather than that at 1.53). Choose a column of 1.54. At time  $t_0$  the appropriate trajectory has selected *one* cell in that column. This same trajectory, at time  $t_1$ , may pass onto the

$TX_a$	$\hat{x}_{\alpha\beta}$	$\hat{x}'_{\beta\beta'}$	$\hat{x}''_{\beta\beta''}$	.....	$\hat{x}_{\gamma\delta}^*$	Transform configuration set values	TIME
$hT_\alpha$	$R(\text{mod } \alpha)$	$R'(\text{mod } \alpha)$	$R''(\text{mod } \alpha)$	.....	$R^*(\text{mod } \alpha)$	First transformation of highest modulus	
$hT_\beta$	$R(\text{mod } \beta)$	$R'(\text{mod } \beta)$	$R''(\text{mod } \beta)$	.....	$R^*(\text{mod } \beta)$	Second transformation of next highest modulus	
	.	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
	.	.	.	.	.		
$hT_2$	$R(\text{mod } 2)$	$R'(\text{mod } 2)$	$R''(\text{mod } 2)$	.....	$R^*(\text{mod } 2)$	Final transformation of modulo 2	



same column in 1.66, but to any number of cells. This machine therefore has a number of outputs: a set  $H$  of them for *each* configuration, and each set  $H$  being not necessarily the same as the others.

1.70 *The T-Machine* is the brain artefact device which receives, processes and analyses sensory input. The whole of this model to this point has been concerned with its specification in the most general form. Here is a summary of the T-Machine mechanism, beginning with a set of sensory configurations crystallized at time  $t_0$ , and following these inputs through their development to outputs at time  $t_0 + \delta t$ .

$$[S_a] = \{S_b, S_c, S_d, \dots, S_v\} \quad (1.52) \quad t_0$$

$$\sim X_a = \{\hat{x}, \hat{x}', \hat{x}'', \dots, \hat{x}^*\} \quad (1.521) \quad t_0$$

$$\approx M\mathbf{T}_g : \{\hat{x}_{\alpha g}, \hat{x}_{\beta g'}, \hat{x}_{\gamma g''}, \dots, \hat{x}_{\nu g}^*\} \quad (1.56)$$

$$\approx H\mathbf{T}_{R_H} : \{R_{H1}, R_{H2}', R_{H3}'', \dots, R_{H\mu}^*\} \quad (1.67) \quad t_0 + \delta t.$$

1.71 The T-Machine registers the set of all available sensory inputs  $s$ , which are then blended, statistically transformed, and homomorphically mapped onto a set  $[\Xi]$  of outputs  $\xi$ . This set is equivalent to the configuration set, and not to the elemental inputs  $s$ . Similarly the configuration set is equivalent to the output set, and not to the elemental outputs  $\xi$ . For just as a number of inputs  $s$  combine to form an input configuration  $S$ , so does an output configuration  $\Xi$  disperse into a number of outputs  $\xi$ —e.g. different homomorphic mappings using remainders from different moduli. Thus the output value (for example)  $R_{H\mu}^*$  is a set of mappings using (for example) modulo 5, modulo 3 and modulo 2. These values would then be allotted to output channels  $\xi_1$ ,  $\xi_2$  and  $\xi_3$ , forming together the output configuration  $\Xi$ . The same construction applies to output values drawn from higher levels in the T-Machine, which may also map onto  $\Xi$ .

1.711 Thus we have the output configuration defined as:

$$\Xi_b = \{\xi_1, \xi_2, \dots, \xi_n\},$$

1.712 and the set of output configurations for time  $t_0$ :

$$[\Xi_a] = \{\Xi_b \dots \Xi_\nu\},$$

1.713 and the equivalence of structured and well-ordered input/output:

$$[S_n] \sim [\Xi_n].$$

1.72 All processes carried out in the brain and the artefact are bedevilled by Indeterminacy, stochastic behaviour, noise, error, and the arbitrariness of particular transformations. Yet the output is coherent, precise and robust. It is contended that this is achieved by redundancy in the form of multiplexing.

1.721 In the crudest sense, this means the proliferation of connecting links in the neural network, and of nodes by which they are connected.

1.722 But it also means what McCulloch has called "the redundancy of potential command": the ability of any richly interconnected ganglion to assume a central role in any network, and for this centrality to change location freely.

1.723 These mechanisms are reflected in the T-Machine, and it is noted that the outputs are expected to be reliable, despite all forms of disturbance. This, it has been said, is achieved by the learning of paths of facilitation through the machine.

1.73 A necessary part of the behaviour of the T-Machine is therefore that a variety of trajectories should be equifinal. An output  $\xi_n$ , for instance, should be invariant under many different transformations of a particular subset of  $\mathcal{S}$ . This is the formal condition for a satisfactory empirical exemplification of a T-Machine.

1.731 It is this formal condition which makes it possible to envisage a brain that works, or an artefact that is constructable. For it means that a T-Machine can in principle escape from the limitations of arbitrary decisions taken during its construction.

## 2. ACTIVE SENSIBILITY

2.00. *General.* The model has so far dealt with afferent impulses and a cerebral mechanism for sensation. It must now pass to efferent impulses and the notion of activity.

2.10 *The Motor Cortex* is visualized as complementary to the sensory cortex and disposed (as it is in the brain) in close association with it. Thus a central sulcus is envisaged as separating the postcentral gyrus (the sensory cortex) from the precentral gyrus (the motor cortex).

2.101 A knowledge of recent research into cortical activity will be assumed at this point. For the task of introducing modifications into the model that would dispel the crudity of the previous

assertion would be disproportionate to the general level of refinement in the model at large.

2.102 Thus, for example, an interaction between the sensory and motor areas will be taken for granted insofar as it is known that "20 per cent of stimulation experiments on the cortex of the postcentral gyrus results in movement rather than in sensation"—Sholl.<sup>(8)</sup>

2.103 But this simplification is no greater than that which underlay much neurophysiology until recent times; and it is hoped that the model can be developed in a way which, while not formally recognizing these discoveries, will not actually conflict with them.

2.104 The following postulate is made as, *inter alia*, offering ready means in the future for studying such sensorimotor ambiguities in brain function.

2.11 It is postulated that there is a set-theoretic logical equivalence between the sensory and the motor cortical architectonic.

2.111 *Sensations* will be regarded as paralleled by directions; and the sensory channels hitherto designated  $s$  are equivalent to directive channels (or directives)  $d$ , each of which is an element of the set  $\mathfrak{D}$ .

2.112 Note: the numbering of these paragraphs now follows that of section 1.00. Thus a full understanding of (for example) 2.12 implies reference to 1.12.

2.12 To every element  $d$  of  $\mathfrak{D}$  there corresponds a measure set  $\langle x \rangle$ .

2.13  $\mathfrak{D} = \{d_1, d_2, d_3, \dots, d_n\} \sim \mathfrak{S}$ .

2.131 Directives deriving from the organism itself belong to the segment  $\mathfrak{D}_{a_v} (\sim \mathfrak{S}_{s_v})$  determined by  $d_v$ .

2.132 Directives deriving from the environment belong to the complement of  $\mathfrak{D}_{a_v}$  in  $\mathfrak{D}$ , namely the remainder subset.

2.133 Thus, for example:

$$\begin{aligned} d_1 &\in \mathfrak{D}_{a_v} \\ d_n &\in \mathfrak{D} - \mathfrak{D}_{a_v} \end{aligned}$$

2.20–2.24 *Motor Configurations*. The arguments of 1.20 to 1.24 apply, *mutatis mutandis*.

2.25 A motor configuration is a subset of  $\mathfrak{D}$ , namely  $D$ , of elements  $d$ .

$$D = \{d_\gamma, d_\delta, \dots, d_\theta\}$$

for all subsets of  $\mathfrak{D}$ .

2.26 The totality of possible motor configurations is the set of the subsets of  $\mathfrak{D}$ , namely  $\mathfrak{U}(\mathfrak{D})$ .

$$2.27 \mathfrak{U}(\mathfrak{D}) = \{D_1, D_2, D_3, \dots, D_{2^n}\} \sim \mathfrak{U}(\mathfrak{S}).$$

2.30 *The Fourth Layer of the Motor Cortex* is at least partly defined by a set of many-one mappings of  $\mathfrak{D}$  into  $\mathfrak{U}(\mathfrak{D})$ . One such mapping is  $D_b$ , a motor configuration.

2.31 At time  $t_0$ , there exists in the fourth layer a particular set  $[D]$  of subsets  $D$  of the set  $\mathfrak{D}$ , equivalent to  $[S]$ :

$$[S] \sim . [D] = \{D_b, \dots, D_v\} . \subset \mathfrak{U}(\mathfrak{D}) \sim \mathfrak{U}(\mathfrak{S}).$$

2.32 And the total possible states of the fourth layer is given by the set of the sets of the subsets of  $\mathfrak{D}$ :

$$\mathfrak{U}\mathfrak{U}(\mathfrak{D}) = \{[D_1], [D_2], [D_3], \dots, [D_{2^n}]\} \sim \mathfrak{U}\mathfrak{U}(\mathfrak{S}).$$

2.33 The arguments of 1.33 to 1.672 apply, *mutatis mutandis*.

2.672 The range of transformations  $T$  are now paralleled by an equivalent range  $V$ .

2.70 *The V-Machine* is the brain artefact device which receives, processes and analyses motor input. Its mechanism is summarized by equivalence with the T-Machine as follows:

$$\begin{array}{ll} [D_a] = \{D_b, D_c, D_d, \dots, D_v\} & t_0 \\ \sim X_a = \{\hat{x}, \hat{x}', \hat{x}'', \dots, \hat{x}^*\} & t_0 \\ \rightrightarrows M \vee g : \{\hat{x}\alpha_g, \hat{x}\beta_g', \hat{x}\gamma_g'', \dots, \hat{x}\gamma_g^*\} & \downarrow \\ \rightrightarrows H \vee R_H : \{R_{H1}, R_{H2}', R_{H3}'', \dots, R_{H\mu}^*\} & t_0 + \delta t. \end{array}$$

2.71 The output of the V-Machine is homomorphically mapped onto a set  $[Z]$  of outputs  $\zeta$ . This set is equivalent to the configuration set  $[D]$ , but the inputs  $d$  are not equivalent to the outputs  $\zeta$  (cf. 1.71).

2.711 Thus the output configuration is defined as:

$$Z_b = \{\zeta_1, \zeta_2, \dots, \zeta_n\},$$

2.712 and the set of output configurations for time  $t_0$ :

$$[Z_a] = \{Z_b, \dots, Z_v\},$$

2.713 and the equivalence of structured and well-ordered input/output:

$$[D_n] \sim [Z_n].$$

2.72–2.731 The arguments of 1.72 to 1.731 apply, *mutatis mutandis*.

## 3. INTEGRATION AT THE THALAMIC LEVEL

3.00 *General.* Both the T-Machine and the V-Machine are devices for organizing experience: afferent and efferent experience respectively. Each is capable of organizing this experience under two modes: organismal and environmental. This fourfold system must clearly be intricately and intimately balanced.

3.01 Notwithstanding possible interactions at the cortical level itself (cf. 2.102) a deeper level integration is required, and this is postulated to occur at the thalamic level.

3.10 Four output sets will thus be considered, reflecting the sensory organismal  $[\Xi_0]$  and environmental  $[\Xi_E]$  activities, and the corresponding motor activities  $[Z_0]$  and  $[Z_E]$  of the cortex.

3.101 The system incorporating these sets is clearly some form of Ashbean ultrastable machine,<sup>(1,9)</sup> and Ashby's lead will be followed in the use of Bourbaki's set notation<sup>(10)</sup> at this point.

3.1011 This permits consideration of a "product" of the four sets that is neither an orthodox union nor an intersection but an ordered "superset".

3.102 Output configurations are the elements of the four sets defining the system, which is written:

$$[\Xi_0] \times [\Xi_E] \times [Z_0] \times [Z_E] = U_{t_0}.$$

3.103 The element  $u \in U_{t_0}$  is a structural state of the whole system, designating an ordered quadruplet of elemental configurations:

$$u_{t_0} = (\Xi_0^n, \Xi_E^n, Z_0^n, Z_E^n).$$

3.104 It remains to decide whether the T-Machine is in fact capable of distinguishing between organismal and environmental subsets of its output. These aspects were distinguished in  $\mathfrak{S}$  as the segment determined by  $s_v$  and the remainder subset, but there is in practice considerable ambiguity in allocating inputs to these destinations. The T-Machine itself will similarly involve ambiguities in attempting to preserve the distinction, on account of the law of indeterminacy of configuration structure (*vide* 1.43). Even so, the division can be maintained in  $\mathfrak{S}$  for most  $s$ , and in  $\Omega(\mathfrak{S})$  for most  $S$ ; thus an approximately accurate distinction can be preserved for most  $\Xi$ .

3.1041 Most dichotomies of function in the real world turn out to be ambiguous over at least some area of the phase-space, and



this is especially so in the brain (cf. sensory and motor cortex). To reject a scheme of classification on this ground is neurotic, not rigorous science, if it holds in the main, is useful, and is not treated as if it admitted of no exceptions.

3.1042 In this case especially, the cortex is at least partly localized (cf. the well-known cortical homunculi of Wilder Penfield), and the segments under discussion are not entirely logical abstractions but have a certain unity in the brain and the machine.

3.20 *The U-Machine* is an assembly competent to handle the four-fold superset  $U$  of quadruplet elements  $u$  in such a way that the organism and the environment are and remain in mutually acceptable states.

3.21 To this end, the U-Machine has a set of outputs  $\Omega$  that represents all possible interactions of the brain with the state of the world—which itself provides the inexhaustible variety feeding in through the T- and V-Machines.

3.22 Efferent impulses themselves originate in the motor cortex, and change the behaviour of the organism directly (“move left arm”; “switch off second engine”) and of the environment indirectly (“open the door”; “write to customer”). The U-Machine *monitors* all these activities in relation to the sensory gestalten. Thus the output of the U-Machine is a monitoring signal that feeds back approbation and disapprobation, through the world picture itself, into its own input.

3.221 The output set  $\Omega$  is thus a *reward function*. It tends to reinforce input patterns conducive to survival, and to break down patterns that are not. This is achieved in a *directed* rather than a random way, through the highly conditioned T- and V-Machines.

3.2211 This mechanism greatly modifies the concept of random mutation that informs the Ashbean homeostat and Neo-Darwinian genetics alike. It offers a procedure for constructing in the T- and V-Machines something analagous to the mechanism Waddington<sup>(11)</sup> has invoked for embryological development: an “epigenetic landscape”.

3.23 The U-Machine may be described thus at time  $t_0$ :

$$U_{t_0} = [\Xi_0] \times [\Xi_E] \times [Z_0] \times [Z_E] \quad (\text{vide } 3.102),$$

3.231 the element of  $u \in U_{t_0}$  is a structural state:

$$u_n = (\Xi_{0n}, \Xi_{En}, Z_{0n}, Z_{En}).$$



3.232 and the elements  $\hat{u}_n$  is a computable value corresponding to  $u_n$ , whose elements are computable values  $\hat{x}$  of the ordered quadruplet of the elemental configurations of  $u$ :

$$\hat{u}_n = (\hat{x}_{A_n}, \hat{x}_{B_n}, \hat{x}_{C_n}, \hat{x}_{D_n}) \sim u_n.$$

3.2321 The computable value  $\hat{\chi}_n$  is analogous to  $\hat{x}_n$  in 1.4. In particular,

$$\hat{\chi}_n(t_0) = f(\chi, \chi', \chi'' \dots \chi^*)$$

where

$$\Xi_n(t_0) = \{\xi, \xi_1, \xi_2, \dots \xi_n\} \sim \{\chi, \chi', \chi'', \dots \chi^n\}$$

and

$$\chi = x \vee x_{\alpha g} \vee R_{H_1} \quad (\text{vide } 1.70)$$

as determined by the T-Machine.

3.24 The U-Machine may be described thus in general:

$$U = \bigcup_{t \in I} U_{t_i}$$

the union of the sets  $(U_{t_i})_{i \in I}$  where  $I$  is the set of all possible states through time of the Bourbakian product of the four sets of output configurations (cf. 3.23).

3.241 The element  $u \in U$  is a structural state:

$$u_{t_0} = \bigcup_{n \in J} u_n(t_0) \quad \text{and} \quad [u] = \bigcup_{n \in J} [u]_n$$

where  $J$  is the set of all possible states through time of each output configuration (cf. 3.231), and  $[u] = ([\Xi_0], \text{etc.})$ ,

3.242 and the element  $\hat{u}$  is a computable value for the whole system corresponding to  $u$ :

$$\hat{u}_{t_0} = \bigcup_{n \in J} \hat{u}_n(t_0) \quad \text{and} \quad [\hat{u}] = \bigcup_{n \in J} [\hat{u}]_n$$

where  $[\hat{u}] = ([\hat{\chi}_A], \text{etc.})$ .

3.25 Throughout this formulation, time has been treated as quantized. To discuss the operation of the U-Machine, its state will be described at  $t_0$ , and time will then be advanced by quanta.

3.251 There is a computable value  $\hat{\chi}$  referring to any output configuration (vide 3.232). The subscripts  $A, B, C, D$  have already been used to signify the four parts of the ordered quadruplet, and these are maintained in what follows. We now consider the full

sets of configurations at time  $t_0$  (e.g.  $[\Xi_0]$ ) rather than each configuration separately (e.g.  $\Xi_{0n}$ ), and the set of computable numbers  $\hat{\chi}$  referring to each such set will be written  $[\hat{\chi}_A]$ , which is the quantized state of the assembly

$$A = \bigcup_t [\Xi_0]_t.$$

3.2511  $[\hat{\chi}]$  is strictly a vector of the set of configurations, but it is convenient to regard it as a single value. This is perfectly possible; for example  $[\hat{\chi}]$  is itself a computable value if every set of values  $\hat{\chi}$  is allotted a unique Goedel number.

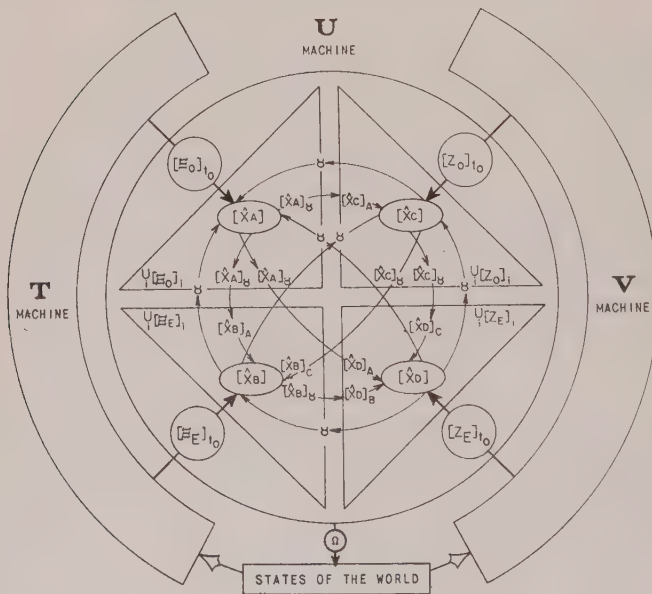


FIG. 2

3.25 Operation of the U-Machine. The U-Machine (*vide* Fig. 2) is an Ashbean homeostat.<sup>(9)</sup> A selection of its more important couplings is shown in the figure. These are similar in form, and it will suffice to discuss the operation of one coupling: that between the motor organism (inputs  $[Z_0]_{t_0}$ ) and the sensory organism (inputs  $[\Xi_0]_{t_0}$ ).

3.251 The motor organism has a set of states

$$\bigcup_i [Z_0]_i (= C),$$

and at time  $t_0$  is in the state  $[\hat{\chi}_C]$ . The sensory organism is in the state  $[\hat{\chi}_A]$ , an element of its set of states

$$\bigcup_i [\Xi_0]_i (= A).$$

The motor state will be assumed to be different from the sensory state: that is, the brain is proposing that the organism should act and will only implement this intention if the proposed state is acceptable to the sensory cortex (with its learnt patterns and ability to forecast).

3.252 To do this,  $C$  must inform the sensory cortex  $A$  that it regards the intended state  $[\hat{\chi}_C]$  as favourable to itself. This is judged from internal criteria: the learning of the V-Machine.  $A$  then decides, by its own T-Machine criteria, whether the state is acceptable or not.

$$\begin{aligned} 3.2521 \quad [\hat{\chi}_C] &\rightarrow \gamma_1 && \text{for an unacceptable state} \\ &\rightarrow \gamma_2 && \text{for an acceptable state} \end{aligned}$$

where survival value is measured by the set  $\Gamma = \{\gamma_1, \gamma_2\}$ , providing the operation  $g$ .

3.253 There is a pair of mappings of  $A$  in  $A$ :  $[\hat{\chi}_A]_{\gamma_1}$  and  $[\hat{\chi}_A]_{\gamma_2}$  such that the  $g$  operation gives an invariant point for an acceptable state:

$$[\chi_A]_{\gamma_2}([\hat{\chi}_A + 1]) = ([\hat{\chi}_A + 1])$$

where time has advanced one quantum at the  $g$  operation, and the new state of  $A$  is written  $[\hat{\chi}_A + 1]$ . For an unacceptable state there is no invariant point:

$$[\hat{\chi}_A]_{\gamma_1}([\hat{\chi}_A + 1]) \neq ([\hat{\chi}_A + 1]).$$

3.254 There is a set of mappings of  $C$  in  $C$ :  $[\hat{\chi}_C]_A$  which adjusts  $C$  towards  $A$  if it is invoked. It will be invoked if  $A$  changes state, otherwise not.

3.255 The state of this (simplified) U-Machine  $C \times A$  at time  $t_0$  is defined by the ordered couple:

$$\hat{u} = ([\hat{\chi}_C], [\hat{\chi}_A])$$

the value of which element changes at each quantum of time during which the operations  $[\hat{\chi}_A]_\gamma$  and  $[\hat{\chi}_C]_A$  are induced simultaneously.

3.256 This is the transformation:

$$([\hat{\chi}_C], [\hat{\chi}_A]) \rightarrow ([\hat{\chi}_C]_A([\hat{\chi}_C + 1]), [\hat{\chi}_A]_{g([\hat{\chi}_C + 1])}([\hat{\chi}_A + 1])).$$

3.257 The equilibril state is reached when the whole machine  $C \times A$  is quiescent, which may be defined as:

$$[\hat{\chi}_A]_{g([\hat{\chi}_C + p])}([\hat{\chi}_A + q]) = ([\hat{\chi}_A + q])$$

so that

$$g([\hat{\chi}_C + q]) = \gamma_2$$

or

$$[\chi_C + q] \in {}^{-1}g(\gamma_2)$$

after  $q$  quanta of time passed.

3.258 Note: Ashby remarked <sup>(9)</sup>—corrected version) that his mechanism should be developed to use a random variable; but the machine he was describing, once having been disturbed, had no further input and was left to reach equilibrium. It will be appreciated that in the U-Machine random perturbations arrive continuously from the world via the T- and V-Machines, and can be handled.

3.259 At the equilibril state, the machine is given as:

$$[\hat{u}]_{t+q} = ([\hat{\chi}_C + q], (\hat{\chi}_A + q)).$$

This expression denotes the *preferred state of the system* for given  $T$  and  $V$  input configuration sets.

3.2591 To extend this operating description to cover all possible couplings of the machine would be firstly to write down the whole of this section (3.25) many times with different symbols—a work of supererogation indeed. Secondly, however, it would require that all the operations by which the states of each assembly are mapped onto themselves be traced right through the whole system. It is easy to *visualize* the reverberating effect of the modification of one coupling throughout the system, but difficult to formalize it. For instance, although a *uniform* reverberation throughout the system can be visualized, it is clear that some pathways will be differentially blocked by vigorous local action which will engulf the tenuous

effects of distant activity. Thus I tentatively propose that the reverberations could be treated as random walks through the network of operations that induce the critical mappings; they might be treated formally as Markov chains.

3.26 *The Output of the U-Machine* is a set  $\Omega$  of elements  $\omega$  onto which are mapped the succession of Bourbakien elements  $[\hat{u}]$  of the product set which stands for the whole system (given for  $t_0$  at 3.23).

3.261 Thus the set of elements  $[\hat{u}]$  determining the U-Machine and the set of elements  $\omega$  determining the output are linked through their mutual equivalence to

$$\prod_{t \in K} [\hat{x}]_t$$

where  $K$  is any possible assembly (such as  $[\Xi_0]$ ) included in a U-Machine.

3.262 Just as the vectors of the configuration sets  $[\hat{x}]$  can be regarded as unique Goedel numbers (*vide* 3.2511), so can their product. Thus the monitoring signal carried by the set  $\Omega$  has a measure corresponding to each element  $\omega$ . The channel capacity required for the transmission of  $\Omega$  must be very high indeed, and an estimate of the variety of this measure ought to be obtainable.

3.2621 In the T-Machine, all possible states of the fourth cortical layer are given by  $\mathcal{U}(\mathcal{S})$  of cardinal number:

$$2^{2^{|\mathcal{S}|}} \quad (\textit{vide} 1.32)$$

and in the V-Machine the equivalent cardinal is:

$$2^{2^{|\mathcal{D}|}} \quad (\textit{vide} 2.32).$$

3.2622 But  $\mathcal{S} \sim \mathcal{D}$  (2.13) so  $|\mathcal{S}| = |\mathcal{D}| = n$  where  $n$  is the total number of possible inputs. Therefore the maximal *structural* variety from the two input machines which converges on the U-Machine is:

$$2(2^{2^n}).$$

3.2623 However, the information carrying capacity of this

population of input networks is affected by the infinitely denumerable measure sets each of cardinal number  $\aleph_0$ . Moreover, the U-Machine proceeds to select from both input sources conjointly. The expression for the channel capacity required for output is elusive (but nonetheless should be susceptible to statement).

3.2624 If this expression could be determined, and since the approximate channel capacity for the brain's actual input and output mechanisms can be estimated (cf. Bowman<sup>(12)</sup>), then an approximate measure of variety reduction in the brain could be obtained. This is interesting because it would throw light on the modulus used by the T- and V-Machines in reducing continuous measuring to discrete scales (attempted calculations suggest, for example, that the transfinite cardinal must in practice be reduced to a cardinal of 4 or 5). It might also explain (via pattern recognition via sensory configurations) the severe limitation on human ability to discriminate between more than a few (say five) points on a scale of subjective judgment. Such arguments seem important, though not strictly relevant to present purposes.

#### 4. MONITORING AT THE RETICULAR LEVEL

4.00 *General.* Without intending to minimize the importance of the cortex itself, there is much to be learnt from recent work on the reticular formation of the brain stem. This must have its place in the model and the artefact, although little can be said at present.

4.10 *The ascending reticular formation* must surely accept an important segment of the output set  $\Omega$ , as well as direct cortical connections. In the organism it has the task of mediating hypothalamic and bulbar control of visceral function, and the monitoring output of the U-Machine would certainly be adapted to an analogous function in the institutional brain artefact.

4.11 *The ascending reticular formation* offers precisely the mechanism required for arousal of the T-U-V system which, in the absence of significant cortical activity must be quiescent. It is clear that in a "routine" situation the T- and V-inputs to the U-Machine would be fairly static; the U-Machine would find and



retain an equilibrium, and the entropy of the whole system would steadily rise. A decreasing tendency for anything at all to happen would be revealed. An input signal referring to danger (for example) could then be accepted at the cortical level and engulfed: by this time, it is envisaged, the T-U-V system would be *habituated*; it would be cycling at a steady sleepy rhythm. But, since the ascending reticular system receives collaterals from afferent paths, and if it is capable of monitoring these afferent impulses (as it appears to be), a danger signal could be fed straight into the U-Machine and destroy its equilibrium.

4.12 These considerations, which are eminently appropriate to this theory, assume a knowledge of the work of such people as Magoun.<sup>(13)</sup> They connect importantly with the idea (*vide* 1.722) of "the redundancy of potential command", and McCulloch has explicitly referred this name to the reticular formation. But I have not succeeded in pursuing the formal statements of this general theory into this area, because the anatomical and physiological picture hitherto accepted is undergoing such rapid and drastic change.

4.121 For example, the diffuse organization of the reticular formation for which a preliminary model was constructed turns out to be a myth. It is "subdivided into several regions which differ with regard to their cytoarchitecture, fibre connections, and intrinsic organization" according to Brodal,<sup>(14)</sup> regions which are certainly interdependent.

4.13 *The R-Machine*. Even so, and even in the absence of any formal account, an R-Machine will be postulated and is constructable. For it is certainly possible to envisage control loops which cannot be closed within the T-U-V system. The R-Machine will be defined in a preliminary way as the mechanism for such closure.

4.131 Gastaut<sup>(15)</sup> has for example proposed that conditioned learning (with which a useful brain artefact must be concerned) cannot be closed within the cortex, as experimental cortical lesions have not prevented closure. If the afferent collaterals already mentioned excite the reticular formation for an unconditioned stimulus, and there are connections between the reticular formation and the cortex as well at the thalamic level, then a subtle conditioning mechanism is available. Instead of the establishment of direct facilitated pathways in the cortex itself, there is a thalamo-cortical

mapping onto the R-Machine which is injecting the random perturbations of unconditioned stimuli into the U-Machine. Thus the process of conditioning becomes a homeostatic struggle for dominance between an organized and a random response at the thalamic level, which will eventually be settled by an equilibrium which perfectly measures the relevance of the conditioning to the general experience of the organism.

4.1311 This is my own interpretation for this theory of Gastaut's proposal, although the R-Machine mechanism shown in Fig. 3 is Gastaut's own model imposed on mine. It does not show the pathways through the association cortex which Gastaut shows, but these are implicit in my definition of the T- and V-Machines.

4.132 Mention was made (at 3.221) of a reward function which was conceived as a role of the output  $\Omega$ . This is another candidate for monitoring at the reticular level which can be made firm in a specification for a constructable R-Machine (despite its unformalized structure). It is an R-closure alone that can stimulate or inhibit the "metabolism" of the U-Machine.

4.1321 This thought derives especially from the arousal mechanism of the ascending reticular formation (4.11) and also the conditioning mechanism (4.131).

4.1322 Moreover, the mechanism of pain seems important as implying an inverse reward. Central pain is readily contained within this model: cf. the evidence adduced by Noordenbos<sup>(16)</sup> that lesions occurring anywhere along the afferent pathways in the neuraxis can cause spontaneous pain without stimulation of the area in which it is localized. Visceral and referred pain can also be visualized in the model. But in view of the arguments above, it may be that pain arising from peripheral receptors presents no special problem to the model either, as might at first appear.

4.133 Thus we are led to consider an extension of the concept of a reward function already given, and to consider in the theory an *algedonic control system* which would have wide implications for all the functions of the artefact. This is a tentative proposal, and cannot be elaborated here.

4.2 A general picture of the whole theory, showing how the parts fit together, is provided at Fig. 3. The temptation to make the outline look like a coronal section of the living brain was irresistible and I apologize to cerebra everywhere for such insolence.



FIG. 3



## PART III

*From the Special Theory:**An Exemplification of the T-Machine and Prospectus  
of the Artefact*

## 1. THE CHAIN OF SYSTEMS

The company is a manufacturer of steel rods. It takes billets of steel ( $2\frac{1}{4}$  in<sup>2</sup>, in various steel qualities) and rolls them down to a small round cross-section (in a variety of sizes). There are two strands to the mill: that is, two lots of steel run side by side through the mills. The rods, which are red-hot, emerge at a speed of approximately 70 m.p.h., and are then coiled, cooled and inspected for despatch.

The basic plant consists of a heating furnace for the billets, and a means of feeding them into the mill; three interconnected trains of rolling mills (roughing, intermediate and finishing) arranged like a flattened letter S, the steel looping round the two bends in order to change direction twice; suitable coiling mechanisms capable of dealing with a continuous output from the mill; machines for removing the coils and circulating them on a conveyor while they cool; and a means of retrieving and stacking them thereafter. This is the "producing system" shown on Fig. 4.

This producing system is connected to the outside world through two stocking systems. The input stocks consist of billets, and buffer the producing system against the vagaries of the environmental supplying system. The output stocks consist of finished rods, and feed the environmental consuming system—the market.

The couplings within the company are quasi-deterministic. Input stocks must eventually be fed into the producing system to become output stocks. But the quantities and qualities to be charged are subject to decision procedures, and these themselves do not specify which individual billets (or rods) are to be used. The two environmental couplings, however, are entirely probabilistic: there is no guarantee about supply (from steelmaking companies) nor about demand (from the market). These couplings have nonetheless been

intensively studied by operational research methods, and can be effectively described as stochastic processes.

Feedback loops operate throughout this chain of systems, and the four obvious ones are shown in the diagram. More complicated feedbacks, which do not involve the plant itself, couple the supply and demand systems directly, in that potential changes in either require consultations with the other before they become actual.

## 2. A CONVENTION OF THIS EXPOSITION

We are dealing with an actual and particular exemplification of the T-Machine which, as explained in Part I, has been created for research purposes. This means that a large number of decisions has had to be taken about the structure of the "sensory cortex"—decisions which, ultimately, the brain artefact itself is intended to take by its multiple multiplexing techniques. The research team in the field has, however, taken these decisions on an informed basis, by operational research methods. In short, what has been done is an O.R. study of the company; and what has resulted is an O.R. model of both the company and the T-Machine.

It would be wearisome continually to qualify statements made here with accounts of how each statement should be expressed in T-Machine terms. This has already been done in Part II. All that is needed is a single device for indicating the points which have been taken as "known", following the operational research, which could not in fact be known without an actual T-Machine. For example, in this first exemplification there are nineteen sensations. There are not "really" nineteen sensations; as a matter of fact, further experimental exemplifications have already brought the number of sensations considered in this work up to thirty-six, and there is nothing absolute about that number either.

To show exactly where the argument stands in this respect, the following convention is adopted. For this first exemplification, there *are* nineteen sensations: I "solemnly declare" this to be the "correct" number; which means "I pretend that a T-Machine is in operation that turns out to be using these nineteen sensations". The ordinary, if uncommon, English word "asseverate" means "solemnly declare", and will be used here as a technical term having



the meaning just explained. It is claimed that all asseverations made in this Part can be eliminated from the picture by mechanisms formally described in Part II.

### 3. THE EXEMPLIFICATION OF THE T-MACHINE

It is asseverated that there are nineteen sensations, eleven deriving from the company and eight from the environment. These are named in the small circles of Fig. 4. Arrangements were made to measure these sensations continually, and a large amount of data was (and still is) collected.

From twelve of these sensations an adequate accounting system was devised, representing a model of the conscious process of deducing facts from information available. The costing system is regarded as reflecting the company's internal operations: as the diagram shows, environmental information determines the input costs from which the total cost is generated by the plant; and this total is compared with further environmental information about the sales value of the product to yield a profit (or loss). The financial accounts, however, regard the organism from the viewpoint of the environment. Here the sums of assets and liabilities as judged by the outside world are compared to yield a balance-sheet surplus (or deficit). Accounting is the conscious level of deduction about the state of affairs. (It involves no asseverations, for it is itself conventional. Nevertheless, the balance-sheet is accepted by business and the law alike as a measure of survival-value, in which it somewhat resembles a proposal form for life insurance as a measure of longevity.)

The exemplification then proceeds to elucidate the processes of judgment, which do not rely on conscious deduction. This is the level of sensory configuration in the brain model, and constitutes the T-Machine proper.

It is asseverated that there are twelve sensory configurations, and that six of these are primarily concerned with the company, and six with the environment. Each one blends together a number of sensory inputs. It is asseverated that the appropriate inputs are known, and that a robust mathematical function defining each is available.

The six company configurations are quantified by the following asseverated functions, on each of which a few notes are appended:

(a) *Arrival function*. The optimal level of arrivals is defined *post hoc* as the level of actual billet consumption, and the mathematical model was devised to measure the success of actual arrivals as a forecasting procedure.

(b) *Billet stock utility function*. This function compares actual stocks at this moment with optimal stocks calculated from an O.R. model which makes the necessary distinctions between classes of steel. Stock utility is diminished if the company is either under or over insured against possible mishaps. Thus this function sinks below its ideal value of unity in either case.

(c) *Activity function*. This is based on a complicated model of production, and seeks a pure measure of efficiency when all the variable factors have been taken into account. It is capable of distinguishing between technical and human reasons for changes in the level of activity.

(d) *Yield function*. This again computes the optimal yield in every case, taking into account the losses that are inevitable from oxidation and the cropping of both billet and coil ends, and compares this with the actual yield.

(e) *Rod stock utility functions*. Rod stocks arise from a number of causes: the need for inspection, rejected material, orders cancelled, the optimal buffers required to meet delivery promises. There is also a complicated component of the stock due to a cyclical effect in matching rolling programmes to demand. All are modelled.

(f) *Departures function*. Basically, this measures the complementarity between activity and despatch. But the function is loaded by the complement of the ratio between carriage costs for each order (which are borne by the company) and internal profit. It therefore provides a judgment from the point of view of the organism about the exploitation of the effective market.

The six environmental configurations are likewise quantified by asseverated functions, on which these notes are offered:

(g) *Supply function*. A complicated model of supply measures the efficiency of delivery from each supplier, in terms of both average reliability and variability about the average, and includes a measure of the quality of material supplied.

(h) *Expenses function*. This is a measure of the economic utility of "overheads"—non-production expenses.

(i) *Plant function*. By measuring the changing availability and price of new plant, this function expresses the environmental desirability of investing in new plant at any given moment.

(j) *Money function*. This function is also based on measures of change in both availability and price, this time of money itself. It reflects the level of capital expenditure in the industry, and includes a model of the (local) effects of change in bank rate.

(k) *Labour function*. Again, the model is based on availability and cost. Absenteeism and the unemployment level are taken into account.

(l) *Demand function*. This is the most complicated function of all, and attempts to measure the market and its profitability from all relevant input sensations and deductions. The amount of forward ordering, the profitability of each order (adjusted for such matters as carriage costs, rebates and quantity allowances), and "the solidity of the order book", are all involved

The full O.R. models asseverated for each of these twelve functions are not given here: they would be out of place. But enough has been said to show the scope of the exemplified T-Machine, the method of blending information inside a configuration to provide commensurate pure numbers as output, and the fact that the functions all depend on ratios (it is asseverated) of some measure of expected behaviour to the actual behaviour.

This last point is important, since it incorporates in this exemplification the essential "black box" treatment of unknowns and imponderables common to all cybernetic machines. For a model of performance in any field may be inadequate: predictions and judgments based upon it will be effectual only insofar as the model *is* adequate. But in exceedingly complex and probabilistic systems *no* analytic model can possibly be adequate. The answer to this paradox, which I have used successfully for ten years, is to load the raw predictions of any analytic model with a continuous feedback measuring its own efficiency as a predictor. In this way everything that went unrecognized in the analytic work, everything that proved too subtle to handle, even the errors incurred in making calculations, is "black boxed" into an unanalysable weighting which is error-correcting.

Values for these twelve functions are computable in principle every eight hours, although in some cases (e.g. the Money Function) changes must occur much less frequently. At present, practical O. & M. work has resulted in a reasonably efficient clerical system for collecting these data, and for processing them. (In future, it is hoped to obtain many of these inputs automatically.) The result is that a set of twelve values becomes available, in present practice, every day. These are plotted on boards in an Operations Room for the benefit of management, as a by-product of this research, and various orthodox methods are used to analyse them, as mentioned in Part I.

So far, this account of a first exemplification has considered the operation of a T-Machine to the level of the fourth layer of the sensory cortex. This gives rise, it is asseverated, to an "encephalogram" of twelve readings. A sample of such a trace appears at Appendix I, where a whole year's actual information is recorded by daily plots.

After this, it is asseverated that three statistical transformations are necessary. Firstly, a transformation is made to improve the statistical homogeneity of the recorded values of each function, according to what is asseverated to be a transformation that corrects each distribution of sample values towards the Gaussian form. Secondly, a transformation is imposed to improve statistical homogeneity across the set of functions, and scale transforms are obtained which can be asseverated to have this effect. Thirdly, chance fluctuations are removed from the time series by filtering each of them through a criterion of noise, the level of probability which is significant being asseverated by the study of samples.

The operations of the T-Machine are completed, as proposed in Part II, by the consideration of a set of group homomorphisms preserving the structure of each statistically transformed time series as against its scalar information. To this end it is asseverated that three different homomorphic transformations will prove valuable. Two of them, giving remainders modulo 5 and modulo 2, are shown, for the original year's data, at Appendices 2 and 3. These are final outputs of this exemplification of the T-Machine, to which an exemplified set of possible transformations has been applied. Formal comparisons of the information conveyed by the original encephalogram (Appendix 1) with the final forms now



proposed are still going on, and the actual records are provided in this case to facilitate informal comparisons.

#### 4. THE NEXT STEP IN THE EXEMPLIFICATION

The value found for each function for each time quantum is an estimate, which may be depicted as the mean of a probability distribution. As time passes, a collection of such distributions is stored away for each function, forming a quantized chronological memory of relevant experience. The whole set of these records for the company functions on the one hand and the environmental functions on the other constitutes a generalized gestalt memory, as depicted in Fig. 4.

It is next asseverated that although changes in mean value through time within these gestalten may be important, may be learnt, and may be optimized, by the T-Machine, the main component of judgment lies in a function of statistical variance. This is a mechanism at present under close study, and further empirical insight into it will eventually become available through the operations room already mentioned.

Meanwhile, the rest of the artefact proposed in Part II may be exemplified by the model homeostatic system depicted in Fig. 4. This shows a simplified form of the interactions taking place in the U-Machine, and deals only with the company-environment homeostat derived from the sensory configurations. The range of behaviour through time of each generalized gestalt memory defines two phase spaces in which the company and the environment can respectively operate. It is clear that a set of preferred states is learnt by each: a learning transform applied to the vector provided by the functions "at this moment" will map the set of solutions into itself, seeking invariant points which will define the two preferred states sets (depicted in Fig. 4 as  $n$ -dimensional polyhedra). The formal mechanism for this process has been given in Part II, as also the mutually vetoing system by which the homeostatic loop in the diagram continues to operate until both company and environmental points in phase-space (representing vectors of functions) lie in the appropriate preferred states set.

The operations of the R-Machine, with its arousal mechanism,

reward function, and so on, are depicted in Fig. 4 as an "algedonic control" exciting or inhibiting the activity of the homeostat. (The correlate in consciousness of R-Machine activity is surely correctly suggested by the diagram to be *feeling*, rather than sensation or judgment.)

The empirical research continues. It will be remembered that the T-Machine was said to be set-theoretically equivalent to the V-Machine. Thus the behaviour of both these machines, considered as providing inputs to a U-Machine homeostat in which the V-decisions are approximated (via *d*-directions and *s*-sensations) as T-configurations after a time-lag, may be studied in the operations room. For the display there is the output of the (exemplification of the) T-Machine, while projected moves discussed on the display are momentarily outputs of the V-Machine; management itself plays the role of the U-Machine, monitored by its own staff in the role of R-Machine.

Hence some sort of exemplification of the whole artefact is now in being. The T- and V-Machines have genuine exemplifications, while the U- and R-Machines are spurious exemplifications at the present time. (For these are artefacts of the artefacts of real managerial systems—namely the real managerial systems themselves!) However, it is now possible to proceed to the empirical study of the U- and R-Machines in the context of *actual* exemplifications of the T- and V-Machines: that is the important point.

As far as the construction of cybernetic machinery is concerned, it is clear that the first component to transcend the status of a mere exemplification must be the U-Machine. For exemplifications of T- and V-input are already available, and can be fed to a U-Machine in parallel with their equivalent reporting to management—which can then provide an R-Machine monitoring service to the cybernetic U-Machine. Having succeeded in operating the cybernetic U-Machine, the research will turn to constructing cybernetic T- and V-Machines. Only at the last stage would the R-Machine be built. After this, management would be free for the first time in history to manage, not the company in the language of the organism, but the T-U-V(R) control assembly in a metalanguage. (That word is used strictly, for at last it would become possible to discuss T-U-V(R) theorems as undecidable in the direct control language. Managers who attempt this feat today are



discussing their *own* brains, insights and motives, and are sent to see psychiatrists.)

## 5. A PROSPECTUS OF THE ARTEFACT

Experiments have been conducted for a number of years by a number of people into possible artefacts having self-organizing properties. Most of them have been concerned with devices using standard electronic techniques (that is, up to and including the use of transistors) and standard computer technology (that is, up to and including the most advanced machines for business and industry). A great deal has been learnt from these researches, but I consider one outcome to be that these standard approaches (as I have just called them) are unsuitable for the artefact considered in this paper.

There is no need to provide detailed objections to such approaches here. The most cursory consideration of the sheer size of a "standard" computer-like artefact for controlling a cybernetic factory rules them out. But, following some study of the logical status of electronic brain analogues which was reported four years ago,<sup>(17)</sup> I came to the conclusion that the real secret lay not in pragmatic considerations of constructability, but in the logical structure of *the analogue fabric itself*.<sup>(4)</sup> This is not a trivial statement, in my view. As a constructor of machines man has become accustomed to regard his materials as inert lumps of matter which have to be fashioned and assembled to make a useful system. He does not normally think first of materials as having an intrinsically high variety which has to be constrained.

But there are new developments in solid state physics which may resolve the objections to "standard" electronics. With micro-modules (as announced by the Radio Corporation of America), an electronic component is envisaged perhaps for the first time as a wafer of material of uniform size and shape that has been constrained rather than fashioned to behave in a desired way. Even so, the emphasis is still on the size criterion rather than the fabric criterion: micro-modules must still be assembled to a circuit design. A further stage is reached by molecular electronics (as announced by Westinghouse) in which differential behaviour is obtained between domains of molecules inside single crystals.

Again, however, the assembly must be designed topologically; and again there is the emphasis on size. Now this emphasis clearly derives from the military and astronomical relevance of these developments, and is unobjectionable; but the cybernetician will need to make an intellectual effort to adjust the emphasis towards the changed analogue of fabric that is implied. The effort required is so great that so far I am unable to see how "design" can be eliminated from molecular electronics: the U-Machine must be enabled to construct its own components, and this fluid and evolutionary self-designing process should not be irreversible. Besides, the constraining techniques so far available ("diffusion, plating, electron-beam machining, etching, cutting, radiation, alloying, and photographic processes") involve massive equipment that could hardly be visualized as operated by the U-Machine to change its own internal mechanism.

Probably the most adaptable fabric for our purposes which has yet been made to work in an actual cybernetic machine is the colloid developed by Gordon Pask, and referred to previously. Electrochemical machines in general offer an apparently more amenable fabric than the physical semiconductors, and other workers in Britain are researching on similar lines. George's experiments with cotton threads soaked in sodium hydroxide, which measure conditional probabilities by their changing conductivity as impulses are passed through them, are cases in point. Some people have given thought to the use of organic materials, such as the lipids, to provide semi-permeable membranes as a delicate analogue fabric; and in general the interfaces between aqueous liquids offer a means for topological constraint of an undifferentiated or high-variety fabric.

This is not an exhaustive review of other people's work: these ideas are drawn together to lead into what I regard as a list of important features for any proposed analogue fabric for an industrial brain artefact. A high-variety material is required which can be topologically constrained, and reversibly, by simple low-energy inputs. Structuring of the fabric thus obtained must supply requisite variety for absorbing input variety. The structuring and its associated measures of information must be "readable": not indeed in the (by now) trivial sense of offering a digitized output, but as mapping itself onto an external situation from which feedback can

be supplied to the inputs. None of these activities needs to be a linear function, nor even a definable function, of input. The whole assembly is a black box, and needs no designing. In its solutions to problems simply *grow*, as Pask's metallic threads grow. And to bring the issue back to solid-state physics, if a ribbon crystal of germanium can be made to *grow* from a melt with differentiated domains (as is said to be possible), then by including the *melt* in the system semiconductors become possible fabrics again.

Now all this is said to contend that we may have been looking for the wrong thing in cybernetic research into possible artefacts. There may have been too much concern with assembling to a design, and with forcing materials to conform to that intention. If my interpretation of the facts is correct, almost any undifferentiated stuff will serve as a fabric for the U-Machine, and I would like briefly to catalogue the several fronts on which we are pursuing this thought in the Department.

In terms of electronics, K. D. Tocher is building a large apparatus for constructing an almost infinite variety of black boxes out of eighty units consisting of non-linear electrical networks. Inputs and outputs can be selected arbitrarily from possible nodes of the system, which is driven in terms of information by an extremely fast random-number generator with thermionic diodes as sources of noise. It is intended to explore homeostatic structures in regard to the time taken to reach equilibrium with this machine.

The formal logical properties of systems that grow are being investigated by R. A. Cuninghame-Green on a Pegasus computer. The programme under development is intended to provide a specification for a suitable analogue fabric for an artefact such as a U-Machine, which can then be quoted in the search for materials.

But so far the most valuable lead in my own mind concerns the use of organic systems themselves. In mentioning lipids just now I quoted an organic substance; anyone could be forgiven for asking why a material such as lecithin should be considered when it is so difficult to control compared with various sterile substances of similar molecular properties. But I now speak boldly of animals and animalcules themselves. Do not colonies of living things, which are reproductive (and therefore "self-repairing" as a fabric), with

their tropism as individuals and their taxis as groups, offer a fabric that meets the specification recently set down? Even molecular electronics cannot rival a living organism for size, much less for cost.

In pursuit of this idea I have undertaken a number of experiments, notably with human beings as being readily available and capable of introspection when the experiment is over. The object has been to make use of the natural behaviour of a system with living components to find equilibria in another system of which these components are ignorant. In this way I have obtained solutions to simultaneous linear equations from pairs of subjects who imagined that they were seeking an equilibrium in an exceedingly simple letter game—the ostensible object of the experiment. Children who have never heard of an equation can reach the same results; and I see no reason why (for instance) mice should not be similarly employed if the letter game is translated into a “cheese game”. The theory behind this work has been the transformation of the actual problem language into homomorphic mappings inside another kind of, and simpler, problem language the handling of which is natural to the participants. This is to amplify their intelligence in exactly Ashby’s sense.<sup>(9)</sup> The apparatus used was an algedonic system registering pain and pleasure for the participants. This is easily constructed for human beings, who are simply told to entertain a red light as meaning pleasure and a green light as meaning pain. For animals the algedonic mechanism would require more thought, although I have elsewhere<sup>(4)</sup> reported somewhat unconvincing trials with the freshwater crustacean *Daphnia*.

The point of the discursive review in this section is simply to emphasize the importance of the analogue of fabric in constructing artefacts of the kind discussed in this paper at large. This is not the place to record all the work done in this field meticulously, but I have tried to give an indication of the breadth of approach that seems necessary. Before long a decision will be taken as to which fabric to use in the first attempt to build a U-Machine in actual hardware (or colloid, or protein), and I look to the Symposium for advice.

\*

\*

\*

The research here described is incomplete, and the account of it uneven. While some aspects are, I trust, properly explored, others are sketchy: they fade into shadows where nothing is yet clear and distinct. But the project, though unfinished, has the merit of being very much alive. A self-organizing system must always be alive and incomplete. For completion is another name for death.

#### ACKNOWLEDGMENTS

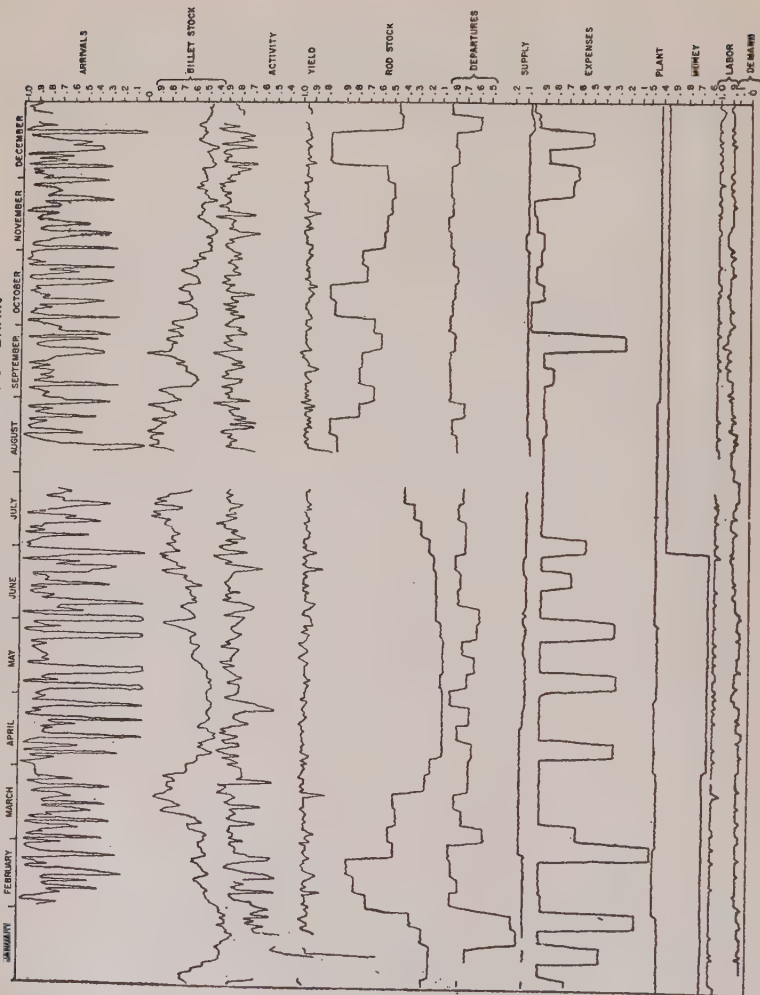
The company in which this work has been pursued is the Templeborough Rolling Mills Limited. The enthusiastic support of its managing director (Mr. S. R. Howes), of its general manager (Mr. H. Sadler) and of its other senior executives and officials, is acknowledged with gratitude.

The research is under the direction of the author, but the detailed results given in Part III were obtained by a project team consisting of three operational research workers: Mr. T. P. Conway, Miss H. J. Hirst and Miss M. D. Scott. This team is led by Mr. D. A. Hopkins, who is also the author's chief assistant in this field.

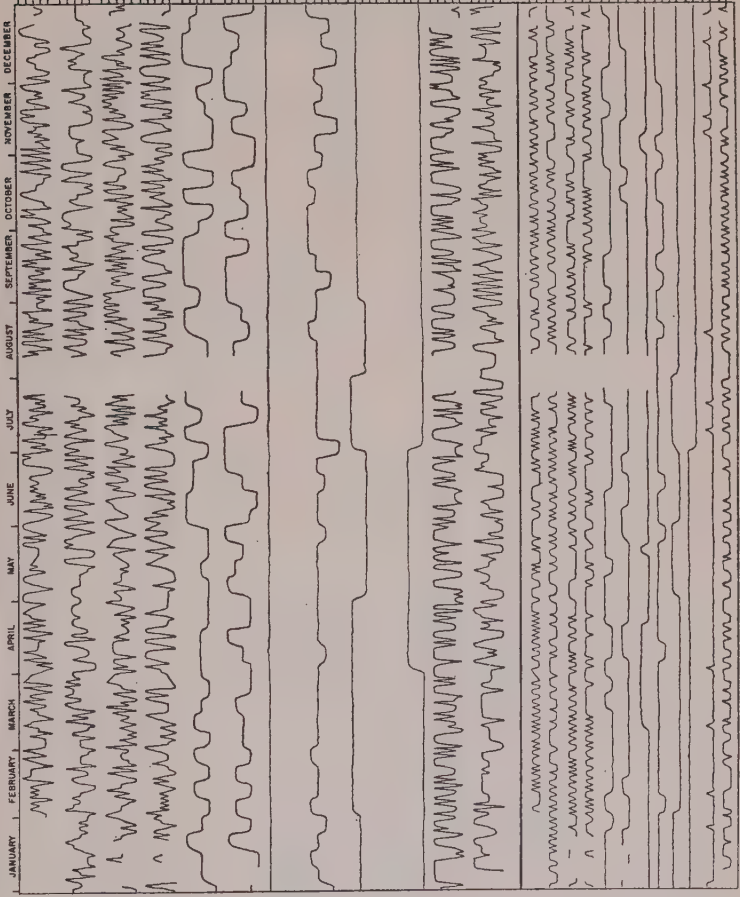
Acknowledgement is also due to the United Steel Companies Limited, who sponsor the Department of Operational Research and Cybernetics, and with whose permission this research is being undertaken for Templeborough Rolling Mills and also reported here.



## APPENDIX I: TRIAL "ENCEPHALOGRAM" OF THE COMPANY'S "BRAIN"



APPENDICES 2 & 3: HOMOMORPHIC TRANSFORMATIONS OF "ENCEPHALOGRAM



APPENDIX 2: T: R(mod 5)

APPENDIX 3: T: R(mod 2)

## REFERENCES

1. W. ROSS ASHBY, *Design for a Brain* (2nd edition), Chapman & Hall, London (1960).
2. W. ROSS ASHBY, The mechanism of habituation. *Mechanisation of Thought Processes*, (N.P.L. Symposium No. 10), H.M.S.O., London (1959).
3. GORDON PASK, Physical analogues to the growth of a concept. *Mechanisation of Thought Processes*, (N.P.L. Symposium No. 10), H.M.S.O., London (1959).
4. STAFFORD BEER, *Cybernetics and Management*, English Universities Press, London (1959).
5. J. C. ECCLES, Interpretation of action potentials evoked in the cerebral cortex. *E.E.G. & Clin. Neurophysiol.* **3**, pp. 499-564 (1951).
6. R. LORENTE DE NÓ, Cerebral cortex: architecture, intracortical connections, motor projections. *Physiology of the Nervous System* (J. F. FULTON), Oxford University Press (1943).
7. NORBERT WIENER, *Cybernetics*, Wiley, New York (1948).
8. D. A. SHOLL, *The Organization of the Cerebral Cortex*, Methuen, London (1956).
9. W. ROSS ASHBY, Design for an intelligence amplifier. *Automata Studies* (Annals of Mathematics Studies No. 34), Princeton University Press (1956).
10. N. BOURBAKI, *Théorie des Ensembles*; fascicule de résultats, A.S.E.I. No. 1141, Hermann, Paris (1951).
11. C. H. WADDINGTON, *Strategy of the Genes*, Allen & Unwin, London (1957).
12. JOHN R. BOWMAN, Reduction of the number of possible Boolean functions. *Transactions of Ninth American Conference on Cybernetics* (VON FOERSTER), Josiah Macy Jr. Foundation, New York (1953).
13. H. W. MAGOUN, *The Waking Brain*, Thomas, Illinois (1958).
14. ALF BRODAL, *The Reticular Formation of the Brain Stem*, Oliver & Boyd, Edinburgh (1957).
15. H. GASTAUT, Neurophysiological basis of conditioned reflexes and behaviour. *Neurological Basis of Behaviour*, Ciba Symposium, Ciba, London, (1957).
16. W. NOORDENBOS, *Pain*, Elsevier, Amsterdam (1959).
17. STAFFORD BEER, A technical consideration of the cybernetic analogue for planning and programming. *Proceedings of the First International Conference on Cybernetics*, Namur (1956).

## DISCUSSION

WILLIS: Something you said bothers me a little bit. Maybe I didn't understand how you were using it.

The number two to the two to the  $N$  is a very big number, as Warren McCulloch pointed out last night in discussing the five-input neurones. I think it is pretty clear that if you want to look at all the functions of  $N$  variables, of which there are  $2^{2^N}$ , you could probably do it when  $N$  is five.

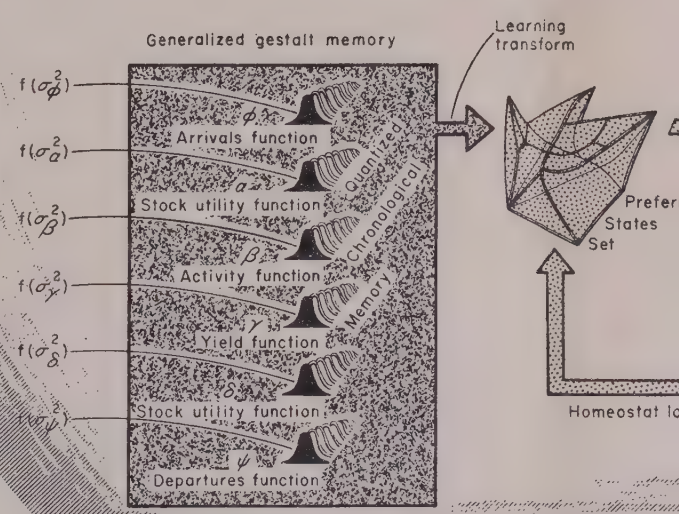
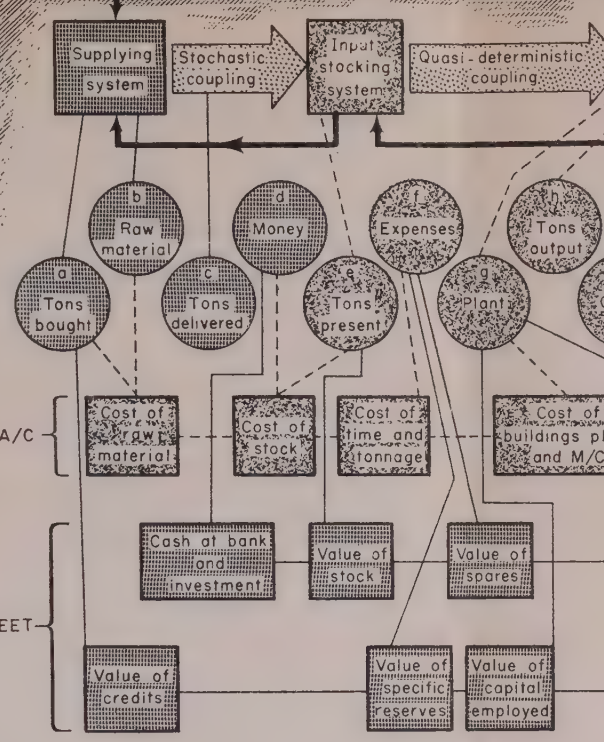
SENSATIONS

PROFIT AND LOSS A/C

DEDUCTIONS

BALANCE SHEET

JUDGMENTS







Two to the two to the fifth is about ten to the tenth. With a big computer which could look at functions at the rate of a million per second, the job could be done in a little under two hours. If you go to six variables, there are about  $2^{64}$  functions, about ten to the twentieth. A little calculation here shows me that if you have 30,000 of these computers and run them at full speed and are willing to wait a little bit more than your lifetime, you can do it. If you go to more than six variables, the thing becomes absurd.

Now, what are you going to do about it? This is the universe of your functions. How do you win?

BEER: Do let me tell you this.

VON FOERSTER: I think Jack Cowan would like to inject something.

COWAN: It is simply this. What you are assuming is that you have a two-valued non-redundant Boolean logic. What Stafford Beer has been talking about is the evolution of a set of systems in which constraints (or redundancies) are introduced. Because of Ashby's results, we know that such constraints can always be used to improve the effectiveness of search procedures. It follows that simple trial-and-error scanning, Markoff scanning, can be bettered by search procedures which make use of these constraints. In fact, there exist certain many-valued logics which are so constrained in structure as to admit the definition of several kinds of measures or distance-to-go criteria, so that these may be used for the computation of "trends". We may conceive of a reasonable search procedure as one which uses distance-to-go criteria to get into a region that is very likely to contain a solution, and then "hill-climbs" to the solution, using the Markovian technique. This process would be much faster than the simple Markovian technique, and so your comments are not quite appropriate.

I would like to make a comment on what I think is an important part of Beer's paper, and that is the indeterminacy of representation. It seems to me, at first sight, that this is but a restatement of Gabor's uncertainty principle: that the analysis of any physical signal, be it time function or frequency function, into time *and* frequency cannot be performed with arbitrary precision. That is, although analysis into time *or* frequency can be performed with any degree of accuracy, it cannot be simultaneously carried out in both beyond a certain limited accuracy. If this is so, then what is to be gained by calling this principle by a new name and going into the business of well-ordering denumerable sets?

BEER: I haven't denied the similarity. It is certainly analogous.

VON FOERSTER: I think that was a very helpful comment because you can now refer to some facts which are already existent. When Stafford spoke about this, he struggled a little bit: he felt he had to defend very strongly his third level of uncertainty. This has been established. So now you have a wonderful way to point out where this is going. You can lift it from Gabor's uncertainty principle and from information theory.

BEER: But, Heinz, then people will say it is not right. They will say it is not relevant. I have tried to bring this out in its immediate context. I agree it is the same point in principle, but I have tried to bring it out here in the brain model as such, and I think this is worth doing.

May I say something about Willis's point and Cowan's comment on it—which was, of course, absolutely right. We have got to keep very clear heads about the proliferation of variety. We know from Ashby's clear and beautiful statement that we must have "requisite variety" in a control system.

Now the kind of thing we are trying to control in a brain, or in a company,

or in an economy, is proliferating variety. The amount of variety here is simply colossal; and so I begin by saying I must be right in talking about an equivalent amount of variety which we must generate in a control.

Now you say, yes, but you cannot provide it in a computer, and of course, you cannot. This is why Gordon Pask and I have been working for years on what I always call the analog of fabric. So much work has been put in on the analog of mechanism, and on the analog of uncertainty (through statistics, quasi-pseudo-random numbers and so on). But we have not done enough on the analog of fabric itself. What we need, in a nutshell, is a high-variety, undifferentiated fabric which we can *constrain*. We do not want a lot of bits and pieces which we have got to put together. Because once we settle for the second, we have got to have a blueprint. We have got to *design* the damn thing; and that is just what we do not want to do.

COWAN: The point questionable to me was that once you find a path through the phase space Willis mentioned, and reach a point corresponding to a homomorphic image of the environment, do you really have anything in the nature of a meaningful measure of the environment in question?

BEER: I do not suppose I have, and I do not care. You see, the thing is, if we preoccupy ourselves with transfer functions, and being able to define things, and the getting of meaningful measures, we are constructing a descriptive science—not a control science. In order to talk coherently I have established homomorphs of reality, because these we know and understand; but in fact, it seems to me that our conditions for the real control system are very, very simple. You take a line from a control box into the world, where there must be some effect, any effect; and you take a line out of the world and into the control box, where there must be some effect. And this system will run to stability. I think this is what Ashby has shown; and we do not need to know what those transfer functions are. We do not care if they are non-linear, for example. Electronics people get so worked up about those. Why do they matter? If we design a system to be self-organizing, why take these analytical steps to organize it ourselves?

WILLIS: Let me make one more comment on this. I think that the same thing that licks you when you try to do this with computers is going to lick you no matter how you do it. I think the numbers are just too big. Now we have a point of disagreement.

BEER: All right, but I assert my right to reply. The way I have approached this, at the unconstrained, undifferentiated block of stuff level, the "computation" is going on automatically and without access at a molecular level. Pask's own machines—

WILLIS: My point is you don't have enough molecules.

BEER: Well, if you are really going to exhaust the molecules, I admit we are in trouble!

WILLIS: When  $N$  approaches 100 I think you may even begin to exhaust the molecules in the universe.

VON FOERSTER: I think one of the essential problems that Stafford has in insisting on this particular discussion is how to link two to the two to the  $N$  internally.

WILLIS: You cannot do it. You must do something else.

BEER: We must get my contention straight and precise. I do not say you must link every one of two to the two to the  $N$  with every other. I do say you must create a definite, limited linkage. But on the other hand you are not going to say how and where the linkage will occur in advance. My T-Machine model is

constrained all right, but only by its own experiences, its own epigenetic landscape; it is not to be constrained by the ignorant designer who has no idea what life in the future will be like.

WILLIS: You have to look at some subset. I am convinced of that.

BEER: Yes, but you must not choose it yourself. This is what I am saying. The "metalanguage machine" must choose it.

WILLIS: That may be true, but I think I shall be able to show you in my talk that such a thing is impossible when  $N$  is large.

BEER: The concept I have brought forward on this point is taken from genetics. After all, consider how this same point arises in genetics. Think of the variety that we constitute as human beings, and think what a task it would be to permute the available variety all down the line. We do not. We constitute a subset of available variety as you say. And the way that is achieved without designing in detail in the chromosomes, is a mechanism I have pinched from Waddington, the geneticist; the epigenetic landscape. This is something which constrains. It is a response surface which constrains the total, but not in a determinate way. This I think is a very valuable piece of borrowing.

BOWMAN: I have spent about two hours with your manuscript, but I did not bring it with me. I did not understand that you were going to speak this morning. There were a few little points that disturbed me. Let me just pick on some of the, I will admit, trivial ones, just a brief moment here.

Section 1.12: "to every element  $s$  of  $S$  there corresponds a dense, denumerably infinite set". Why denumerably? Must we handle everything in a digital machine? I think we are going to get into some difficulties of ordering if we speak of a dense denumerably infinite set, such as the set of rational numbers. You cannot put those in an order that makes much sense, even though they are denumerable.

That was just one small point that stops me. Just why do you introduce the word denumerable there?

BEER: Yes. I am aware of this difficulty. My answer is simply that I have done this in order to hang on to the concept of well-ordering, and nothing seems to be lost by quantizing a continuum for the purpose.

MULLIN: No problem; he just wants to invoke some properties of inseparable matrix space; it is just a convenient mathematical artifice. There are many mathematical results which are known which invoke this principle. It is a well-known mathematical principle; it is called separability; all that stuff can be run together.

BOWMAN: I see. Well, I was just a little bit disturbed because I believe, somehow, that the illumination of this room is a truly continuous non-denumerable variable that can take on transcendental values, not necessarily—

NOVIKOFF: That is a debatable point.

BOWMAN: That is the kind of thing I was anxious to get some talk started on. You can make your own hypotheses, of course, with the end in sight. We are all allowed to do that, but I was anxious to bring that point out.

BEER: Fair enough.

ASHBY: May I ask for clarification in regard to the general proposal? That is, in what form are you seeing the ultimate goal or range of permissible activities?

For instance, suppose we are considering a steel mill, and the personnel manager so handles things that strikes break out everywhere, everybody arms themselves with crowbars, and the sale of steel goes up high, making a bigger profit. Where does it come in your formulation, the decision about whether

this is a good thing or a bad thing? Are you assuming the system is just growing, as it were, or is there something that says no, we will not have that, and we will have this?

VON FOERSTER. Before you answer this, I think there is another question coming up from Sherwood.

SHERWOOD: If I may, because this is contingent on what Ashby said. I think your brain, it seems, has no religion. What is the factory for? If steel bar goes out of fashion, is it going to go into textiles or chemistry? If so, then how would it reorganize itself? It is built only for making steel. Now, if it is built only for making steel, do you want to make a big profit, or do you want to make high quality steel? Do you want to produce small expensive amounts, or large cheap amounts, and so on and so forth? This always comes back, I take it, into the factory as a feedback of some sort; but if the bottom falls out of the steel market, do you go into housing instead and give up airplanes and so on and so forth?

That is one difficulty; because the order, if any, that is imposed upon the factory is, in a way perhaps, a religion.

BEER: Yes, but this order is going to be self-imposed. You see this is one of the things I do not want—

SHERWOOD: But in the first place, the factory was designed to make steel.

BEER: The plant was; but not the company.

Now the point here, I think, is this; to answer Ross Ashby first, it seems to me that we just do not know enough about teleology yet. This is one of the things I want to investigate with this kind of model: are what we call purposes and goals in such systems merely names for equilibrium states—an inevitable consequence of trying to describe a system of this kind?

SHERWOOD: But you did say at one point the whole thing was only for survival, or homeostasis; but does it want to be evil or good?

BEER: That is part of survival, surely: does the environment favour behaviour that you would designate evil or good? That is a value judgment.

But the point I am trying to bring out is just this: what this machinery will *do* can only be a direct function of what it *knows*. This is game-theoretic talk, and it is perfectly acceptable, I think. This is a game, with a certain amount of information.

Now, if it only has inputs about steel, and the steel industry collapses, all that this organism can say to itself is "I am for the high jump"; and it may predict this sooner than men would. And in that case I think we should want to expand its knowledge, so that it could decide to do something else. This seems to be perfectly possible. But at the moment my model is thus limited; and therefore at the moment the answer to Ashby is that this is what management in the model exists for, to take this kind of superior decision. It can say: yes, well, this steel factory is running all right; but I am going to make ice cream now.

So it is that there is a whole hierarchy of languages here in which we can talk. And the point about this system, as of all systems, is that you can always redraw the boundaries to suck in a few more of the hierarchic levels; whereupon you have got a different kettle of fish entirely, which is a fascinating thought.

But the point about survival that I would like to make is this, that when I was criticizing the way that we do these things now, and complained about optimizing for one factor, we clearly have not got a performance criterion, such as control engineers like to have in systems of this kind, for viable



organisms with an  $n$ -variable system. A lot of the things in industry which are recognized as goals are incompatible; some are downright contradictory. Now in these circumstances the only thing I can do is to put a ring around this lot of contingent aims, and talk about survival as far as I can interpret it for them all.

VON FOERSTER: May I interrupt you just a moment? I think you should distinguish between life, or being alive, and survival. These are two different tasks, funnily enough. You can put something in, which will live, but probably not survive, because it cannot take in changes of the environment, their being too rapid or something of that sort. I think that the religion pointed out is just one of the sub-routines used in order to solve the survival problem. The religion becomes the sub-task which comes up as being a fair thing to do at the moment in order to keep on with the survival. The fact of life alone does not do it; but if you have a survival criterion, you may come up with an ice cream factory, funnily enough, if the organism is complex enough.

BOWMAN: Where does the criterion come from?

VON FOERSTER: Oh, from the problem; you set it up in there.

MCCULLOCH: I think the point on religion is somewhat different, if I may say so. I think the crucial thing is that the machine, if it can become aware of its epigenetic landscape, takes that more seriously than it takes itself, and this is good for survival. That is the crucial point.

BEER: Yes, and to the observer, this might *look* like a machine acquiring a set of value judgments. I mean we do not know how to talk in this area really, do we?

VON FOERSTER: We cannot yet.

BOWMAN: I do not want to take up too much time here, but the conversation has drifted on to your idea of a more-or-less homogeneous fabric of a machine, subject to some constraints.

Let me pick that up. As stated, it seemed to me a profound and advanced, but a very vague concept. Can I go to the opposite and speak along the same lines, perhaps from an engineering standpoint? It is routine now to design a general-purpose digital calculator, or a computer, and we start out with a block diagram, which ordinarily means very little to anyone except the man who made the sketch, because he usually forgets to put a description of the function of the blocks and the test. Without exception though, he will have a rectangle marked "arithmetic unit", and he will have one marked "store", or as we say here, "memory". Those are often on separate sides of the page. I believe the idea of a constrained fabric very nearly breaks down the distinction between those two black boxes.

To take a particularly simple example of a machine that I have worked with at length, that machine has a very large, very slow magnetic drum store. The arithmetic unit is, oh, something you could hold in the palm or your hand, it is a two-bit adder, and a one pulse delay line.

Now, if we wish to perform a multiplication (this I should say also, is a purely binary machine), the multiplier rather automatically becomes a part of the arithmetic unit. It gives orders: shift, add, shift, do not add, shift, do not add, shift, add, shift, add, shift, add, depending upon the bits of that factor.

Now the whole operation of multiplication in a binary machine is a lot more complicated than that. We have a parity check. We have an establishment of sign. We have a round-off problem. Routine within routine; the *erm* microprogramming has come into use.



Now, all of those microprograms are on the drum, stored in exactly the same way numbers are, and provide a constraint as to what the machine will do. And I merely suggest that the distinction between operator and operant has become far too sharp. And I think you have come a long way toward breaking that down when you speak of a homogeneous but constrained fabric.

BEER: Yes, I hope so. You see this analogy with computers I do not like for two reasons. (I do not mean Bowman's exposition, I mean the analogy.) For one thing, I really cannot stomach, although I do not know that everybody agrees with this, I cannot stomach the idea of memory as a bit put in a locker to be called for like a parcel in a cloakroom.

Now memory in my kind of system, you notice, is really a path of facilitation through a phase space. You get onto this path in a few places, and *whew*, you should be off to the destination of recognition. This is much nearer to a physiological memory to me. And the other big point I would like to make, about the big electronic machines, which I think are just dinosaurs—

BOWMAN: Subject to the same fate?

BEER: I think so; that is what I meant. They are preoccupied with digital access. Now why is this? It is always possible, given an output channel which you can fit on somewhere, to say what is happening just there, and to get an enormous printout. Now we are not concerned with digital access, but with outcomes. Why do we pay so much money to make it available?

In the sort of machines that Gordon and I have been concerned with, you cannot get at the intermediate answer. If you take one of Gordon's dishes of colloid, you may be effectively inverting a matrix of order twenty thousand. The cost of the computer is perhaps ten cents. The only trouble is you do not know what the answer is.

Now this sounds absurdly naïve, but it is not, you know, because you do not *want* the answer. What you want to do is to *use* the answer. So why ever digitize it? If the molecules in this thing are "taking up" the answer pattern, then we have all the variety we need to feed impulses back to control something. Failing to recognize this has been a terrible trap in the development of control technology, and I think the exponential increase in cost of control systems in industry is a direct result of thinking you have not got anything unless you can measure it in the measuring-rod sense. It is doubtless true that science is measurement and all that sort of talk; but this is being made to mean that there is no measure unless you can produce the digit. This, to me, is a *non sequitur*.

ASHBY: I am reminded of what Dr. Turing said: that half the business in higher mathematics is to persuade people to accept, not what they think they want, but what they really want.

BOWMAN: I can give you a fine, first-hand example where a chemist, and a very good one, came down and gave us a statement of a curve he wanted plotted; and after many tens of machine hours we gave him reams of numbers. Then he sat down and laboriously put dots on graph paper.

I asked him, well, what are you trying to do here? Oh, well, I have to find the point of inflection of this curve. He wanted one number, and the machine would have given that to him, but he had to have the full curve and do the rest of it manually. He just did not know how to ask a question.

BEER: May I quote my favorite example of this digital access issue where the brain is concerned, because I would like the comments of the brain specialists present on this. Obviously we can do colossal calculations in our heads. We have got all the requisite variety. I often dodge about the traffic in Piccadilly

Circus, and I have not yet been run over; and this is a wonderful piece of calculation. But whenever I try and get *digital* access to my own brain, I cannot have it. If you ask me how far off that wall is, I should be twenty, perhaps fifty per cent out; hopeless at it. And the favorite example that I want to quote is this.

We draw a picture on the blackboard of a road receding into the distance that is to have telegraph poles on it. We might then say: the first two poles will be one foot apart, would you please give me the numbers which will define the rest of the intervals? So we try. Well, we get the answer right to the extent we know the intervals are monotonic decreasing, and that is about it. If we take those estimated digits and plot them, we can see at once that it is wrong.

But we can say to almost anyone: draw those poles in, and he will get them dead right by eye. So, I argue, we have apparatus which can do projective geometry in our heads, even if we have never heard of projective geometry. We can *do* it, but we cannot get the digits out.

Now, does this merely mean (this is the question to the neurologists) that we have not got the digital output channeled to the right point; or does it mean that in some sense it does not actually exist?

SHERWOOD: No, the answer to that is when you are trying to find out the measurement, you have got to work from a symbol to an object. You have got a symbol in your head.

BEER: Well, this is just a mapping then?

VON FOERSTER: No, it must be identification of a symbol. An interpretation problem. The symbol means what you are seeking, looking for; and you need the interpretation of the symbol, which is a process. You have to re-evaluate lots of different possible states: what it *could* mean, and things of that sort.

SHERWOOD: There are at least two translations in different languages.

VON FOERSTER: Yes.

BEER: What I am getting at is this, it is a fact that those distances when you finally draw them will be very accurate as worked out on a computer by projective geometry. What does that mean?

SHERWOOD: You are asking me to tell you the thermodynamics, and the accurate application of forces, and what have you. If you can jump across the brook and land on a stone, which is on the other side at Warren's farm, you cannot do that. You all can do the jump; but most of us do not know what it takes to calculate the forces acting on the foot, or what is needed for the balance. It is sheer physics; that means we have to translate an action into a science or into a different language entirely. And then, similarly, it is no good telling what one ought to do in terms of engineering.

MCCULLOCH: If you take any good cat, hold it upside down, and drop it that much off the floor, it lands on all fours. If you are a professor of physics, it would take you a long time to persuade a class that it is conserving angular momentum.

BOWMAN: He can roll himself into a doughnut, rotate and unroll; that is a conservation of everything.

BEER: This is Torus the cat, not Taurus the bull.

VON FOERSTER: George Zopf wants to say something.

ZOPF: Whether we pick the cat or Sherwood's problem of leaping accurately to a stone, I think what Beer is suggesting is that we cannot examine the mental processes and say, here is where the calculation of distance is reported, here is where the calculation of momentum is reported, and so on, and say that

then these are put together to define what the motor act should be. In other words there are no representations of these sub-calculations individually.

MCCULLOCH: We are asking questions like in which of the vacuum tubes of my radio is Rudy Vallee's voice.

BEER: No, no. This is not what I meant at all.

ROSEN: I want to go back to something you raised a little while ago.

You have described several types of self-organizing machines made of some fabric where the useful results that come out of a machine are as a result of constraints. I am just wondering, and this is my own intuition, whether the system of constraints that you must know and apply to a fabric, does not constitute as much of a problem as the problem of what happens in each one of your boxes representing a machine, and therefore we may have just transferred the problem, setting perhaps just as difficult a problem in defining specific restrictions.

BEER: I think the mistake you make is to say we must know the system of constraints. We do not know them, that is the whole point.

ROSEN: I am sorry, but the implication that I got from many remarks that were made, was that the purpose of the machine was the operation of the particular plant. Once you state these objectives, if you can state them——

BEER: You cannot. This is the whole point, you see. One manager will tell you we are here to make a profit. Another manager will tell you we are here as a social service. And all of them act as if they were there to fish for trout. You cannot state these things.

ROSEN: If you cannot state them, and these are the constraints that are applied externally onto these big black boxes that have the capability of doing some organization within them, I do not see how you are going to apply any constraints that will permit an optimum set of self-organizing procedures to give you some results.

BEER: These inputs are the constraints. They are what constrains.

COWAN: The mistake Rosen is making is thinking that these constraints are absolute constraints. You really have to think of them in a relative sense.

ROSEN: Are they variable; are these constraints variables?

COWAN: They would become variable eventually.

ROSEN: If they do, then you have the same problem of determining which of these sets of variables are going to mean something, if this problem has any meaning at all. I will accept the fact that if the constraints can be stated in a simpler form than the internal mechanism of the machine, we have simplified the over-all problem. I have not seen this.

MCCULLOCH: Why do we need to state the constraints?

ROSEN: Why go through the procedure of making these self-organizing machines?

COWAN: It would not be self-organizing if you state the constraints.

ROSEN: But I think we started by saying we wanted to improve the operation of an industry or a plant. That is how we started.

VON FOERSTER: It means dropping defined constraints.

ROSEN: Well, if you drop defined constraints, do you not start out with as difficult a system as you are trying to solve?

VON FOERSTER: Maybe Anatol Rapoport can help us here.

RAPOPORT: About what?

VON FOERSTER: About the discussion here.

RAPOPORT: I am sorry, I am trying my best to follow what is being said. I pass.

ROSEN: Let us all pass.

VON FOERSTER: Let us not yet pass. We have fifteen minutes yet, we can work something out.

ASHBY: One method for getting a decision is simply to have someone who says, "I am the manager; you ask me whether a thing is good or bad and I'll tell you. I don't give reasons, just give me the proposition and I'll mark it good or bad".

There is no *necessity* for a specification in detail of what is going on internally; what is necessary is the decision.

BEER: Exactly. This is the algedonic loop defined in my talk.

ROSEN: Can I answer that? This is precisely the thing. Now it is the manager with his brain, that determines whether or not he likes things or he does not. So we have in fact transferred the problem to a mechanism which is self-organizing, and just as complex, or even more complex, than the one we are plotting.

ASHBY: I am assuming that he is a person, in a sense an automaton, selected by the shareholders. They know the sort of guy he is, they know the sort of decisions he will give; to them he is just a very complex function over labor, money, profits, expansion—a very complicated function. All they want is just that he will be the deciding function of these variables and will say this combination is bad, this combination is good, and so on. He is not organizing; he is defining what Sommerhoff calls the "focal condition".

ROSEN: He is putting the constraints on the system: here there is an organism that is functioning to put constraints on the system, and it is as complex as the one we are making.

COWAN: The point is the system will remove him and substitute another if he is not giving the right sort of results.

ROSEN: You hope that somewhere the problem begins to simplify progressively, and does not keep on the same level of complexity.

VON FOERSTER: I think this is really a point of two different liberals talking. One thing I think Dr. Rosen was getting at was he would like to explain the problem perfectly clearly; as McCulloch pointed out yesterday, perfectly clear set responses you can expect from a very clear set of stimuli—change of light, straight edge coming about, and things like that.

Fair questions, of course. But, what, would we like to explain the frog as it is, or would we like to explain how it became that frog? The latest specification, the real constraints which we have in the frog, are the result of something happening. Nobody defined upstairs how the frog should come about, and these constraints are not laid down.

If you have a running factory then, maybe an outside observer, several outside observers, may come to a definite conclusion about its properties. It is *this* type of system. And then you come to the evolving ice cream factory, and they approach the whole thing as a different kind of a problem. And I think this was the kind of thing that Stafford was trying to point out, the transition problem of the one into the other. He says I have a steel mill, but this steel mill has no brain, so now let us get it into a state where it may finally have a brain. This is, of course, a gross evolution process. It is a thing which would evolve according to its internal constraints and become a better steel mill. And this would be just the thing that your Board of Trustees would say: let's become a better steel mill.





## WARREN S. McCULLOCH

*Research Laboratory of Electronics,  
Massachusetts Institute of Technology*

### SYMBOLIC REPRESENTATION OF THE NEURON AS AN UNRELIABLE FUNCTION

I am going to do worse than chair this session; I am going to speak for a minute first, because I know that by no means all of you are familiar with our ideographs.

The years from 1952 to 1957 I worked single handed on these problems, and it was not until 1959 that Manuel Blum joined me. The consequence is that I developed a symbolism which is idiosyncratic, to say the least. It came out of Euler's use of circles, to convey the relations of classes, and through Venn's diagrams, and through the use of a jot in such a diagram to stand for a proposition which was true. So the ideograms are actually an extension of Wittgenstein's truth tables.

Let me begin then by stating what I mean. I started out by using circles (Fig. 1a). Then I got lazy, I did not bother to draw any more than a part of the circle (Fig. 1b), so they became symbols that looked like the chiasm of Fig. 1c.

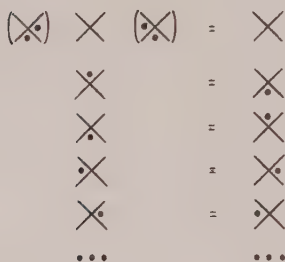
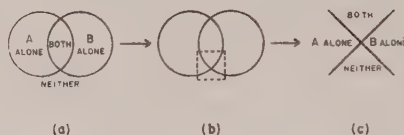
With the use of these diagrams, it is possible, by putting a jot in any one, two, three or four places, or leaving it blank, to write any logical function of two variables immediately. They operate on each other as follows:

Suppose I have three symbols as in Fig. 1d, in which  $A$  and  $B$  enter into the lateral two, and the central symbol operates upon these two symbols. I will take a simple case. Suppose I have jot  $A$ , and suppose I have jot  $B$ . All right, and if I have a jot in the "both" position in the operator symbol, it will take only the common jot, which is neither  $A$  nor  $B$ . If the jot is in the "neither" position, it will put a jot in the empty space, which is the top space, empty in both; if it is on the left, it puts in what is in the left symbol alone, and if it is operating on the other side, it will put in what is in the right symbol alone.

These rules can be extended immediately to any number of variables by drawing increasingly complicated forms.

ASHBY: I take it that the central "X" has a different meaning than the "X's" on the right and the left?

McCULLOCH: This central "X" is the diagram for the operation of an output neuron, and these are two input neurons, if you will. It is making this function of those two as arguments.



(d)

FIG. 1

ASHBY: You are now treating the right-hand "X" as a jot in relation to that middle "X"?

McCULLOCH: Yes.

ASHBY: So you could write the "X" with an "X" above and an "X" below?

McCULLOCH: Yes. It would not matter where I wrote the symbol. I just happened to put it between. It is that which is computing on the basis of the output of the other two.

Here (Fig. 2) I have drawn these jots inside the neurons, so that you see what they are doing. The input-output function is on the right. The numbers in the symbol in the output neuron are places where jots may appear as thresholds shift.

Now, I do not want to go and spend any more time on this, but we will be using this kind of symbolism. We are thinking about a realization in terms of neurons, in which as thresholds shift, the

function computed is off. So, if I have a very simple case as in Fig. 2, playing on one output neuron from *A* and from *B*, and the threshold happens to be four, then nothing fires it. If the threshold is three, it is fired for *A* and *B*. Theta equals two, for *A* alone, as well as for both.

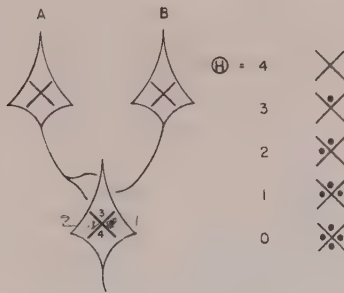


FIG. 2

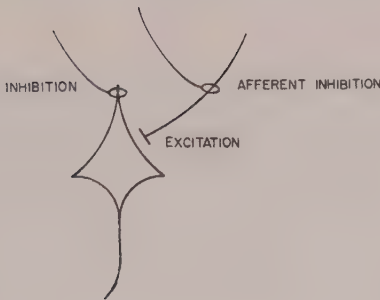


FIG. 3

Theta equals one, for both, for *A* alone and for *B* alone; and for zero, it will go off and I cannot stop it.

Now this is how simple the symbolism is, and it can be extended to any number of arguments by Selfridge's and Minsky's Venn diagrams, which are very simple constructions.

All right, now what we are thinking about, what I have been working on, up until this last year, has been on the various ways that I could combine these neurons. They will have the following properties inevitably (Fig. 3). They may be excited by afferents

delivered to the cell body. They may be inhibited by affairs which in general ascend apical dendrites. These are, therefore, the common symbols that we use.

Next, we do know that when an impulse occurs in some fibers, it can prevent others from reaching the cell. And this is that form of inhibition that is knocked out by strychnine. We know where it occurs.

Now what I have examined here are changes in the value of theta which may be moving together in a group of neurons or in scattered neurons; examining changes in the strength of the signals, examining changes in the synapsis itself, let us say if it gets two feet when it should only have one, and examining those properties, primarily, which might make trouble in a computer composed of these neurons.

That is: a change in the threshold, a change in the strength of the signals, and a change in the connections of the neurons. And it is very easy to build with neurons, for three or more inputs per neuron, extremely rugged circuits. They will stand all kinds of disturbance.

What I had not gone into, and what we have now gone into, is the troubles that occur when there is an error appearing anywhere in the output of this neuron, regardless of all this working properly. That is to say it appears as noise on the input to some other neuron, noise generated quite apart from this part of the neuron behaving properly or improperly.

And it is concerning this that we are about to present a job which we have worked on together, so much that I do not think any one of us knows who did what, and I will ask Manuel to start the ball rolling.

## MANUEL BLUM

*Research Laboratory of Electronics,  
Massachusetts Institute of Technology*

# PROPERTIES OF A NEURON WITH MANY INPUTS\*†

## 1. INTRODUCTION

A formal neuron is a logical device with well-defined properties. This paper gives some theorems about the neuron which clarify the interesting properties of circuits using them as components.

In his paper on Probabilistic Logic,<sup>(7)</sup> John von Neumann posed and attempted a solution of the problem of reliability in nets of computer components with two or three inputs. However, his solution required better components than could be expected in the brain.

The search for a solution was continued by Warren S. McCulloch.<sup>(1,2,3)</sup> Dr. McCulloch postulated a formal neuron, which we shall simply call a "neuron", and gave physiological evidence for the choice of this model. He connected the neurons in nets with outputs more reliable than the outputs of the individual neurons. This paper is a mathematical investigation of many-input neurons such as are contained in these nets.

The FORMAL NEURON is a computer component with the following properties:

1. It receives fibers from  $\delta$  inputs and has one output.
2. Each input and the single output may be either ON or OFF.

---

\* This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research), under contract with the Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts.

† Section 1 is an introduction taken almost exclusively from the work of Dr. Warren S. McCulloch. I wish to thank him, Herman Berendsen, Jack Cowan, Gene Prange, Leo Verbeek, and my brothers Simon, George and Michael, without whom this work would never have been completed.



3. Fibers from an input may divide, but may not combine with other fibers.

4. A fiber may excite a neuron with a positive unit (+1) of excitation (excitatory fiber) or excite a neuron with a negative unit (-1) of excitation (inhibitory fiber). A fiber may also inhibit a signal which passes through another fiber (Fig. 1).

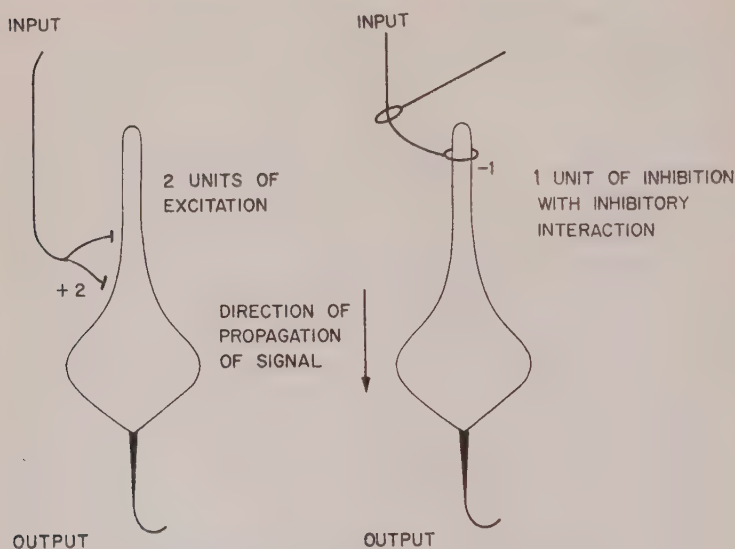


FIG. 1. Representation of neuron and input fibers.

















5. Signals may travel in only one direction through the neuron.

6. There is a unit time delay in the transmission of a signal through the connection between input fiber and neuron.

7. If the neuron makes no error, it fires when the arithmetic sum of excitatory and inhibitory signals to it exceeds some specified THRESHOLD ( $\theta$ ).

The calculus of propositions can be written by using “.” for “AND”, “v” for “OR” and “~” for “NOT”. VENN DIAGRAMS are more transparent schematic representations of propositions. Logical variables are represented by areas, and the intersection of several areas represents the logical product of corresponding variables (Fig. 2). Logical functions are represented in a Venn

diagram by jots in appropriate spaces. Thus we write the sixteen logical functions for a calculus of two variables as follows:

	"O" or "contradiction"		$\sim B$
	$A \cdot \sim B$		$A \cdot B \vee \sim A \cdot \sim B$
	$A \cdot B$ or "AND"		$A \cdot \sim B \vee \sim A \cdot B$
	$\sim A \cdot B$		$A \vee \sim B$
	$\sim A \cdot \sim B$ or "Sheffer stroke"		$A \vee B$ or inclusive "OR"
	$A$		$\sim A \vee B$
	$B$		$\sim A \vee \sim B$ or "Sheffer stroke"
	$\sim A$		"1" or "tautology"

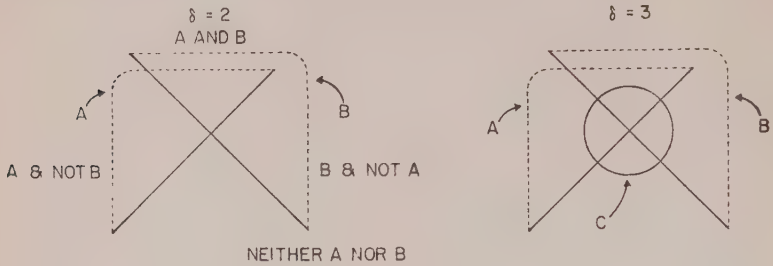
Venn diagrams are used to represent the logical function computed by a neuron. This function is determined by the connectivity of fibers on the neuron and by the threshold of the neuron. In general, it is necessary to draw a different Venn diagram for each of several possible values of  $\theta$  (see Fig. 3 for examples).

Each neuron has  $\delta$  inputs, and each input may be in either of two possible states: ON or OFF. Thus there are  $2^\delta$  possible INPUT CONFIGURATIONS to a neuron and each configuration is represented by a space in the Venn diagram corresponding to that neuron.

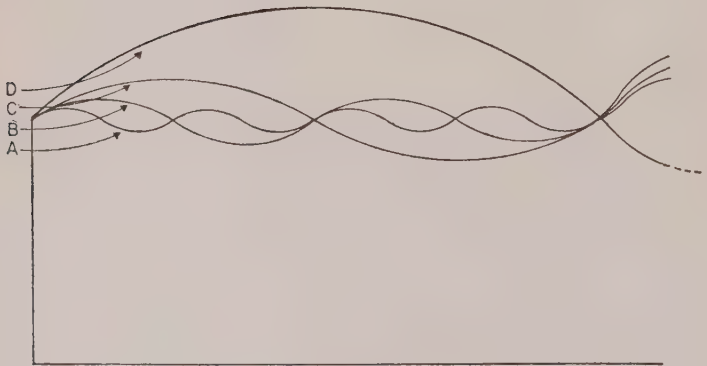
It is convenient to use only a single Venn diagram for the logical functions computed by a neuron as it varies in threshold. There are two useful ways of doing this:

(1) For each input configuration the neuron receives some numerical value of excitation. Let this number be placed in the Venn diagram in the space corresponding to the input configuration. Each neuron of Fig. 3 may now be represented by a single Venn diagram as in Fig. 4.

(2) Numbers may be placed in the Venn diagram to represent the order in which jots appear as the threshold of the neuron decreases. Each neuron of Fig. 3 may then be represented by a single Venn diagram as in Fig. 5.



(a)



(b)

FIG. 2. (a) Venn diagrams for 2 and 3 variables.

$\delta$  = number of variables.

(b) Minsky-Selfridge diagram which extends the Venn diagram to any number of variables.

It is easy to distinguish between the Venn diagrams of Figs. 4 and 5, since the former have a zero in the  $\sim A \cdot \sim B$  space while the latter do not.

*McCulloch nets*<sup>(4)</sup>

Let  $\delta$  inputs be connected to a rank of  $\delta$  neurons, and let the outputs of these neurons be connected to a second rank of neurons. Continue this procedure for as many ranks as desired and let the

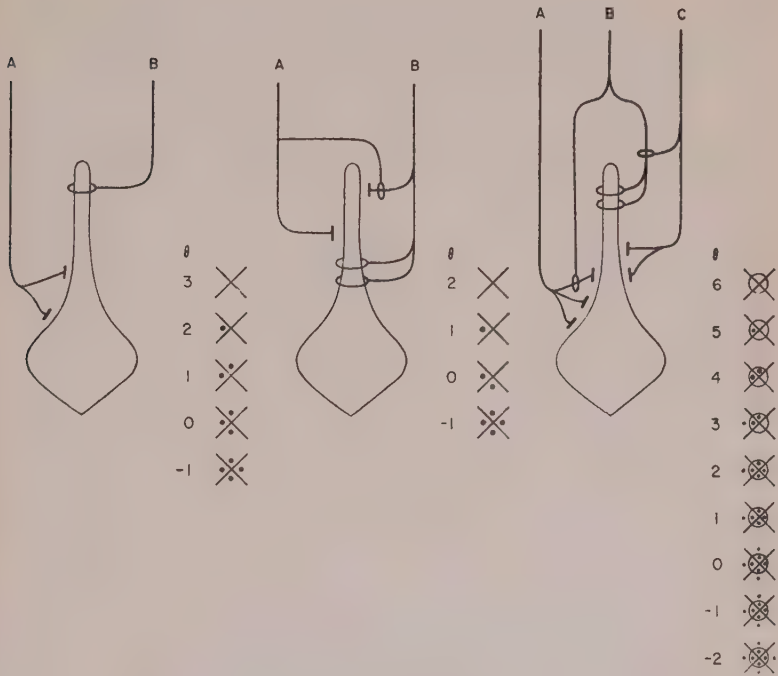


FIG. 3. Neurons and corresponding Venn diagrams for different thresholds.

last rank consist of a single output neuron. This configuration, with four or more ranks and no feedback, is a McCulloch net (Fig. 6).

*Rules for Manipulating Venn Diagrams in McCulloch Nets:*

We deal specifically with the case of  $\delta = 2$ . First we require a definition: In a row of Venn diagrams, a COMMUNITY SET is the set of all spaces which represent a single input configuration to the

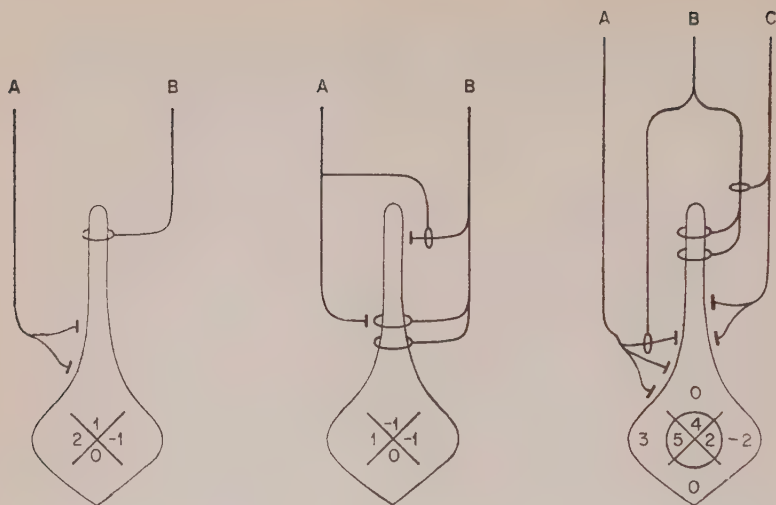


FIG. 4. The space of a Venn diagram contains the excitatory value of the corresponding input configuration.

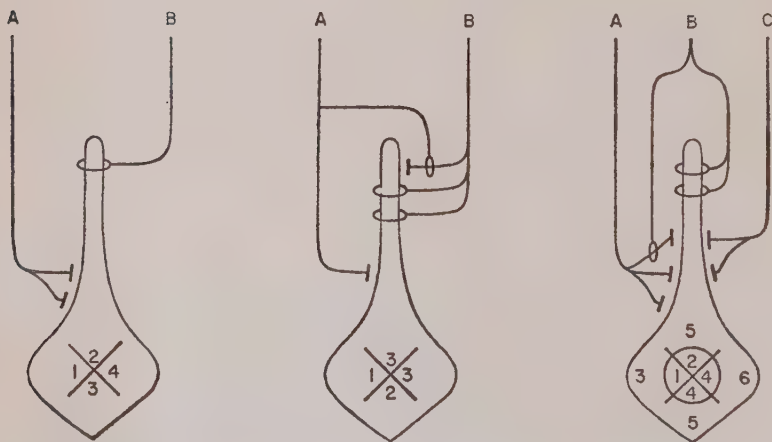


FIG. 5. Numbers represent the order of appearance of jots as threshold decreases.



rank of neurons. Thus the following set of  $A \cdot \sim B$  spaces is a community set:



The logical function computed by a neuron in a net is written in Venn form inside that neuron. The output of a neuron is written in Venn form inside brackets to one side of the neuron. This output Venn contains a jot for each input configuration which eventually fires the neuron.

In the following rules for dealing with Venn diagrams, the reader should refer to Fig. 6.

*Rules:*

(1) The output of a neuron in the first rank is the same as the logical function computed by it.

(2) The output of a neuron in the second rank may be computed as follows:

First look at the  $A \cdot B$  space of the Venn diagram in one of the neurons of the second rank. If this space contains no jot, then pass to the  $A \cdot \sim B$  space of the same Venn. If the  $A \cdot B$  space does contain a jot, then look to the output Venns of the first rank. Find each community set which has a jot in the space of the left output Venn and a jot in the space of the right output Venn, and place jot(s) in the output Venn if that neuron of the second rank in space(s) corresponding to each such community set.

Then look at the  $A \cdot \sim B$  space of the Venn in the neuron of the second rank. If it contains no jot, pass on to the  $\sim A \cdot \sim B$  space. If, however, it does contain a jot, find each community set in the first rank of output Venns which has a jot in the space of the left Venn and no jot in the space of the right Venn. Put jots in the output Venn of that neuron of the second rank in spaces corresponding to these community sets.

Then pass to the  $\sim A \cdot \sim B$  space of the Venn in the neuron of the second rank. Carry out the same procedure as above by searching for those community sets which have no jots in corresponding spaces of the two output Venns of the first rank. Then pass to the  $\sim A \cdot B$  space of the Venn in the neuron and continue as above by searching for those community sets of the first rank

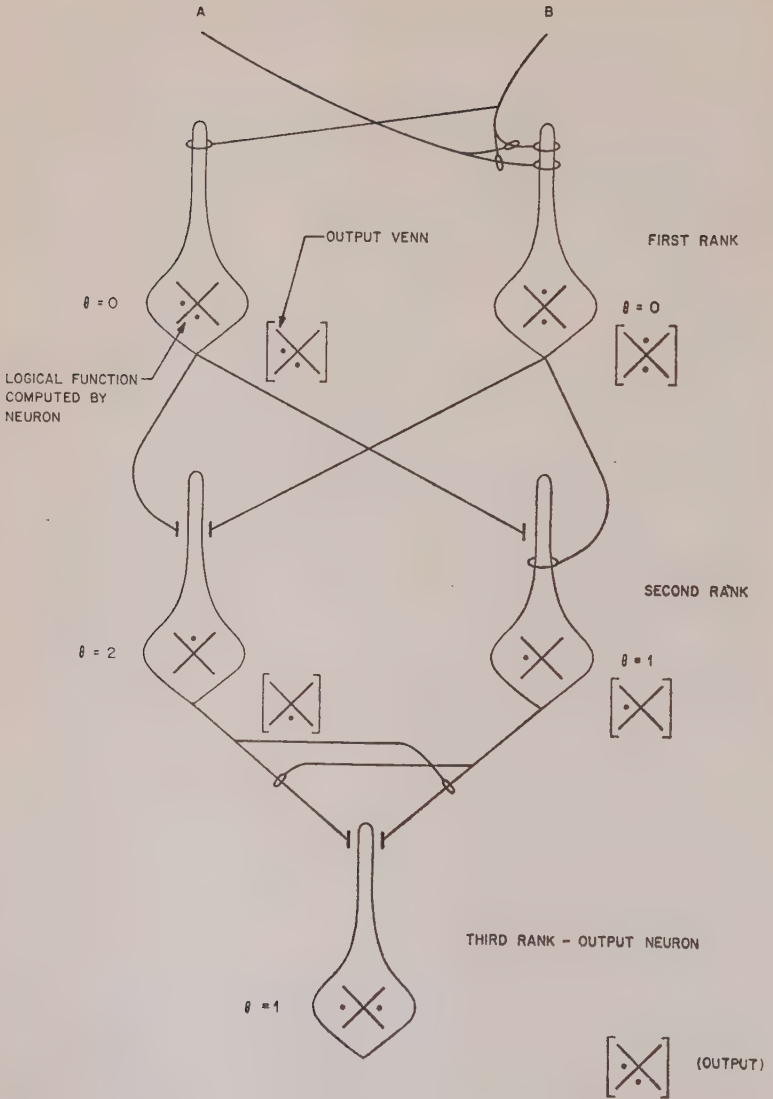


FIG. 6. McCulloch net for  $\delta = 2$ .

which have a jot in the space of the right output Venn and no jot in the space of the left output Venn.

(3) The output of a neuron in the third rank is computed in the same way as the output of a neuron of the second rank.

This procedure is easily generalized to all values of  $\delta$  (Fig. 8).

#### *Errors in Nets:*

Several types of nets have been investigated. We can only mention them and give a few examples. The interested reader is referred to several fine papers on the subject.

A LOGICALLY STABLE NET is a net in which the neurons change threshold simultaneously.<sup>(4)</sup> An example is given in Fig. 7. We have found it is always possible to construct a logically stable net to compute any desired function for  $(1-2^{-\delta}) \cdot 10^2$  per cent range of threshold of the component neurons. We have also found the exact range of threshold that neurons in a logically stable net may have when the logical functions in Venn form, computed by each and every neuron, are constrained to have the same number of jots. As simple bound, we can show that for  $\delta \geq 3$ , a range of threshold of  $(1-1/\delta) \cdot 10^2$  per cent is always obtainable.

Nets of neurons having thresholds which fluctuate independently of each other have been studied by Gene Prange of the Cambridge Air Force Research Center. He has obtained bounds on the permissible fluctuations of threshold of the neurons in a net which computes an error-free output. These bounds are for all  $\delta$ . We use his notation in Fig. 8 where jots, blanks and dashes in a Venn diagram represent, respectively, input configurations for which the neuron always fires, never fires, or fires with error.

W. S. McCulloch has studied changes in signal strength and in synopsis of fibers to a neuron.<sup>(5)</sup> This includes the possibility of error due to faulty connections of the fibers to a neuron. Another means of studying this type of error, proposed by Jack Cowan, is to view the arrangement of fibers to a neuron as a special case of Shannon-Moore relay networks.<sup>(6)</sup>

Leo Verbeek, whose paper appears in this volume, has dealt with errors occurring in the output fiber of a neuron. This type of error is the same as that considered by von Neumann,<sup>(7)</sup> but Verbeek's attack on the problem of reliability yields different and highly interesting results.

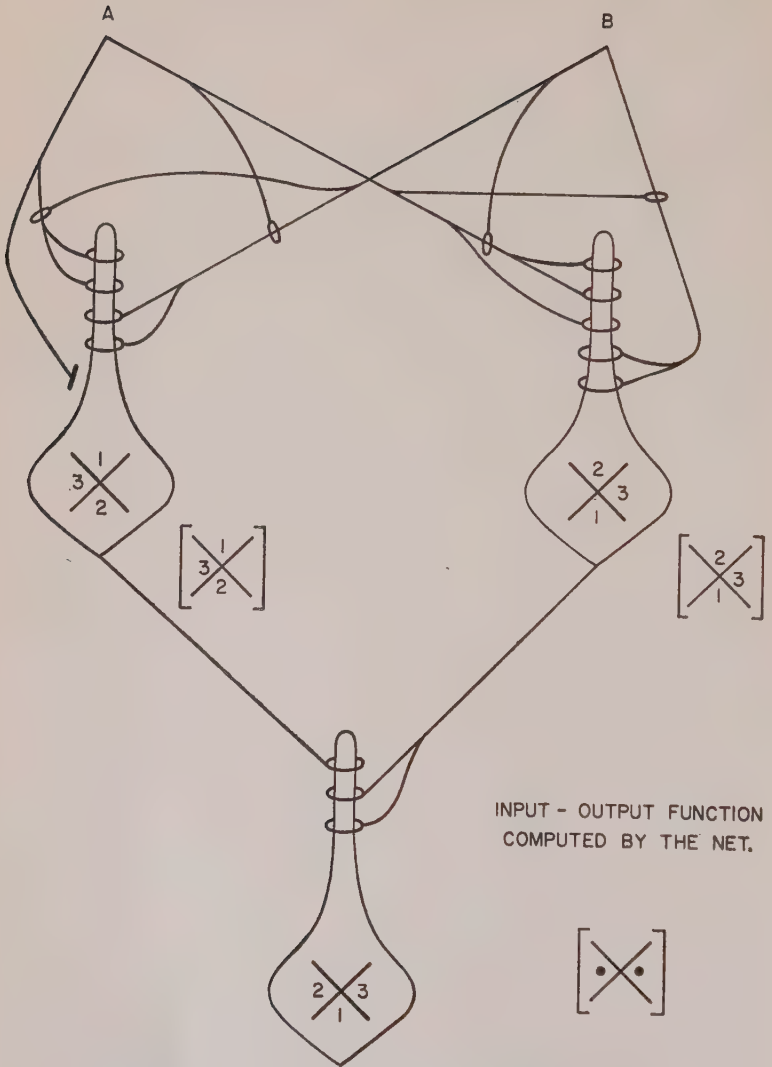


FIG. 7. McCulloch net, logically stable over 75 per cent of range of threshold.

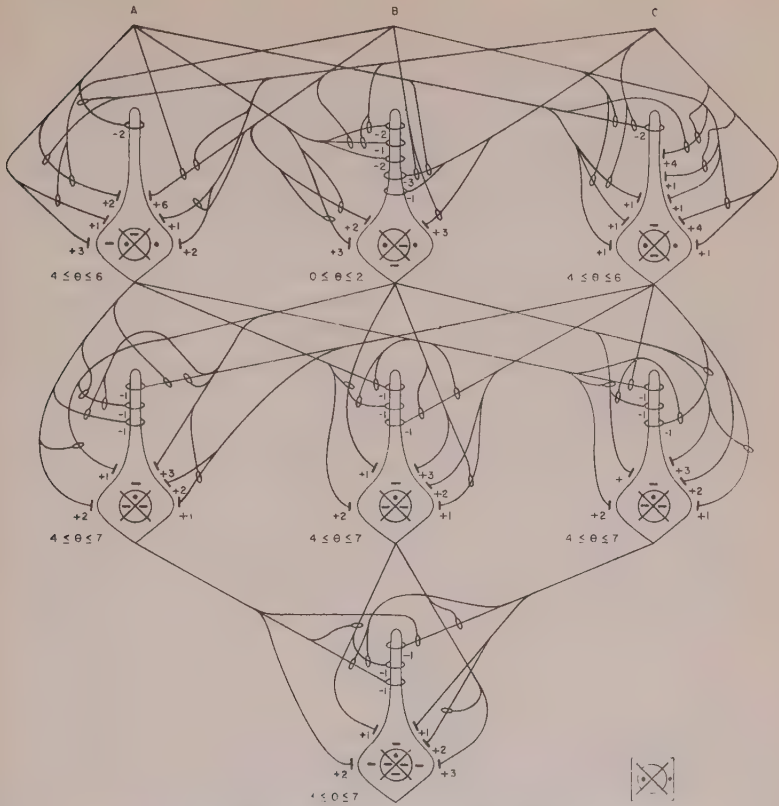


FIG. 8. Stable net of neurons with thresholds that fluctuate independently.

## 2. GENERALITY OF THE NEURON AS A COMPUTER COMPONENT

Until now we have had to construct the neurons in any net we wished to make, for we have had no way of being certain that there does exist a neuron which will compute the desired logical functions for every step in threshold. We now show that for each Venn diagram containing excitation values, as in Fig. 4, there must always exist a neuron. In addition, we shall give an algorithm for the construction of a neuron from a Venn diagram, and investigate the problem of constructing neurons which are cheapest,



in the sense that they receive a minimum number of fibers from the input.

A neuron will be constructed for the ( $\delta = 3$ )-Venn diagram of Fig. 10. This method of construction will then be extended to any  $\delta$ -Venn diagram.

Let  $A$  be the sum of excitations to the neuron when the input configuration is  $A \cdot \sim B \cdot \sim C$  (in this case,  $A = 1$ ). Let  $AB$  be the sum of excitations to the neuron when the input configuration is  $A \cdot B \cdot \sim C$  (in this case,  $AB = -2$ ). All other sums of excitation are similarly defined. The fibers from input  $A$  may be separated into four parts:

(1)  $A_{BC}$  is the sum of excitations from input  $A$ , from fibers NOT inhibited by either inputs  $B$  or  $C$ .

(2)  $A_B$  is the sum of excitations from input  $A$  from fibers NOT inhibited by input  $B$  but which may be inhibited by input  $C$ .

(3)  $A_C$  is defined similarly to  $A_B$ .

(4)  $A$  is simply the total sum of excitations from input  $A$ .

To make clear the reason for the definitions, we consider a particular 2-Venn diagram which we label as follows:



Imagine that a fiber from input  $A$  must break up into four parts and pass through this 2-Venn before ending on a neuron, as in Fig. 9. After each fiber leaves the Venn diagram, it may break up into many parts and become either excitatory or inhibitory on the neuron. The two boundaries of the areas of the Venn diagram represent inhibitions of the fibers through them from inputs  $B$  and  $C$ . In a space of the 2-Venn we write a number which equals the numerical value of excitation of the neuron from the fiber passing through that space. As an example, we write for the 2-Venn of Fig. 9 representing fibers from input  $A$  on the neuron:



and our definitions become:  $A_{BC} = 3$ ,  $A_B = 3 + 1 = 4$ ,  $A_C = 2 + 3 = 5$ ,  $A = 2 + 1 + 3 - 1 = 5$ . All three inputs to a neuron may be represented in this manner.

Returning to our problem of constructing a neuron for Fig. 10, we write, for example,  $AB = A_B + B_A$ , since only those fibers from  $A$  not inhibited by  $B$ , and those fibers from  $B$  not inhibited by  $A$ , will excite the neuron when  $A \cdot B \cdot C$  is the input configuration. Clearly, the  $\sim A \cdot \sim B \cdot \sim C$  space of the Venn diagram must contain a zero, since for this input configuration, the neuron receives no excitation.

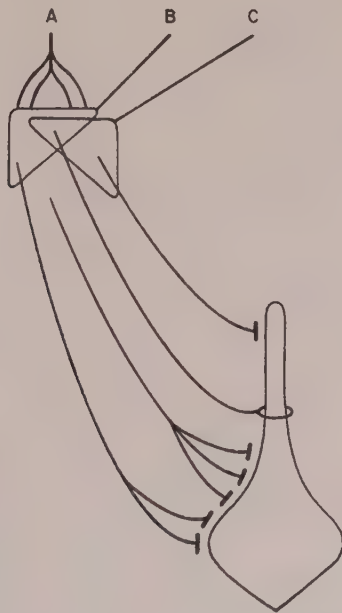


FIG. 9. Inhibitions from  $B$  and  $C$  may be viewed as forming a Venn diagram.

The equations of Fig. 10 indicate that five variables may be chosen arbitrarily. For example, let:

$$\begin{aligned}
 A_B &= -2 \quad \text{which implies } B_A = 0, \\
 A_C &= 1 \quad \text{which implies } C_A = 2, \\
 B_C &= -1 \quad \text{which implies } C_B = 2, \\
 A_{BC} &= B_{AC} = 0 \quad \text{which implies } C_{AB} = 2.
 \end{aligned}$$

For each input to the neuron there exists a 2-Venn that describes the number of excitations and inhibitions on a neuron and

how these fibers are inhibited. Thus, if there exist three 2-Venns that satisfy the above equations, then the neuron may certainly be constructed. Suppose the 2-Venns are filled by first satisfying

Equations	Variables
(1) $A = 1 = A$	
(2) $B = -1 = B$	
(3) $C = 2 = C$	
(4) $AB = -2 = A_B + B_A$	
(5) $AC = 3 = A_C + C_A$	
(6) $BC = 1 = B_C + C_B$	
(7) $ABC = 2 = A_{BC} + B_{AC} + C_{AB}$	

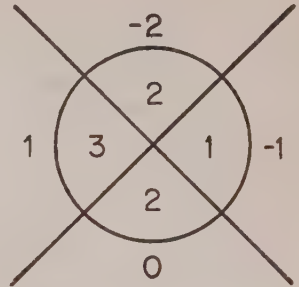
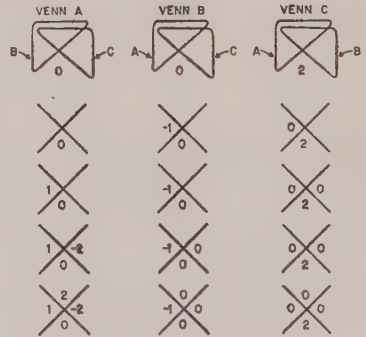


FIG. 10. An example of a Venn diagram for which a neuron must be constructed.

equation (7), then equation (6), and so on up to equation (1). This *might* be done as follows:

equation satisfied

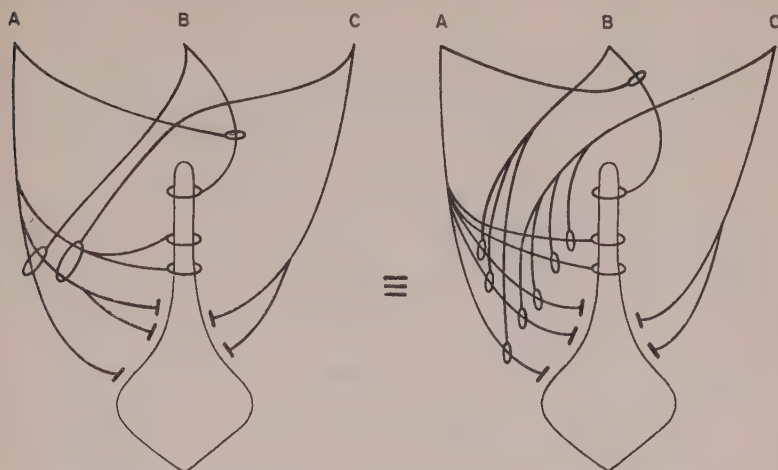
- 7
- 6
- 5
- 4
- 3, 2, 1



It is readily seen that the number in each space of every 2-Venn is determined by the value assigned to each of the twelve variables of equations (1)–(7). Moreover, there is always a sufficient number of spaces in the 2-Venns to satisfy the equations since the number of variables is precisely equal to the number of spaces in the three 2-Venns. In our example, the 2-Venns are:



The neuron which may be constructed directly from these Venns is:



In general, the number of variables such as  $A$ ,  $A_B$ ,  $C_{AB}$ , etc., will be

$$\binom{\delta}{1} + 2\binom{\delta}{2} + \dots + \delta\binom{\delta}{\delta}.$$

A  $(\delta - 1)$ -Venn has  $2^{\delta-1}$  spaces, therefore the total number of spaces in  $\delta$  Venns of this type is  $\delta(2^{\delta-1})$ . It may be shown that:

$$\binom{\delta}{1} + 2\binom{\delta}{2} + \dots + \delta\binom{\delta}{\delta} = \delta(2^{\delta-1}).$$

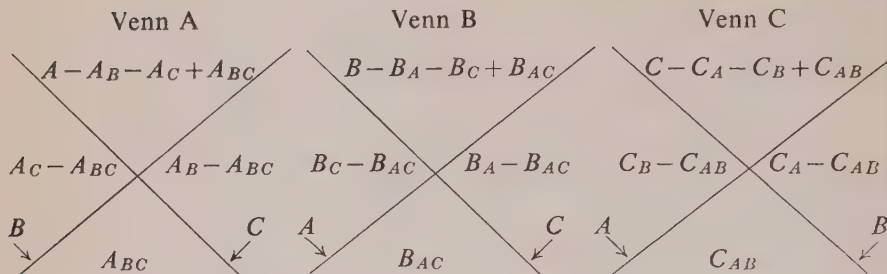
Consequently, the variables, such as those of equations (1)–(7), may always be chosen at will to construct the required  $(\delta - 1)$ -Venns, and a neuron may always be constructed for any such choice.

As a result, we have the

*Theorem.* Excitatory and inhibitory fibers to a neuron plus inhibitory interaction among them is sufficient for the construction of a neuron which computes the logical functions, for changing thresholds, described by any Venn diagram which requires spontaneous firing of the neuron for  $\theta \leq 0$ .

An interesting problem is the construction of neurons which are cheapest in the sense that they receive a minimum number of

fibers from the input. Suppose, for example, that excitatory and inhibitory fibers and inhibitions of input fibers cost 1c apiece. In the special case of  $\delta = 3$ , the three 2-Venns are:



and the total cost ( $P$ ) of fibers to a neuron is:

$$\begin{aligned}
 P = & |A_{BC}| + 2|A_C - A_{BC}| + 2|A_B - A_{BC}| + 3|A - A_B - A_C + A_{BC}| + \\
 & + |B_{AC}| + 2|B_C - B_{AC}| + 2|B_A - B_{AC}| + 3|B - B_A - B_C + B_{AC}| + \\
 & + |C_{AB}| + 2|C_B - C_{AB}| + 2|C_A - C_{AB}| + \\
 & + 3|C - C_A - C_B + C_{AB}|.
 \end{aligned}$$

Substituting from equations (1)–(7), this becomes:

$$\begin{aligned}
 P = & |A_{BC}| + 2|A_C - A_{BC}| + 2|A_B - A_{BC}| + 3|A - A_B - A_C + A_{BC}| + \\
 & + |B_{AC}| + 2|B_C - B_{AC}| + 2|A_B - A_B - B_{AC}| + \\
 & + 3|B + A_B - AB - B_C + B_{AC}| + |ABC - A_{BC} - B_{AC}| + \\
 & + 2|BC - B_C + A_{BC} + B_{AC} - ABC| + \\
 & + 2|AC - A_C + A_{BC} + B_{AC} - ABC| + \\
 & + 3|C + A_C - AC + B_C - BC + ABC - \\
 & - A_{BC} - B_{AC}|,
 \end{aligned}$$

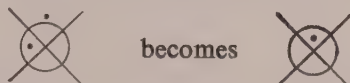
which may be solved to obtain the cheapest neuron. We have no explicit method for solving this equation for all  $\delta$ .

### 3. TRANSFORMATIONS

Transformations change the logical functions computed by a net and/or its component neurons, but do not change the reliability of the net.



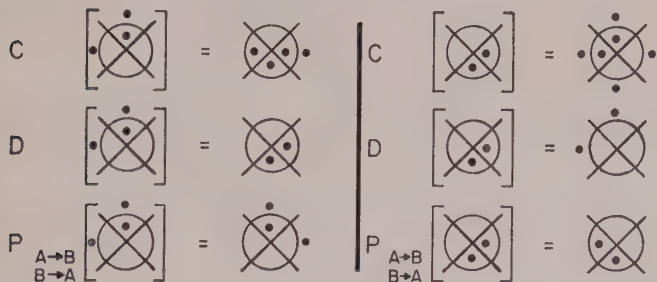
Any logical function may be represented using only “.”, “v” and “~”. Let  $X$  be a logical variable representing an input which is ON, and  $\sim Y$  the negation of a variable, or an input which is OFF. Then, we may represent a logical function ( $V$ ) by  $V(X, \sim Y, v, \cdot)$ . Thus if  $V(X, \sim Y, v, \cdot)$  represents  $A \cdot B \cdot \sim C \vee A \cdot \sim B \cdot C$ , then  $V(X, Y, \cdot, \cdot)$  represents  $(A \cdot B \cdot C) \cdot (A \cdot B \cdot C) = A \cdot B \cdot C$ . In terms of Venn diagrams,



Using this functional representation, we define complements, duals, and permutations as follows:

$$\begin{aligned}
 CV(X, \sim Y, v, \cdot) &= V(\sim X, Y, \cdot, v) \\
 DV(X, \sim Y, v, \cdot) &= V(\sim X, Y, v, \cdot) \\
 P_{\substack{X \rightarrow R \\ Y \rightarrow S}} V(X, \sim Y, v, \cdot) &= V(R, \sim S, v, \cdot).
 \end{aligned}$$

Examples of these transformations are:



From these defining equations, one obtains:

$$\begin{aligned}
 CDV(X, \sim Y, v, \cdot) &= CV(\sim X, Y, v, \cdot) = V(X, \sim Y, \cdot, v) \\
 DCV(X, \sim Y, v, \cdot) &= DV(\sim X, Y, \cdot, v) = V(X, \sim Y, v, \cdot).
 \end{aligned}$$

Thus,  $CDV + DCV$ . Similarly, one may show that  $CPV = PCV$ ,  $DPV = PDV$  and also that  $CCV = V$ ,  $DDV = V$  and  $PPV = V$ . It is now possible to put a net into a simple equation form and obtain transformations of the net. Let  $V_i$  represent the Venns of the first or input rank of neurons and  $V_o$  the Venn of the output

neuron, and let  $V_r$  be the input-output function computed by the net. A net with a single rank of input neurons and one output neuron is then symbolized by:  $(V_i)V_o = [V_r]$ . For example, the net of Fig. 7 may be written:

$$\left( \begin{array}{c|c} 1 & 2 \\ \hline 3 & 3 \\ \hline 2 & 1 \end{array} \right) \begin{array}{c|c} 2 & 3 \\ \hline 2 & 1 \\ \hline 2 & 3 \end{array} - \left[ \begin{array}{c|c} \cdot & \cdot \\ \hline \cdot & \cdot \\ \hline \cdot & \cdot \end{array} \right]$$

Let  $CV_i$ ,  $DV_i$  and  $PV_i$  respectively denote the complement, dual, and permutation of all Venns in rank  $V_i$ . We shall show that

$$\text{if:} \quad (V_i)V_o = [V_r] \quad (1)$$

$$\text{then:} \quad (V_i)CV_o = [CV_r] \quad (2)$$

$$(DV_i)V_o = [DV_r] \quad (3)$$

$$(CV_i)DV_o = [V_r] \quad (4)$$

$$(PV_i)V_o = [PV_r] \quad (5)$$

We shall say that  $V_o$  READS a community set of  $V_i$ , if the output neuron fires in response to an input configuration corresponding to that community set. In general,  $V_r$  contains a jot for each community set which  $V_o$  reads and a blank for each set that is not read.

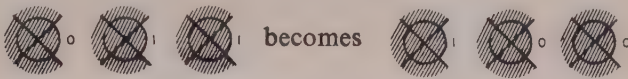
Equation (2) may be seen to hold from the following argument: A community set in  $V_i$  which is read by a jot in  $V_o$  will not be read when the jot becomes a blank. Conversely, a community set which is not read because of a blank in  $V_o$  will be read when the blank becomes a jot. Thus, each blank in  $V_r$  will become a jot and each jot a blank when  $V_o$  is complemented.

Next, we show that equation (5) holds.  $PV_i$  represents a permutation of the spaces in every Venn in  $V_i$ , and this permutation is the same for each such Venn. Therefore, community sets in  $V_i$  are permuted by  $P$ , but not otherwise changed. Since  $V_o$  is not changed, the jots of  $V_r$  must be permuted in correspondence with the community sets of  $V_i$ .

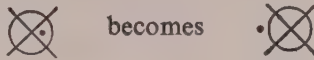
Equation (3) follows directly from equation (5), since duality is only a special case of permutation. Thus

$$DV(X, \sim Y, v, \cdot) = P_{\substack{X \rightarrow \sim X \\ Y \rightarrow \sim Y}} V(X, \sim Y, v, \cdot).$$

Next, we show that equation (4) holds. By taking the complement of  $V_i$ , a community set such as



or, in general, blanks are replaced by ones and ones by blanks in each community set. By taking the dual of  $V_o$ , a space such as



In general, a jot in  $V_o$  reads community set(s) of  $V_i$  corresponding to a particular input configuration. The same community sets are read by  $V_o$  when  $V_i$  and  $V_o$  are transformed as in equation (4).

The following examples are listed in the order of equations (1)–(5) and should clarify their meaning:

$$\left( \begin{array}{c} \circ \\ \times \end{array} \quad \begin{array}{c} \circ \\ \times \end{array} \quad \begin{array}{c} \bullet \\ \times \end{array} \right) \cdot \begin{array}{c} \bullet \\ \times \end{array} = \left[ \begin{array}{c} \bullet \\ \times \end{array} \right] \quad (1)$$

$$\left( \begin{array}{c} \bullet \\ \times \end{array} \quad \begin{array}{c} \bullet \\ \times \end{array} \quad \begin{array}{c} \bullet \\ \times \end{array} \right) \cdot \begin{array}{c} \bullet \\ \times \end{array} = \left[ \begin{array}{c} \bullet \\ \times \end{array} \right] \quad (2)$$

$$\left( \begin{array}{c} \times \\ \circ \end{array} \quad \begin{array}{c} \times \\ \circ \end{array} \quad \begin{array}{c} \times \\ \circ \end{array} \right) \cdot \begin{array}{c} \times \\ \circ \end{array} = \left[ \begin{array}{c} \times \\ \circ \end{array} \right] \quad (3)$$

$$\left( \begin{array}{c} \bullet \\ \times \end{array} \quad \begin{array}{c} \bullet \\ \times \end{array} \quad \begin{array}{c} \bullet \\ \times \end{array} \right) \cdot \begin{array}{c} \bullet \\ \times \end{array} = \left[ \begin{array}{c} \bullet \\ \times \end{array} \right] \quad (4)$$

The permutation of equation (5) will be  $P_{A \rightarrow C}^{C \rightarrow A}$ :

$$\left( \begin{array}{c} \times \\ \circ \end{array} \quad \begin{array}{c} \times \\ \circ \end{array} \quad \begin{array}{c} \times \\ \circ \end{array} \right) \cdot \begin{array}{c} \times \\ \circ \end{array} = \left[ \begin{array}{c} \times \\ \circ \end{array} \right] \quad (5)$$

If dashes are introduced into some of the spaces of a Venn to represent errors, then equations (1)–(5) still hold, and a single reliable set, under transformations, will yield other reliable nets. An even larger number of equations may be obtained by combining 2 or more of equations (1)–(5). For example, transforming equation (2) by equation (3) gives:

$$(DV_i)CV_o = [DCV_r]. \quad (6)$$

Transforming equation (6) by equation (4) gives:

$$(DCV_l)DCV_o = [DCV_r]. \quad (7)$$

Equation (7) has important applications for logically stable nets. To see this, note that:

$$\begin{aligned} DC \left( \begin{array}{c} \times \\ 1 \times 3 \\ 2 \end{array} \right) &= DC \left( \begin{array}{c} \times \\ \cdot \times \rightarrow \cdot \times \rightarrow \cdot \times \end{array} \right) = D \left( \begin{array}{c} \times \\ \cdot \times \rightarrow \cdot \times \rightarrow \cdot \times \end{array} \right) \\ &= \left( \begin{array}{c} \times \\ \cdot \times \rightarrow \cdot \times \rightarrow \cdot \times \end{array} \right) = \left( \begin{array}{c} 3 \\ 2 \times 1 \end{array} \right) \end{aligned}$$

Thus, equation (7) transforms

$$\left( \begin{array}{c} \times \\ 1 \times 3 \\ 2 \end{array} \right) \left( \begin{array}{c} 3 \\ 2 \times 1 \end{array} \right) \left( \begin{array}{c} \times \\ 2 \times 3 \\ 1 \end{array} \right) = \left[ \begin{array}{c} \times \\ \cdot \times \\ \cdot \times \end{array} \right] \quad \text{into} \quad \left( \begin{array}{c} 3 \\ 2 \times 1 \end{array} \right) \left( \begin{array}{c} \times \\ 1 \times 3 \\ 2 \end{array} \right) \left( \begin{array}{c} \times \\ 2 \times 3 \\ 1 \end{array} \right) = \left[ \begin{array}{c} \times \\ \cdot \times \\ \cdot \times \end{array} \right]$$

and, in general, logically stable nets are transformed into different logically stable nets by this equation. Equation (7) will also prove useful with regard to the polypheck to be discussed in Section 4.

#### 4. POLYPHECKS

In 1880, Charles Sanders Peirce<sup>(8)</sup> discovered one of the functions which we today call a Sheffer stroke. A few years later, he discovered the second of these functions. He called them amphecks (from *αμφήκης*, cutting both ways), and showed that either of them is sufficient for the formulation of the Calculus of Propositions. Actually, there are more functions on which the Calculus of Propositions may be based. We call them polyphecks.

The usefulness of the polypheck is that nets, which use a finite number of only one of these functions and time delays, are able to compute all functions.

It has been shown that the “~” and either “·” or “v” are sufficient, given the logical variables, to formulate the Calculus of Propositions. Consequently, any function which can complement and either AND or OR is also sufficient to formulate the Calculus. We shall find all functions, or polyphecks, for all  $\delta$ , which have this property.


The following lemma gives the necessary and sufficient conditions for a function to have the property that it may complement.

*Lemma 1.* Given the logical variables  $A, B, C, \dots$ , necessary and sufficient conditions for a Venn ( $V$ ) to have the property that it may complement any function ( $F$ ) are that it contains a jot in the NONE space ( $\sim B \cdot \sim B \cdot \sim C \dots$ ) and a blank in the ALL space ( $A \cdot B \cdot C \dots$ ).

*Proof.*

Necessary conditions—Suppose there are three logical variables

$A, B,$  and  $C$  or equivalently:  ,  , and  ,

and suppose that  $F =$  is the function to be comple-

mented. Let any number of these Venns (logical variables and/or  $F$ ) be placed in a row in  $V_i$ , and let  $V = V_o$  be the Venn which must operate on  $V_i$ . In general,  $V_o$  is of any value of  $\delta$ , depending on the number of Venns in  $V_i$ . We will show, by a *reductio ad absurdum*, that  $V_o$  must contain a blank in the ALL space and a jot in the NONE space.

Clearly, if  $V_o$  contains a jot in the ALL space, so does  $V_r$  since every Venn in  $V_i$  contains a jot in the ALL space. Next,  $V_r$  together with any number of the logical variables may be placed in another row ( $V_i'$ ) and operated upon by  $V_o$ . The same analysis may be used to show that if  $V_o$  contains a jot in the ALL space, then so does the input-output function  $V_r'$ . By induction, a jot in the ALL space of  $V_o$  always implies a jot in the ALL space of the input-output function. Thus,  $V_o$  cannot complement unless it has a blank in the ALL space. Similarly, if  $V_o$  contains a blank in the NONE space, then so does  $V_r$ , and by induction, the input-output function will always contain a blank in the NONE space. Therefore,  $V_o$  must have a jot in the NONE space in order to complement.

Sufficient conditions—Take any function whatever which must be complemented and put  $\delta$  copies of it in  $V_i$ . The jots of this function cannot be read by  $V_o$ , since  $V_o$  contains a blank in the ALL space. The remaining spaces of this function are the complement and must be read, since  $V_o$  contains a jot in the NONE space. *Q.E.D.*

Any function ( $V$ ) which can complement and which satisfies the



condition  $V \neq CDV$  is a polycheck. For example, when the number of logical variables is two ( $\delta = 2$ ), there are four functions which satisfy the complementation properties of the lemma:

$$\begin{array}{c} \times \\ \cdot \\ \times \end{array} \begin{array}{c} \times \\ \cdot \\ \times \end{array} \begin{array}{c} \times \\ \cdot \\ \times \end{array} \begin{array}{c} \times \\ \cdot \\ \times \end{array} \text{ of these, } \text{dc} \left( \begin{array}{c} \times \\ \cdot \\ \times \end{array} \right) = \begin{array}{c} \times \\ \cdot \\ \times \end{array}, \quad \text{dc} \left( \begin{array}{c} \times \\ \cdot \\ \times \end{array} \right) = \begin{array}{c} \times \\ \cdot \\ \times \end{array}$$

$$\text{while } \text{dc} \left( \begin{array}{c} \times \\ \cdot \\ \times \end{array} \right) = \begin{array}{c} \cdot \\ \times \\ \cdot \end{array} \text{ and } \text{dc} \left( \begin{array}{c} \times \\ \cdot \\ \times \end{array} \right) = \begin{array}{c} \times \\ \cdot \\ \times \end{array}$$

The latter two functions are the Sheffer strokes. To prove that necessary and sufficient conditions for a polycheck ( $V$ ) are: (1)  $V \neq CDV$  and (2) the complementation property, we require a lemma.

Let  $V^*$  be any Venn which contains a single jot. Then, the space of  $V^*$  which contains a jot and the space of  $DV^*$  which contains a jot are DUAL SPACES. It follows from the definition of duality, that for each space in a Venn there is one and only one dual space.

*Lemma 2†:* Let  $d_1$  and  $d_2$  be any two dual spaces in a Venn. Then,  $V = CDV$  if and only if  $d_1$  contains a jot and  $d_2$  contains a blank, or vice versa.

### *Proof*

Suppose  $d_1$  contains a jot and  $d_2$  a blank. In  $DV$ ,  $d_1$  contains a blank and  $d_2$  a jot. In  $CDV$ ,  $d_1$  contains a jot and  $d_2$  a blank.  $d_1$  and  $d_2$  were any two dual spaces. Therefore, every space in the Venn has been considered and  $V = CDV$ . Suppose a pair of dual spaces contains jots (or blanks). In  $DV$ , this pair contains jots (or blanks). In  $CDV$ , this pair contains blanks (or jots). Therefore  $V \neq CDV$ . *Q.E.D.*

### *Theorem*

A logical function ( $V$ ) is a polycheck, that is, given the logical variables,  $A, B, C, \dots$  it will serve to compute any logical function of the Calculus of Propositions, if and only if it contains a jot in the NONE space, a blank in the ALL space, and satisfies the condition  $V \neq CDV$ .

---

† I wish to thank Herman Berendsen of Holland for pointing out the necessity of this lemma and for its proof.

*Proof*

(1) We must show that  $V$  can complement and either AND or OR. By lemma 1,  $V$  can complement if and only if it contains a jot in the NONE space and a blank in the ALL space.

(2) We show that if  $V = CDV$ , then  $V$  cannot be a polycheck.

A logical variable such as  $A$  in Venn form ( $V_A$ ) contains a jot in every space of area  $A$  and a blank in all other spaces. By definition of duality, each pair of dual spaces has one member in the area of  $A$  and one member not in area  $A$ , i.e. in  $\sim A$ . Therefore, each pair of dual spaces in  $V_A$  contains one jot and one blank and by lemma 2 satisfies the condition  $V_A = CDV_A$ . In general, every logical variable  $V_A, V_B, V_C, \dots$ , or in general,  $V_P$ , satisfies the condition  $V_P = CDV_P$ .

Let  $V_i$  contain any number of logical variables in Venn form and let  $V = V_o$  read  $V_i$ . Let the output of the net together with the logical variables be placed in a second rank  $V_i'$  and let  $V_o$  read  $V_i'$ . Repeat this procedure any number of times. We show that there exist functions which cannot be computed by  $V$ .

Let  $V = V_o$  in the equation  $(V_i)V_o = [V_r]$ . By assumption  $V_o = CDV_o$  so that  $(V_i)CDV_o = [V_r]$ . Initially  $V_i$  contains only logical variables so that  $V_i = CDV_i$  and therefore  $(CDV_i)CDV_o = [V_r]$ . By equation (7) of Section 3,  $(CDV_i)CDV_o = [CDV_r]$  and therefore  $V_r = CDV_r$ . Next,  $V_r$  together with any number of the logical variables may be placed in a row of  $V_i$  and be operated upon by  $V_o$ . However, we still have  $V_i' = CDV_i$  and  $V_o = CDV_o$ . Thus, if  $(V_i')V_o = [V_r']$ , then  $V_r' = CDV_r'$ . By induction,  $V_o = CDV_o$  can compute at most complementary-dual functions and therefore cannot be a polycheck.

(3) We show that if  $V$  can complement and if  $V \neq CDV$ , then  $V$  can either AND or OR. Since  $V \neq CDV$ , then by lemma 2, there exists at least one pair of dual spaces  $(d_1, d_2)$  in  $V$  such that  $(d_1 = 1, d_2 = 1)$  or  $(d_1 = 0, d_2 = 0)$ .

(a) Assume  $(d_1 = 0, d_2 = 0)$ . We show that  $V$  can OR. Let  $V = V_o$  in  $(V_i)V_o = [V_r]$ , and let  $V_s$  and  $V_t$  be the Venns which must be OR'ed. The input configuration corresponding to  $d_1$  is represented by a number of Venns in  $V_i$ . The remaining Venns in  $V_i$  represent the input configuration corresponding to  $d_2$ , since  $d_1$  and  $d_2$  are dual spaces. Put  $V_s$  into all Venns of  $V_i$  read by  $d_1$  and put  $V_t$  into all the remaining Venns read by  $d_2$ . Because  $V_o$

contains blanks in the  $d_1, d_2$  and ALL spaces,  $V_o$  does *not* read the jot of  $V_s$  and/or  $V_t$ . Because  $V_o$  contains a jot in the NONE space,  $V_r$  will contain jots only in those spaces corresponding to spaces of both  $V_s$  and  $V_t$  which contain blanks. By complementation of  $V_r$  by  $V_o$  as in lemma 1, we obtain the OR of  $V_s$  and  $V_t$ .

(b) Assume ( $d_1 = 1, d_2 = 1$ ). We show that we can AND  $V_s$  and  $V_t$ . Put  $V_s$  into the places of  $V_t$  corresponding to  $d_1$ , and  $V_t$  into the remaining places corresponding to  $d_2$ . Because  $V_o$  contains a blank in the ALL space, it does not read jots contained in  $V_s$  and  $V_t$ . Because  $V_o$  contains jots in the  $d_1, d_2$ , and NONE spaces, it reads all remaining spaces of  $V_s$  and  $V_t$ . Therefore,  $V_r$  contains blanks only in those spaces corresponding to spaces of both  $V_s$  and  $V_t$  which contain jots, and by complementation of  $V_r$  by  $V_o$ , we obtain the AND of  $V_s$  and  $V_t$ . Q.E.D.

The importance of this theorem becomes clear if we count the number of polypecks for each value of  $\delta$ . The total number of functions with  $\delta$  arguments is  $2^{2^\delta}$ . Of these  $2^{2^\delta-2}$  contain a jot in the NONE space and blank in the ALL space. By lemma 2, a function having the property  $V = CDV$  has a jot in one member of a pair of dual spaces and a blank in the other member. Such functions in Venn forms are therefore half-filled with jots. The first jot may enter any one of the  $2^\delta$  spaces of the Venn, and a blank must be placed in the corresponding dual space. The second jot may enter any one of  $2^\delta-2$  spaces, and a blank must enter the dual space. Thus, the number of Venns ( $Z$ ) for which  $V = CDV$  is

$$Z = \frac{2^\delta(2^\delta-2)\dots 2}{(2^{\delta-1})!}$$

where  $(2^{\delta-1})!$  corresponds to the number of permutations of the  $2^{\delta-1}$  jots. Therefore,  $Z = 2^{2^\delta-1}$ . Of this number, exactly half contain a jot in the NONE space and, therefore, a blank in the ALL space. Thus, the number ( $N$ ) of the polypecks is

$$N = 2^{2^\delta-2} - 2^{2^{\delta-1}-1} = 2^{2^\delta} \left[ \frac{1}{4} - \frac{1}{2 \cdot 2^{2^{\delta-1}}} \right].$$

Most important, if one chooses a function at random, then for large  $\delta$  there is a probability of almost  $\frac{1}{4}$  (exactly  $\frac{1}{4} - (1/2 \cdot 2^{2^{\delta-1}})$ ) that it is a polypeck.

## REFERENCES

1. W. S. McCULLOCH, "What is a number, that a man may know it, and a man, that he may know a number?" Ninth Annual Alfred Korzybski Memorial Lecture of March 12, 1960, to be published in *The Journal of the Institute of General Semantics*.
2. W. S. McCULLOCH, The stability of biological systems. *Homeostatic Mechanisms*, (Brookhaven Symposia in Biology, No. 10) Off. of Tech. Serv. U.S. Dept. of Commerce, Washington (1958).
3. W. S. McCULLOCH, Biological computers. *I.R.E. Trans. E.C.6*. pp. 190-92. (1957).
4. W. S. McCULLOCH, Agathe Tyche of nervous nets—the lucky reckoners. *Mechanization of Thought Processes*, (Natl. Phys. Lab. Symposium No. 10), H.M.S.O. London (1959).
5. WARREN S. McCULLOCH, Infallible nets of fallible formal neurons. *Quarterly Progress Report* of Research Laboratory of Electronics, MIT, pp. 189-96, (July, 1959).
6. E. F. MOORE and C. E. SHANNON, Reliable circuits using less reliable relays. *J. Franklin Inst.*, 262, Part I, pp. 191-208, Part II, pp. 281-97 (1956).
7. J. VON NEUMANN, Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata Studies* (eds. C. E. SHANNON and J. MCCARTHY), Princeton (1956).
8. C. S. PEIRCE, *Collected Papers*, Vol. IV (eds. C. HARTSHORNE and P. WEISS) p.13, pp. 215-16. Harvard (1933).





**L. A. M. VERBEEK**

*Research Laboratory of Electronics,  
Massachusetts Institute of Technology, Cambridge, Massachusetts*

## ON ERROR MINIMIZING NEURONAL NETS\*

### 1. INTRODUCTORY REMARKS

A VARIETY of procedures for analysis and synthesis of networks built from fallible elements are followed in the literature on the subject. The differences in the results obtained are due to differences in the transfer functions of the elements, in the location and in the kind of the errors considered and in the transfer function of the network. It is not our goal to give a survey or an evaluation of the distinct approaches.

We will consider one specific kind of error in formal neurons (threshold devices) and give procedures for minimizing its effect on the functioning of networks constructed from these neurons. The nature and the location of this error is such that the investigation takes into account breakdown of the connections between neurons. That is, the structure of the net may be fallible.

The terminology we use is partly taken from neurophysiology. The introduction of probabilistic expressions and its consequences on the formulation of the results produce a convenient way of describing the happenings.

The stimulation and help of Dr. W. S. McCulloch, Manuel Blum and Jack Cowan was a much appreciated assistance in the investigation.

---

\* This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research); and in part by The Teagle Foundation, Inc., and the National Institutes of Health.

## 2. SINGLE LINE COMPUTATION

A formal neuron is a many-input single-output threshold device the output of which is a logical function of its input. Input and output lines can be in one of two states, "on" or "off", 1 or 0. If a neuron fires its output is "on", otherwise it is "off". An input line can excite (positive stimulation indicated by an arrow on the line) or inhibit (negative stimulation indicated by an  $\circ$  on the line) the neuron which fires if the algebraic sum of its input is equal to or exceeds its threshold  $\theta$ . The input-output function of a neuron is given by the Venn diagram associated with it. For explanation of the Venn diagram symbolism we refer to McCulloch<sup>(1)</sup> and Blum.<sup>(2)</sup> A neuronal net is a structure of interconnected formal neurons, the input-output function of such a network is also given by a Venn diagram.

In our discussion we assume synchrony of states of neurons and unit time delay in their activity, so that we have no timing problems.

There are logical functions which cannot be computed by a single formal neuron, at least not without "interaction of afferents", i.e. mutual inhibition or excitation of input lines, which means logical computation outside neurons (cf. (1) and (2)). Note that there is evidence of inhibition of afferents in biological nervous systems.<sup>(3)</sup> For neurons with two input lines 2 of the 16 possible functions cannot be computed by a single neuron, namely

$$\begin{array}{c} \times \\ \times \end{array} \quad \text{and} \quad \begin{array}{c} \times \\ \circ \end{array}$$

For three input line neurons these numbers are 152 out of the 256 possible functions. Such functions can be computed by simple neuronal nets (cf. for instance Fig. 1).

Errors in logical computations can occur as a consequence of changes in the connections of input lines to a neuron, in the strength of the signal on the input lines, and in the threshold value of a neuron. For a brief survey of this we refer to (2). Here we are not dealing with these sources of error but with malfunction of a neuron such that it sometimes fires when it should not, or fails to fire when it should fire. We will call this axonal error. It is similar to fit or death of biological neurons which suggested our investigation. The axonal error of a formal neuron may be looked upon as

a jump of its threshold value such that it assumes (temporarily or permanently) the value  $+\infty$  (death) or  $-\infty$  (fit). As this axonal error produces erroneous computation on correct arguments we are dealing with probabilistic logic, not with logic of probable arguments. We take the axonal error into account by assigning a probability of error  $\epsilon$  to the output line of each formal neuron and make two assumptions. Firstly we suppose that all neurons in a net have the same value  $\epsilon$  of error probability; if there is reason

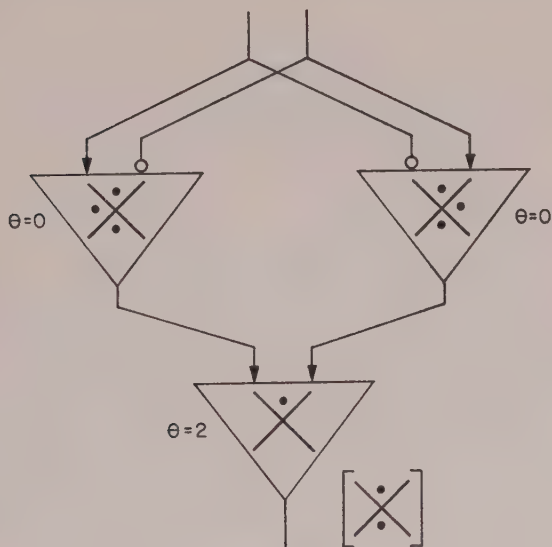


FIG. 1

to expect different values for different neurons we may take the highest value and be on the safe side as far as the results are concerned. Secondly we assume the probability of error to be independent of the previous and momentary activity of the neuron itself and of all other neurons in the net. This stochastic independency is assumed for easy handling of the problem. It is worth noting that complete correlation of errors in the form of simultaneous shifts of threshold values of all neurons in a net is considered and discussed by McCulloch<sup>(1)</sup> and by Blum.<sup>(2)</sup> The axonal error we are concerned with in this paper is identical with

the error von Neumann considered in (4); our subsequent argument follows more or less his discussion.

Let us assign a probability of error  $\eta$  to the output of a logical computer consisting of a neuronal network. This error probability  $\eta$  is due to the probability of error  $\epsilon$  of the formal neurons composing the network, and evidently exceeds the probability of error of the last neuron—the output neuron—of the net, i.e.  $\eta > \epsilon$ . We will now propose a redundant net computing the same function but yielding a final output error probability  $\eta'$  which approaches  $\epsilon$ . The most obvious method of bringing this redundancy of structure in the net is a sort of paralleling. For a discussion of structure of networks with respect to error we refer to Cowan.<sup>(5)</sup> The desired function is computed by the most simple adequate network the output of which has an error probability  $\eta > \epsilon$ . For

example the logical function  $\times$  is computed as in Fig. 1. By

computing this function  $m$ -fold and bringing the  $m$  output lines as input to one majority organ the error probability  $\eta'$  of the final output can be smaller than  $\eta$ . A majority organ is a formal neuron with  $m$  input lines which fires if a majority of these are "on". We only consider  $m$  odd, hence the majority is  $(m+1)/2$  or more. (The result of the procedure with an even number of lines is equal to that for the odd case with one line less.) As each of the  $m$  input lines of the majority organ has a probability of error  $\eta$  the probability  $p$  that a majority of them is erroneous is given by the cumulative binomial probability

$$p = \sum_{k=(m+1)/2}^m \binom{m}{k} \eta^k (1-\eta)^{m-k}.$$

In Fig. 2 the value of  $p$  is plotted versus  $\eta$  for  $m = 1, 3, 7, 25$  and  $49$ . The final error probability  $\eta' = \epsilon(1-p) + (1-\epsilon)p$  is plotted versus  $\eta$  in Fig. 3 for  $\epsilon = 0.2$  and values of  $m$  as in Fig. 2.

From the above arguments and Fig. 3 we can conclude: (a) This method of computing the desired function by paralleling  $m$ -fold and using a majority organ results in  $\eta' < \eta$  if for given  $\epsilon$  the number  $m$  is sufficiently large. This restriction on  $\epsilon$  for given  $m$

can be expressed by

$$\epsilon < \epsilon_c = \frac{1}{2} - \frac{2^{m-2}}{m \binom{m-1}{(m-1)/2}}$$

(b) With given  $m$  and  $\epsilon < \epsilon_c$  further decrease of  $\eta'$  can be obtained by repeating the procedure of using a majority organ. This method

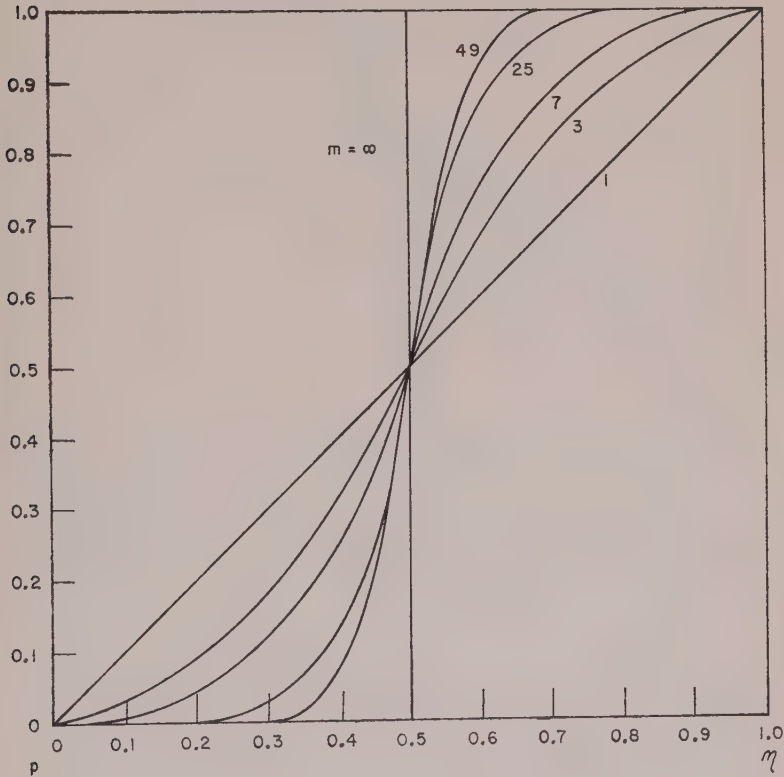


FIG. 2

is shown in Fig. 5. The result of this iteration is shown by the dashed line in Fig. 3 which indicates that the value of the ultimate error probability  $\eta'$  diminishes rapidly with the number of iterations.

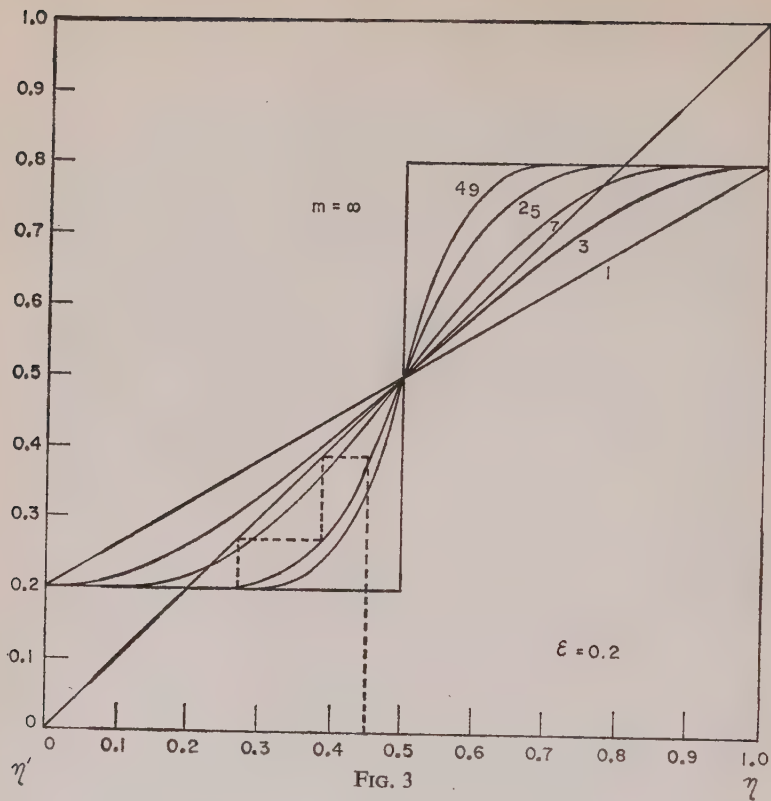


FIG. 3

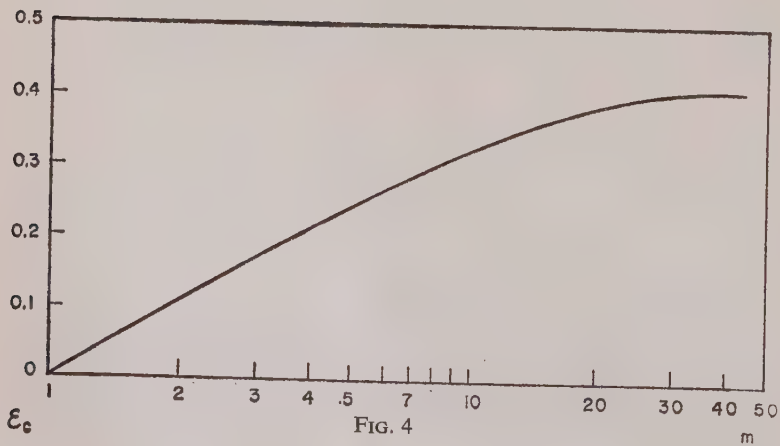


FIG. 4



The above procedure using majority organs with  $m$  input lines is an extension of the considerations von Neumann exposed in (4) where he used only majority organs with three input lines. In

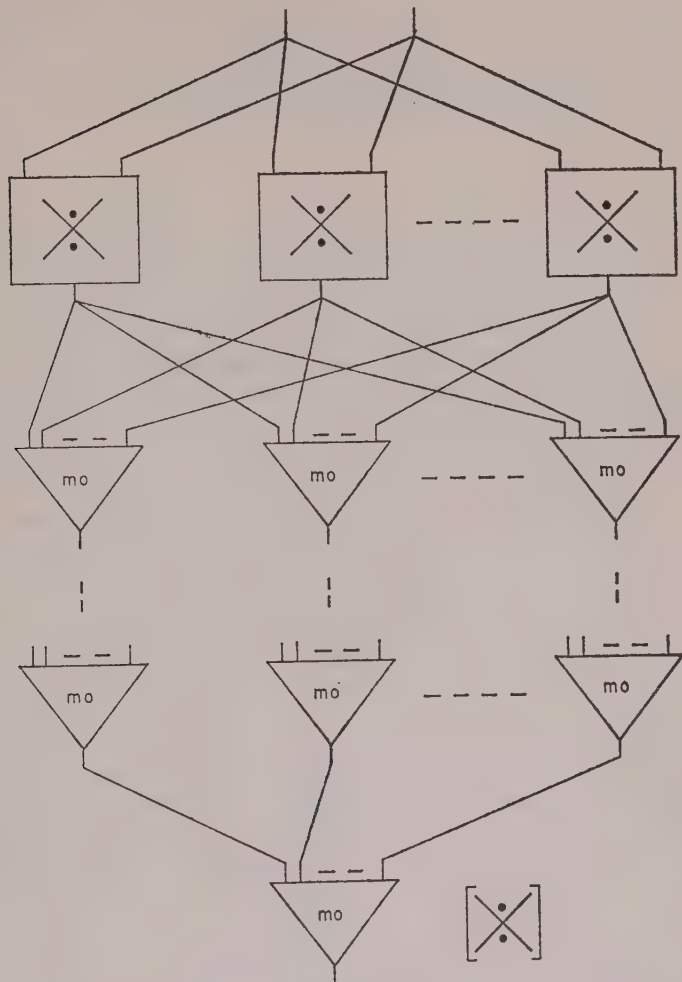


FIG. 5

our scheme a far lower number of elements is sufficient to obtain the same decrease of the total error probability. The reason for this difference lies only in the use of elements with many input lines

in which the effect of an erroneous line is much less than in the case of three input lines. A very important point is that with more input lines to a neuron the maintaining of the threshold level between its limits may be more difficult. The threshold of a majority organ is  $\theta = m/2 \pm < 0.5$ , and this shows that the admissible relative variation is smaller the larger  $m$  is, though the admissible absolute variation is independent of  $m$ .

### 3. BUNDLING AND ALL-TO-ALL PRINCIPLE

From the preceding discussion we know that we can reduce the error probability of a logical computer to the limiting value  $\epsilon$  of the output neuron by paralleling the computation and using a majority organ or, if necessary, an iteration with ranks of majority organs. In order to reduce the final error probability to an arbitrary small value the redundancy in the structure has to be applied in a different manner. Following von Neumann's procedure we therefore introduce "bundling". A two-valued logical variable will be carried by a bundle of  $N$  lines instead of by one single line. A fiduciary level  $\Delta$  ( $0 < \Delta < 0.5$ ) is used to express that the variable is "on" if  $(1 - \Delta)N$  or more lines of the bundle are "on", and the variable is "off" if  $\Delta N$  or less lines are "on". In the intermediate state where the number of excited lines lies between  $\Delta N$  and  $(1 - \Delta)N$  the variable is undetermined, carries no information. Computation with logical variables carried by bundles should use the full possibility this redundancy can deliver. This is achieved if the "all-to-all" principle is applied in the structure of the neuronal net. This means that all lines of all input bundles go to all computing neurons, insuring that all available information in the input bundles goes to all decision elements. For a discussion of this principle from the point of view of many-valued logic we refer to (5). In order to elucidate the idea we take the following example.

Suppose we wish to compute the logical "and",  $\wedge$ , from two bundles each consisting of 10 lines. We bring all lines of each bundle as excitations to a rank of 10 similar neurons as shown in Fig. 6. Assume for a moment that the neurons compute correctly, then the number of lines in error may maximally be  $\Delta N$  if the threshold

of the neurons is set such that  $10 + \Delta N < \theta \leq 20 - 2\Delta N$ . This can be achieved if  $\Delta < 1/3N$ . Let us take  $\Delta N = 3$  from which it follows that  $\theta = 13.5 \pm < 0.5$ . Note that such a constraint on the variations of the threshold may be severe for synthesis of the formal neurons. The larger threshold variations we allow the smaller  $\Delta$  has to be. Now all lines in the output bundle are absolutely error free if at most 3 of the 10 lines of each input bundle

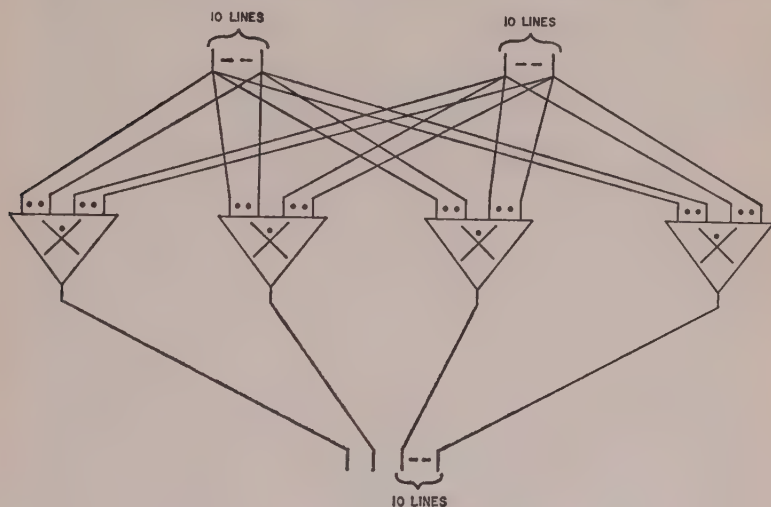


FIG. 6

are in error. If we now take into account the error of the formal neurons by assigning to each of them an independent error probability  $\epsilon$  the probability of malfunction  $P_e$ , i.e. the probability that more than  $\Delta N$  lines of the output bundle are in error, can be calculated by

$$P_e = \sum_{k=4}^{10} \binom{10}{k} \epsilon^k (1-\epsilon)^{10-k}.$$

For  $\epsilon = 0.005$  this results in  $P_e \approx 1.3 \times 10^{-7}$ .

From this example we see that by bundling and the application of the all-to-all principle it is possible to achieve a probability of malfunction  $P_e$  less than the probability of error  $\epsilon$  of the individual neurons.

Von Neumann's scheme consisting of neurons with two input lines, each having a probability of error  $\epsilon = 0.005$ . Using bundles of 5000 lines in a network, i.e. using 15,000 neurons, gives a probability of malfunction  $P_e \approx 4 \times 10^{-6}$ . Our single example is enough to indicate the order of magnitude of the simplification achieved by our scheme. This improvement comes about by the all-to-all principle of connectivity in our scheme which renders it insensitive to a relatively high number of erroneous lines in the bundles ( $\Delta < 1/3$  instead of von Neumann's most favorable  $\Delta = 0.07$ ). On the other hand our scheme uses neurons with many more input lines (20 instead of 2) and this suggests that it is not right to adopt the same  $\epsilon = 0.005$  for both schemes. Relative fluctuation of threshold value is more restricted if a neuron has more input lines, though the absolute fluctuation is not different. In our example  $\theta = 13.5 \pm < 0.5$  whereas in von Neumann's scheme this would be  $-1.5 + < 0.5$ . Though a complete comparison of the two schemes is interesting and tempting, the subject is too complicated to go into here. Moreover Cowan<sup>(5)</sup> deals with this topic.

It may be useful, however, to state that the given example can immediately be generalized to the statement that all logical functions of  $v$  variables (carried by bundles of  $N$  lines each) represented by Venn diagrams with 1 jot (and  $2^v - 1$  zeros) or with 1 zero (and  $2^v - 1$  jots) can be computed error free by a rank of  $N$  infallible neurons if the number of erroneous lines per bundle is  $\Delta N < N/(v+1)$ .

The proof of this statement reads: let the space in the Venn diagram containing the one jot to be the intersection of  $u$  variables. Then we bring those  $u$  excitatory bundles and the  $v-u$  inhibitory bundles to the neurons. Let the maximum number of erroneous lines per bundle be  $\Delta N$ . Then the computation is error free if and only if the threshold of the neurons is set such that

$$(u-1)N + \Delta N < \theta \leq u(N - \Delta N) - (v-u)\Delta N.$$

From this it follows that  $\Delta N < N/(v+1)$ . The proof for functions represented by one zero in the Venn diagram is analogous.

The probability of malfunction  $P_e$  for these networks is just cumulative binomial probability given on the preceding page. As there are many universal elements or polypchecks among these

functions (see (2)), we conclude that any logical computation can be performed by networks using bundling and the all-to-all principle with its inherent high reliability.

#### REFERENCES

1. W. S. McCULLOCH, Agathe Tyche of nervous nets—the lucky reckoners. *Mechanization of Thought Processes*, N.P.L. Symp. No. 10, H.M.S.O., London (1959).
2. M. BLUM, Properties of a neuron with many inputs. This volume, p. 95.
3. HOWLAND, LETTVIN, McCULLOCH, PITTS, WALL, Reflex inhibition by dorsal root interaction. *J. Neurophys.* **18**, pp. 1–17 (1955).
4. J. VON NEUMANN, Probabilistic logic and the synthesis of reliable organisms from unreliable components. *Automata Studies*, Princeton Univ. Press (1956).
5. J. D. COWAN, Many valued logics and reliable automata. This volume, p. 135.

#### DISCUSSION

VON FOERSTER: On this afferent inhibition, why doesn't one make here a neuron which is doing this job?

VERBEEK: That is why I did not talk about this, because it is just introducing new neurons.

VON FOERSTER: That is right, but if you want to introduce more afferent inhibition, you introduce more neurons, and therefore your calculations for  $N$  neurons may also be applicable for every net which may have afferent interaction.

McCULLOCH: The point is this: if we replace afferent inhibition by neurons, we are throwing away something which nature is using. While we are perfectly happy to throw it away for this purpose, we know, when we do so, that we are introducing error—another epsilon—with that neuron we put in between. That is, we have increased the noise in the works by doing so.

BOWMAN: I would like to ask a purely formal mathematical question here. There is a system set up here that looks an awful lot like group theory, but isn't. Have the mathematicians explored at all something that is related to group theory, but in which you combine three, not two, elements to obtain a new element of the group? This appears to me to be new, with the very little information I have on group theory and allied subjects, but I just ask if there have been formal studies of the consequences.

COWAN: I might point out that Karl Menger has shown that certain subsets of these things do form groups, and his son is working on this.

BOWMAN: I don't mean groups in the ordinary sense, because a combination of any two elements does not give rise to a third one. There must be combinations of an odd number of elements to give rise to a new member of the group. This is not group theory, but it comes awfully close to it.

AMAREL: I have had an opportunity to work with the set of these three-to-one relationships. Its structure can be easily described in lattice terms. As a matter of fact, I have used these relationships in my investigation of automatic theory formation, and I plan to discuss this in more detail later, in my talk here.



NOVIKOFF: I await with interest Amarel's discussion that this can be a lattice. It should be pointed out that a lattice is an algebraic system with a binary operation.

The question was asked whether there has been a formal study of ternary or higher  $n$ -ary operations. Asked in its broadest possible sense, the answer is yes. I do know of at least two different ternary systems that have been studied: one is a weakening of the notion of a Lie algebra, which involves putting together three objects to reproduce again; and the other is again a rather abstract algebraic notion which deals with ternary, and, if you wish,  $n$ -ary operations. So there has certainly been some discussion of the mathematical subject of going beyond binary relations.

MULLIN: Getting back to these neural nets, I want to work in a comment on the concept of statistical independence, which underlies their treatment here. I believe, and this is a matter of opinion, that the great progress that has been brought about by the concept of statistical independence in the first fifty years of this century will just about match its retardation effect during the next fifty years.

AMAREL: I wanted to comment on the situation treated by Blum. He is asking the question, "What is the total variation of thresholds that still gives a stable function?" One could ask the question, "Given some probability distribution of thresholds, what is the probability of a correct output?" This is closer to the type of question Verbeek is asking.

MCCULLOCH: Well, the first question, I believe, is directed to what are called logically stable circuits, stable under a common shift of threshold.

AMAREL: The questions are asked in a different way. In the first case you are asking about the total variability that you can allow to the system and still have exactly the answer you are interested in. The second question asks about the probability that the function you are interested in will be correct; this is a different approach to the problem.

Now in the network that you have considered, you have neurons. You assume then that the variation from one function to the next follows a definite order given by the way in which neuron functions correspond to consecutive threshold values. In general, suppose that you have a case where the distribution of states, logical states, does not coincide with that of a neuron: for instance, a case where there could be a transition from "or" or "exclusive or" . . .

MCCULLOCH: I can't devise one.

AMAREL: Well, we can devise elements of this type technologically; these can be certain types of resistor-rectifier gate circuits, for instance. The point I want to make is that for different distributions of logical states (which depend on the particular device we are treating) we have different properties of logical stability. We have studied logical stability in devices other than formal neurons; Maitra of our laboratories has demonstrated logical stability properties in rectifier-resistor gates under certain failure conditions of the rectifiers.

COWAN: This has to do with some work Schutzenberger did last fall at MIT on the particular types of functions you would get from a randomly connected set of switching elements. He found that he could state some kind of ergodic theorem for the system. Only in very special cases did you get anything other than tautology and contradiction out of it.

VON BERTALANFFY: You have some neurodynamics in mind; would you mind giving us one or two cases of the physiological paths they work in?

MCCULLOCH: Yes. We have done a lot of work on the interaction of afferents



which was first detected by Matthews, and then it comes up in the work of Lloyd, where it accounts for inhibition where there is no time for intervening neurons to be playing in the game at all. Now, we ran it down, and showed that in the cat spinal cord, the block occurred at the first point of division of the axon as it enters the dorsal column—it can be as far out as that—and we know that if you knock this out with strychnine, the animal goes into convulsions. We know it is an important gating method. Strychnine doesn't do anything else except to knock out this kind of gate. The body uses it all the time, and we know that with this—this is Manuel Blum's work—and the excitation and inhibition of the cell directly, we can produce a diagram which will bring the jots into any Venn diagram in any prescribed sequence. Without that, nature couldn't have done it.



**JACK D. COWAN**

*Research Laboratory of Electronics  
Massachusetts Institute of Technology*

## MANY-VALUED LOGICS AND RELIABLE AUTOMATA

**Abstract**—Many-valued logics are used to analyse discrete noisy automata. It is shown that the functionally incomplete logic of Lewis, in which not all truth-values are informationally significant, provides an appropriate model for the description of the redundant automata of von Neumann and Elias. A study is made of methods of reducing the redundancy in such automata, and it is shown that the functionally incomplete Lukasiewicz logic, and the functionally complete Post logics provide models for the description of the more efficient automata of McCulloch *et al.* Advantage is taken of the fact that these logics have varying structures corresponding to different ways of coding and processing information. An approach is made to the problem of achieving arbitrarily high reliability of computation with nets of unreliable components, so that these nets are not completely redundant. No solution is obtained to this problem, but it is concluded that the functionally complete Post logics may provide one.

### INTRODUCTION

It has been customary to apply information theory (that is, that body of theory concerned with the generation, storage, transmission and processing of a quantity described by a particular information measure<sup>(1)</sup>) to the study of living organisms, and in particular to the nervous system. What is usually neglected is the fact that the nervous system is a computer, operating in an environment consisting both of meaningful and meaningless stimuli.\* Moreover it is a computer which responds more reliably to highly meaningful stimuli than to less meaningful stimuli, and which preserves this reliability of computation in the presence of both external and internal noise, over an extended period of time. But

---

\*Meaning is here used in the sense of Mackay.<sup>(2,3)</sup>

the solutions so far obtained to the problem of obtaining reliable computation from unreliable components (that is, those solutions aimed at solving the internal noise problem) are not of the same character as those coding solutions, obtained from information theory which permit reliable communication in the presence of noise, apart from one fragmentary result.<sup>(4)</sup> In terms of information theory, the reliability obtained from redundancy of computation is not obtained from redundancy of code or of channel.<sup>(5)</sup> We shall attempt to provide some insights into this rather surprising situation.

Probabilistic logical functions can be represented in a number of ways, among which is McCulloch's chiasmic symbolism.<sup>(6)</sup> Thus the symbol



represents a binary logical functor

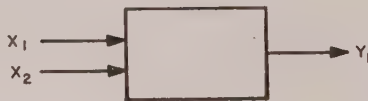


FIG. 1.

whose output  $y_1$  equals "1", with relative frequency  $1 - \epsilon$ , whenever the input is  $x_1$  or  $x_2$  or both, and equals "0" with relative frequency  $1 - \epsilon$ , whenever the input is neither  $x_1$  nor  $x_2$ . Alternatively, the symbol can be interpreted as the function " $x_1$  or  $x_2$ ".



such that with each possible input-output relation, there is associated a precise probability or error  $\epsilon$ . Another representation

of this function can be obtained, using the noisy channel symbolism of Shannon.<sup>(1)</sup>

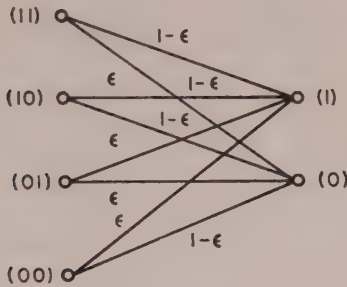


FIG. 2.

It is clear that we are here dealing with computing devices whose function is only probable, not with functions of probable arguments;<sup>(7-9)</sup> and hence the error is always effectively at the output:

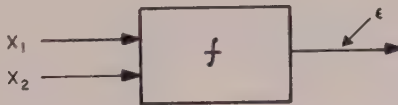


FIG. 3.

The question can now be posed: is there some redundant scheme  $f^+$  which is so organized as to produce an output of "1"s with a relative frequency approaching 1, for a given class of inputs, and an output of "0"s with a relative frequency approaching 1, for the complementary class of inputs? This question has been answered in various ways by von Neumann,<sup>(10)</sup> Elias,<sup>(11)</sup> Petersen,<sup>(12)</sup> Eden<sup>(4)</sup> and McCulloch *et al.*,<sup>(13-15)</sup> whose results, except Eden's, all have the same qualitative character: namely that the probability of error goes to zero only as the ratio of informationally significant elements to total number of elements goes to zero. This ratio is essentially the number of bits per channel symbol (to use information-theoretic terms), or computation rate, and so

these results all have the property that *the probability of error goes to zero with the rate*:

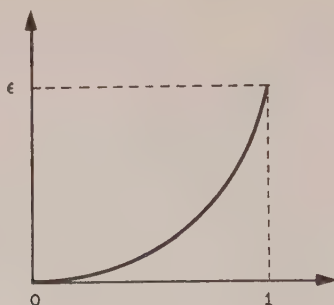


FIG. 4.

If we consider the results of information theory for reliable communication in the presence of noise,<sup>(1)</sup> this property seems rather surprising. In fact the coding theorem for noisy channels can be stated as follows: If, for a communication system, there exists a certain maximum rate for transmission of information (capacity  $C$ ), then for transmission rates less than  $C$ , it is possible to introduce redundancy, *independent of these rates*, so that the probability of error  $P_e$  goes to zero inversely with the redundancy. At rates above  $C$ ,  $P_e$  increases with  $R - C$ :

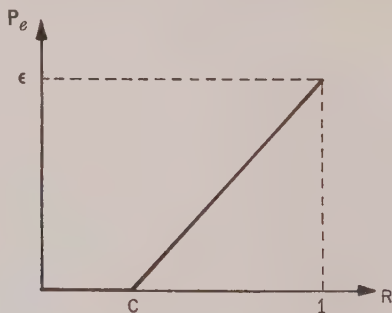


FIG. 5.

On the basis of the results of von Neumann and others<sup>(10-15)</sup> it appears that the maximum rate for reliable computation in the presence of noise is zero. That is, in general, the capacity of a noisy computation channel is zero.



COMPARISON OF ELIAS' AND VON NEUMANN'S SCHEMES

Let us consider some of these investigations in more detail. We shall first consider that of Elias, *loc cit.*, not because of historical precedence but because it illustrates nicely the various requirements of information theoretical coding schemes. The central theme of such schemes is that inputs are matched to noisy channels, in order to minimize the effects of noise present in these channels. That is, it is necessary to select from the set of possible inputs, a subset which produces minimum error probability after transmission, and hence there must exist an encoder. Similarly, there must exist a decoder which reconstructs required outputs from the set of possible channel outputs. Redundancy can exist either in code or in channel, or in both. In terms of the computation model, this means only redundancy of argument. Redundancy of computation (function) is somewhat different, and we shall consider this later. We thus have the scheme illustrated in Fig. 6:

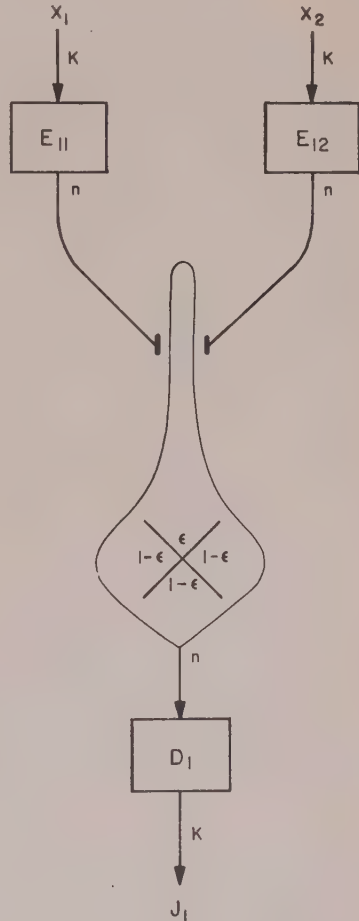



FIG. 6.

A block code is used, and the encoders  $E_{11}$ ,  $E_{12}$  convert input sequences  $x_1$ ,  $x_2$  respectively, each of length  $\kappa$ , into sequences of

length  $n$ . The decoder  $D_1$  reconverts a sequence of length  $n$  into a sequence of length  $\kappa$ . The following assumptions are made: (a) the mapping from input sequences to output sequences is  $\{2 : 1\}$  fixed, and is independent of past inputs and the operation of the computer. (b) The encoders and decoder are noiseless devices, and

no part of the computation  occurs in them. Thus

these devices are  $\{1 : 1\}$  from  $\kappa$  to  $n$ , and  $n$  to  $\kappa$  respectively.

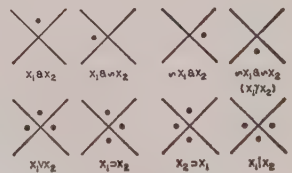
The operation of the code is conveniently described in terms of properties of sequence of binary digits, namely *weight* and *distance*.<sup>(16,17)</sup> The weight of a sequence  $w(a)$  is defined as the numbers of 1's in it. The distance between two sequences  $d(a, b)$  is given by the weight of the sum (modulo 2), i.e.  $d(a, b) = w(a \oplus b)$ . Encoding consists of selecting sequences of length  $n$ , corresponding to the given  $\kappa$ -sequences, whose mutual separation, as given by  $w(a \oplus b)$  is maximal. Decoding consists essentially of identifying certain subsets of the set of  $2^n$  outputs, with sets of  $n$ -sequences corresponding to given  $\kappa$ -sequences on a minimum distance basis. For example, consider the code with  $n = 3$ ,  $R = \frac{1}{3}$ . Encoding is given by  $\{1 \rightarrow 11, 0 \rightarrow 000\}$ , and decoding by  $\{111, 110, 101, 011 \rightarrow 1; 000, 001, 010, 100 \rightarrow 0\}$ . Thus all 3-sequences of weight 2 or more are decoded as 1, and all 3-sequences of weight at most 1 are decoded as 0.

Now the 16 functions of 2-variable, 2-valued logic can be classified according to the parity of the jots in their chiasitic representation,<sup>(6)</sup> i.e. in terms of their weights. Thus we have the 8 even-weight functions and the 8 odd-weight functions:

Important functions are the modulo 2 adder  $x_1 \oplus x_2$ , the Pierce ampheck  $x_1 \gamma x_2$ , and the Sheffer stroke function  $x_1 | x_2$ , the last



8 even-weight functions



8 odd-weight functions

two of which are universal elements. Elias' results can now be stated. These are that an information-theoretic code can be obtained only for the even-weight functions. Thus if  $n$  is the block length, and  $\kappa$  is the number of informationally significant digits per block,  $\kappa \leq n-1$ , and  $R \leq 1-1/n$  may be held constant, such that  $\lim_{n \rightarrow \infty} P_e = 0$  for  $R < C$ , the channel capacity. In fact

$P_e = 2^{-\alpha n f(R,C)}$  ( $\alpha = \text{const}$ ). For the odd weight functions, however,  $\kappa \leq n/2$ , and hence  $R \leq \frac{1}{2}$ . Furthermore  $P_e = 2^{-\alpha/R}$ , so that  $\lim_{R \rightarrow 0} P_e = 0$ . That is,  $P_e$  goes to zero with  $R$ . since universal elements are odd-weight functions, this latter result holds true in general. On the basis of this, Elias hypothesized that under the given assumptions, the computational capacity is zero.

Von Neumann's construction of a reliable automaton from less reliable components (VNC) may be compared with this. We shall briefly summarize the essential features of VNC. The basic functions used were the Sheffer stroke organ, and the ternary median operator or majority organ

$$(x \& y) \vee (y \& z) \vee (z \& x) \equiv (x \vee y) \& (y \vee z) \& (z \vee x).$$

We shall confine our investigation to schemes based on the former organ. The results obtained using majority organs differ only in certain minor details. The single line automaton shown below

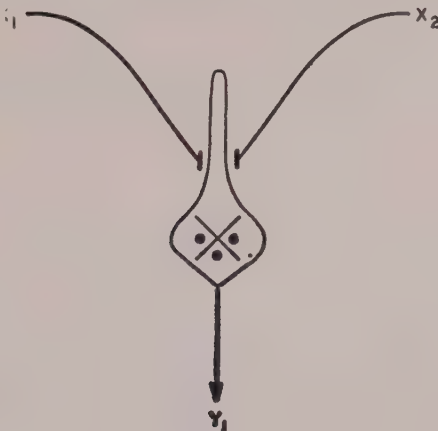


FIG. 7.

is replaced by the following multiple line automaton

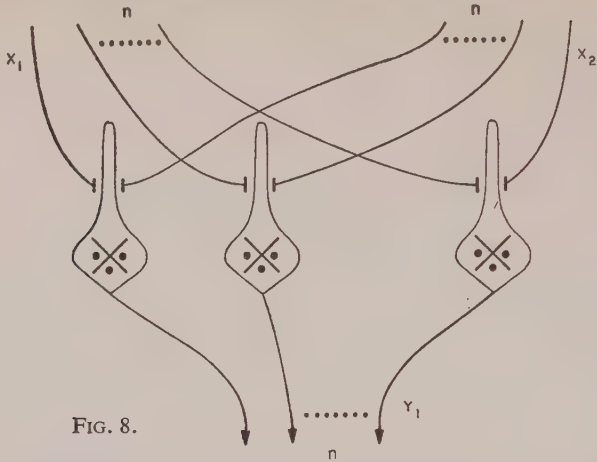


FIG. 8.

that possesses  $n$  lines per bundle. A fiduciary level  $\Delta$  is set so that, if at least  $(1 - \Delta)n$  lines “fire”, this is interpreted as a “1”; if at most  $\Delta n$  lines “fire” this is interpreted as a “0”, and if any intermediate firing level occurs, this is interpreted as a malfunction “0”. An informational constraint has therefore been introduced, producing one bit per bundle; that is, the rate is  $1/n$  bits per channel symbol. By virtue of this method of coding, even though the basic Sheffer-stroke organs are functioning without error, if the input is not all 1’s or 0’s, that which is designated as information  $\{1, 0\}$ , sometimes maps into malfunction  $\{i\}$ . Von Neumann termed this a *degradation of information*, and introduced *so-called restoring organs* to counteract this. The process can be represented as follows:

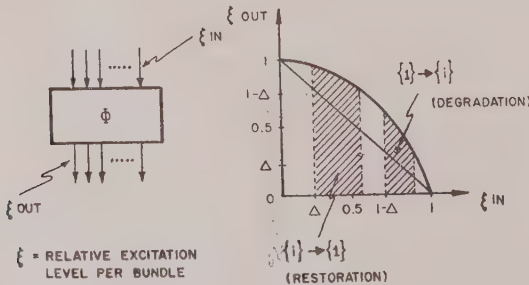


FIG. 9.

A restoring process is evidently required, with the following characteristic:

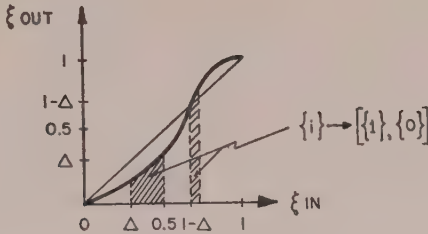


FIG. 10.

Such a restoring operation is obtained by using the following scheme:

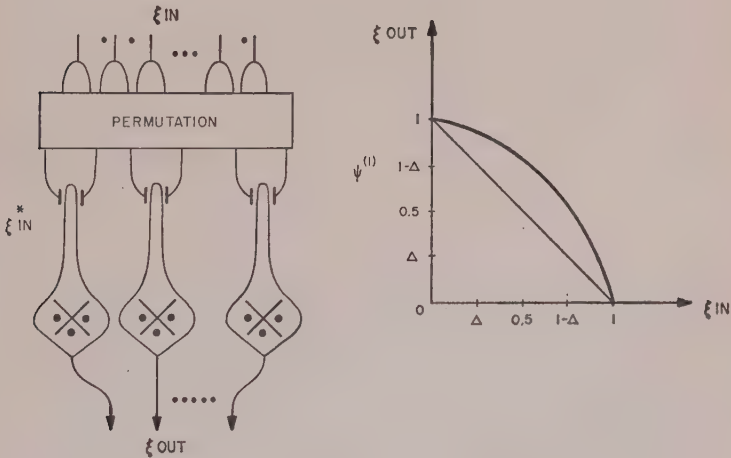


FIG. 11.

and iterating twice, to obtain the required function shown in Fig. 10. Two stages are necessary since  $\psi^{(1)}$  degrades as well as restores. In general  $\psi^{(2\kappa)}$  ( $\kappa$  finite) operations per bundle will restore all degraded information. The nature of the permutation is such that

(a) It ensures that  $\xi_{in}^*$  is a statistically independent random variable so that statistical approximations can be used to derive  $\xi_{out}$ .

(b) It is compatible with the following scheme

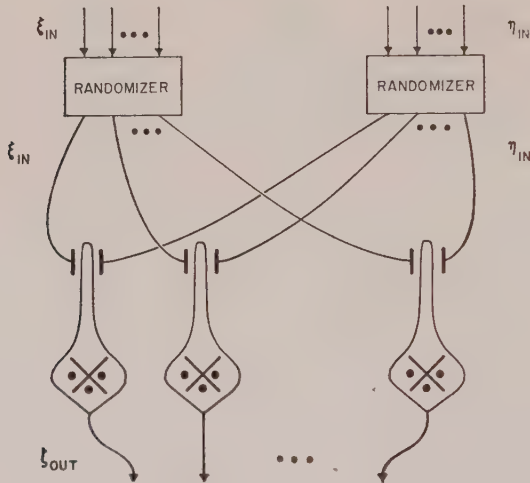


FIG. 12.

which ensures that  $\xi_{in}^*$ ,  $\eta_{in}^*$  are statistically independent random variables, and can be taken as a special case in which  $\xi_{in} \equiv \eta_{in}$ . The quantitative results for VNC are as follows. Let  $\xi$ ,  $\eta$ ,  $\zeta$  be the relative excitation levels of bundles corresponding to  $x_1$ ,  $x_2$  and  $y_1$  respectively; and suppose that  $\xi$  and  $\eta$  are statistically independent Gaussian random variables.

(a) Suppose the basic organs are noiseless. Then  $\zeta$  is approximately normally distributed, with a mean value given by  $\bar{\zeta} = 1 - \xi\eta$ , and a standard deviation

$$\sigma = [\xi(1-\xi)\eta(1-\eta)]^{1/2}/n^{1/2}.$$

Let  $\zeta_0$  be the relative excitation level of  $\psi^{2\kappa}y_1$ . Then  $\zeta_0$  is given by

$$1 - (1 - (1 - \xi\eta)^2)^2 = 1 - (2\xi\eta - \xi^2\eta^2)^2.$$



(b) Suppose the basic organs are noisy, with error probability  $\epsilon$ . Then

$$\xi = (1 - \xi\eta) + 2\epsilon(\xi\eta - \frac{1}{2})$$

and

$$\sigma = \{[(1 - 2\epsilon)^2 \xi(1 - \xi)\eta(1 - \eta) + \epsilon(1 - \epsilon)]/n\}^{\frac{1}{2}}$$

Similarly

$$\zeta_0 = 1 - (1 - \{(1 - \xi\eta) + 2\epsilon(\xi\eta - \frac{1}{2})\}^2)^2;$$

with given  $\sigma$ . The desired  $P_e$  is now obtained by choice of  $\Delta, n$  as follows: the required behavior of the net is of course

	1	0	$x_2$
1	0	1	
$x_1$	0	1	1

For sufficiently small  $\epsilon$ ; and sufficiently large  $n$ ,

$$\zeta_0 = 1 - (2\xi\eta - \xi^2\eta^2)^2$$

leading to the following table,

		$\eta$	
		$\geq 1 - \Delta$	$\leq \Delta$
$\xi$ :	$\geq 1 - \Delta$	$0\{\Delta^2\}$	$1 - 0\{\Delta^2\}$
	$\leq \Delta$	$1 - 0\{\Delta^2\}$	$1 - 0\{\Delta^4\}$

For sufficiently small  $\Delta$ , we can achieve the required behavior. It turns out in fact that the most favorable value for  $\Delta$  is 0.007, leading to  $\epsilon \leq 0.0107$ . For  $\epsilon = 0.005$ ,  $\Delta = 0.07$ , the resulting error probability is  $P_e \doteq 2^{-\beta n}$ ,  $\beta \sim 3.5 \times 10^{-3}$  (cf. Blum), that is  $P_e = 2^{-B/R}$ , which is similar to the results obtained in Elias' construction (EC).

We may compare EC and VNC via the information theoretic trilogy of encoder, channel and decoder. We have noted what were

the encoding and decoding functions for EC for VNC, the following simple interpretation results. The operation of encoding consists simply of splitting a line into  $n$  strands, thus mapping 1 into 111...11, and 0 into 000...0. The decoding operation (never explicitly mentioned in von Neumann, *loc. cit.*) consists simply of mapping  $\xi n \geq (1-\Delta)n$  into 1,  $\xi n \leq \Delta n$  into 0, and intermediate  $\xi n$  arbitrarily into 1 or 0. That is, there exist "don't care" conditions for these  $\xi n$ . In coding language, all  $n$ -sequences of weight  $(1-\Delta)n$  or more are decoded as 1, all  $n$ -sequences of weight  $\Delta n$  or less as 0, and all  $n$ -sequences of intermediate weight are decoded arbitrarily either as 1 or as 0.

The reason for the existence of these "don't care" conditions is clear. For large  $n$ ,  $n$ -sequences of intermediate weight occur with a vanishingly small probability, and so the decoder can neglect them. In contrast to this, the decoder EC receives with finite probability,  $n$ -sequences of all possible weights, and from these reconstructs the desired  $\kappa$ -sequences. Thus we may say that VNC achieves a probability of malfunction equal to  $2^{-B/R}$ , before the operation of decoding. Methodologically then, VNC differs from EC in that the former is essentially a combination of what we may term "channel-matching" and "source-matching", while the latter is the conventional "source-matching" scheme of information theory. The structures of the two schemes, however, apart from this difference of decoders, are essentially equivalent, as we shall demonstrate.

We shall conveniently represent both EC and VNC as follows. The sets of possible input configurations may be represented as the lattice points of  $n$ -dimensional hypercubes:

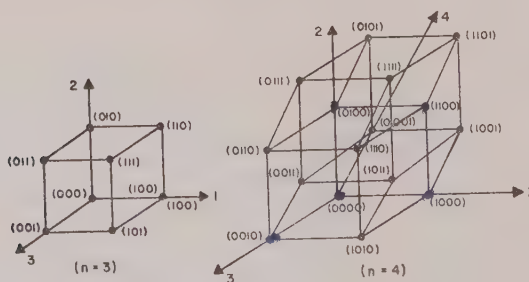


FIG. 13.

or equivalently by Boolean lattices (see Appendix):

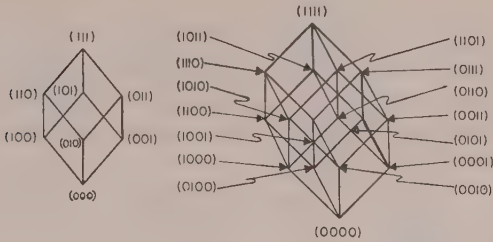
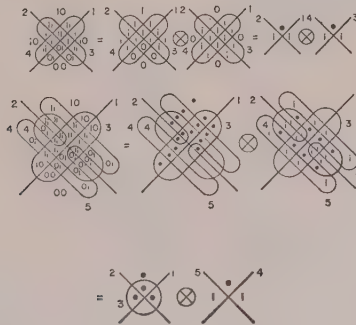


FIG. 14.

VNC and EC may then be represented as in Fig. 15. The essential structural equivalence of the two codes is easily seen.

The apparent complexity of  $D_4''$  and  $D_5''$  is spurious. Clearly



thus  $\{E_4'', D_4''\}$  and  $\{E_5'', D_5''\}$  are simply

$$\{ \subseteq ; \times \otimes \times \}, \{ \subseteq ; \otimes \otimes \}$$

as expected.

Of course, for such small values of  $n$ , the use of don't care conditions produces additional errors in VNC, and the full benefit

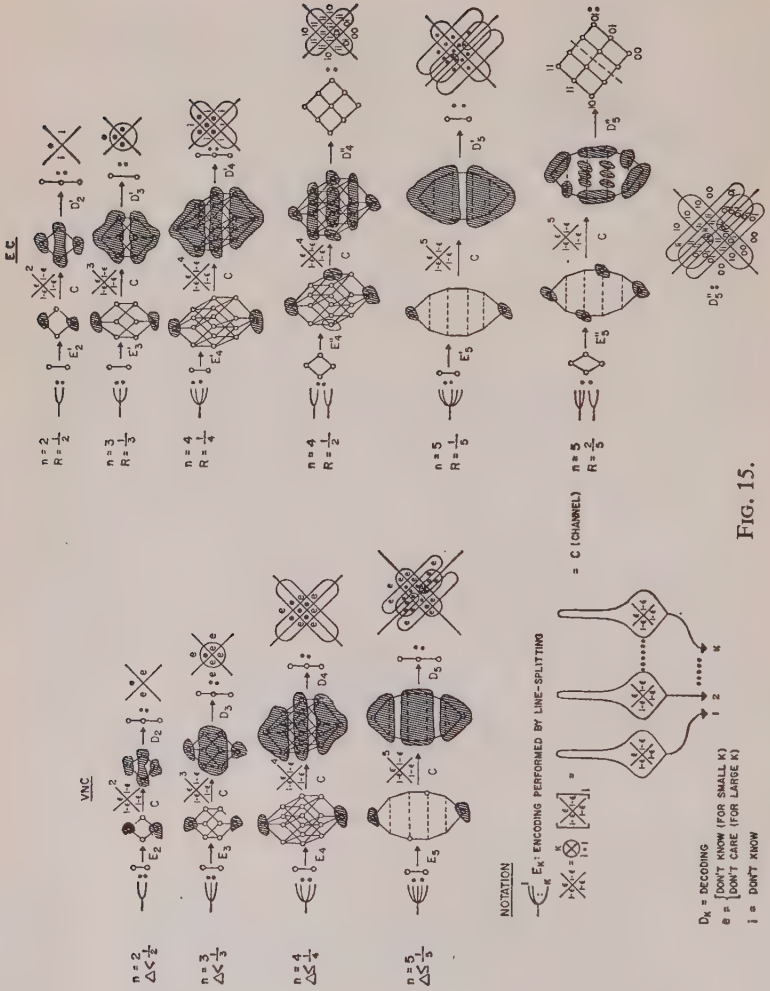


FIG. 15.

of the scheme is not obtained until  $n$  approaches  $10^3$  or  $10^4$ . For small  $n$ , instead of don't-care, we need an extra symbol  $e$  say, for intermediate, as in EC. Thus, for small  $n$ , both schemes utilize 3-state decoders, and for large  $n$ , only EC uses such a decoder, the other uses a 2-state decoder. Thus the only difference between EC and VNC, for computations of *unit logical depth*, consists entirely of a relatively small difference in the structure of the decoder. We may characterize both schemes as follows:

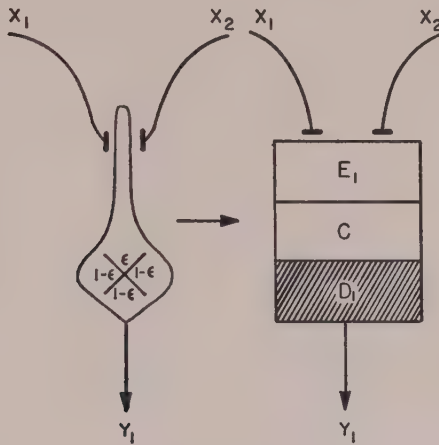


FIG. 16.

The noise in this scheme lies entirely in  $C$ . This presents no problem as far as  $E_1$  is concerned, since this is just line splitting. However,  $D_1$ , as we have seen, is at least as complex as  $C$ , and must be relatively infallible with respect to it. Since we cannot introduce a paradox by encoding and decoding for this organ, we must assume that  $D_1$  has been made relatively infallible by some kind of *microlevel redundancy* (Lofgren<sup>(18)</sup>). This need of an effectively noiseless decoder presents no problem for computations of unit logical depth, but for situations involving computations of arbitrary logical depth, we run into certain problems.

Consider the following computation:

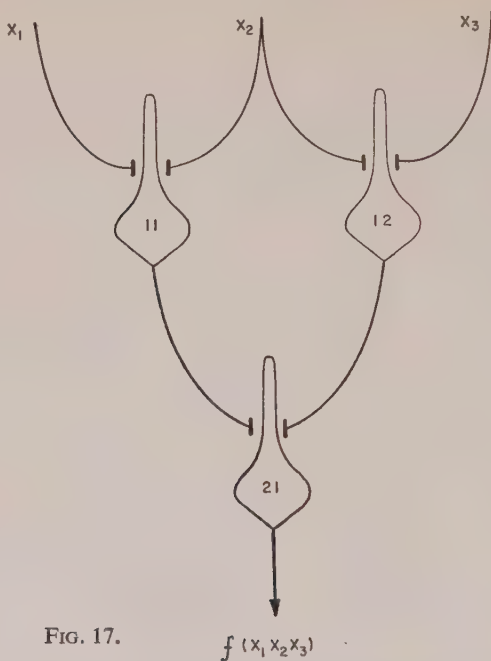


FIG. 17.

Following EC, we should replace each element by a triad, obtaining the following network:

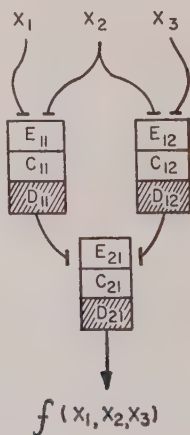


FIG. 18.



It is now necessary that only elements of the computing ranks  $C_{ij}$  be noisy. This presents no problem, as before for the  $E_{ij}$ , since we have assumed line splitting to be noiseless, but now the successive decoding ranks  $D_{ij}$  must be noiseless with respect to  $C_{ij}$ , otherwise errors may propagate through the net. Since the  $D_{ij}$  are at least as complex as the  $C_{ij}$ , this requirement places a rather unsatisfactory constraint on the location of errors in the net, and it would certainly be more satisfactory to permit a homogeneous error distribution throughout the net. In fact VNC allows this, and goes as follows: replace 11 and 12 by  $\{E_{11}, C_{12}\}$  and  $\{E_{12}, C_{12}\}$ , insert restoring organs  $\psi_{ij}^{2\kappa}$ , randomize inputs and outputs of these (cf. p. 144), replace 21 by  $C_{21}$ , randomize, insert  $\psi_{2j}^{2\kappa}$ , then add  $D_{21}$ . We thus obtain the following network:

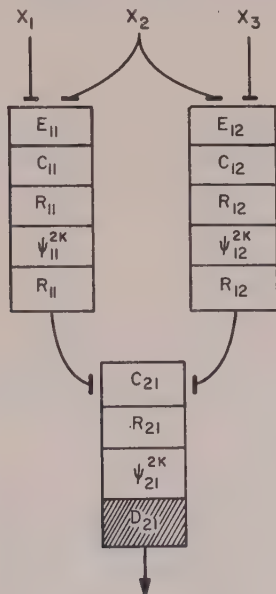


FIG. 19.  $f(x_1, x_2, x_3)$

By VNC, only  $D_{21}$  need be noiseless with respect to  $C_{ij}$ , and all other computing elements— $\psi_{ij}^{2\kappa}$ —may be noisy. This is clearly more satisfactory than EC, and the reasons for it are not obscure. For computations of unit logical depth VNC, as we have noted, achieves  $P_e = 2^{-\beta/R}$ , before the operation of decoding is performed,

by achieving the function shown on p. 145. It is obvious that this can be iterated, to any desired depth, and the final decoder is needed only to map  $n\xi \geq (1-\Delta)n$  into 1 and  $n\xi \leq n\Delta$  into 0. EC fails, because it has to reconstruct  $\kappa$ -sequences from all possible  $n$ -sequences, *at each level*, by noiseless devices. We may, of course, consider  $\psi_{ij}^{2\kappa}$  to be a species of decoder, since it certainly maps from all possible  $n$ -sequences into a restricted subset of these sequences, corresponding to  $\kappa$ -sequences, but *it does not have to be noiseless*.

Thus, in certain aspects, VNC is preferable to EC. However, VNC suffers from the drawback that  $n$  has to be of the order of  $10^3$ – $10^4$  before  $P_e$  becomes really small. Thus for  $n = 2 \cdot 10^4$ ,  $P_e = 2^{-3} \cdot 10^{-3} \cdot 2 \cdot 10^4 \doteq 10^{-18}$ . Furthermore, even in the absence of noise, degradation of information occurs, and has to be combated by the introduction of restoring organs. It would certainly be desirable to construct schemes that use much smaller numbers of components, and do not degrade in the absence of noise. In order to eliminate these aspects, we require to obtain deeper insights into the structure of VNC. For this purpose, a more abstract description is required, and we obtain this by using many-valued propositional calculi, instead of the conventional two-valued calculi. The relevant properties of these calculi are summarized in an appendix.

### MANY-VALUED LOGICAL SCHEMES

Let us consider once more, VNC, and suppose that we have just two lines per bundle:

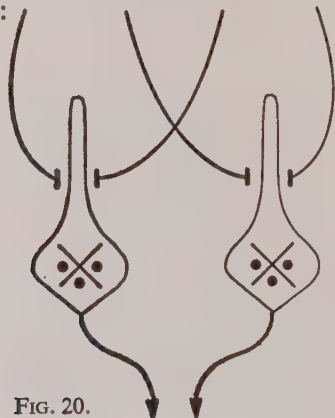


FIG. 20.

We interpret the excitation levels per bundle, not as 1's and 0's via some fiduciary level, but as a complete set of truth-values. Thus (11, 10, 01, 00) correspond to the truth-values (1, 2, 3, 4). Furthermore we recognize this whole net as the direct-product of two 2-valued systems, i.e.:

$$\begin{matrix} \text{X} \\ \text{X} \end{matrix} \otimes \begin{matrix} \text{X} \\ \text{X} \end{matrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^{(2)} = \begin{bmatrix} 00 & 01 & 10 & 11 \\ 01 & 01 & 11 & 11 \\ 10 & 11 & 10 & 11 \\ 11 & 11 & 11 & 11 \end{bmatrix}$$

and so we obtain, under the correspondence

$$(11, 10, 01, 00) \rightarrow (1, 2, 3, 4)$$

the 4-valued logical matrix:

$$\begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

which is a generalized Sheffer-stroke function. Similarly, we may construct generalized "and", "or" and "not" functions, using the nets:

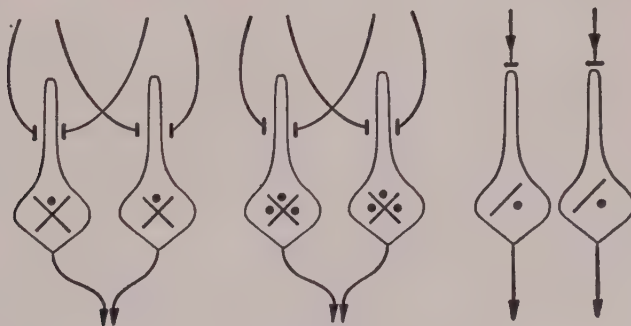


FIG. 21.

which realize the following logical matrices:

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 2 & 4 & 4 \\ 3 & 4 & 3 & 4 \\ 4 & 4 & 4 & 4 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix} \quad \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

But these are just the Lewis functions  $\&$ ,  $*$  and  $\Gamma$ , and in fact the generalized Sheffer function  $|$  is also a Lewis function, as may be seen by inspection of the tables, and conditions in the appendix.

In a trivially obvious fashion, it follows that all von Neumann's multiplexed automata (excluding restoring organs and randomizers) are generated from the generalized Sheffer function  $X|Y$ . Thus

$$\begin{aligned} X|Y &= \neg X * \neg Y, & X \& Y &= (X|X) | (Y|Y), \\ \neg X &= X|X, & X * Y &= \neg(\neg X \& \neg Y). \end{aligned}$$

In terms of the 2-valued Sheffer stroke function  $x|y$ , it is evident that

$$X|Y = (x_1|y_1) \otimes (x_2|y_2)$$

and in general,

$$X|_n Y = \bigotimes_{i=1}^n (x_i|y_i)$$

is a  $2^n$ -valued Sheffer stroke function. This function, or the doublets  $\{\&, \Gamma\}$  or  $\{*, \Gamma\}$ , does not generate all possible  $2^{2^\delta}$  Lewis functions of  $\delta$  variables, but only a subset of  $2^{2^\delta}$  functions. This subset is in fact the set of all possible Boole-Schröder functions of  $\delta$  variables, and is isomorphic to the set of all possible Boolean functions of  $\delta$  variables. It follows that *all executive ranks in VNC are Boole-Schröder functions, and also, of course, Lewis functions.* We note also that the same result follows if we make use of the generalized Pierce ampheck function

$$X\bar{Y} = \bigotimes_{i=1}^n (x_i \downarrow y_i) = \neg X \& \neg Y.$$

The full set of Lewis functions is realized only by the triplet  $\{\&, \Gamma, \diamond\}$ , the last element of which we have not yet considered. However, for  $n = 2, \psi^2$  the restoring organ of VNC, may be constructed as follows:

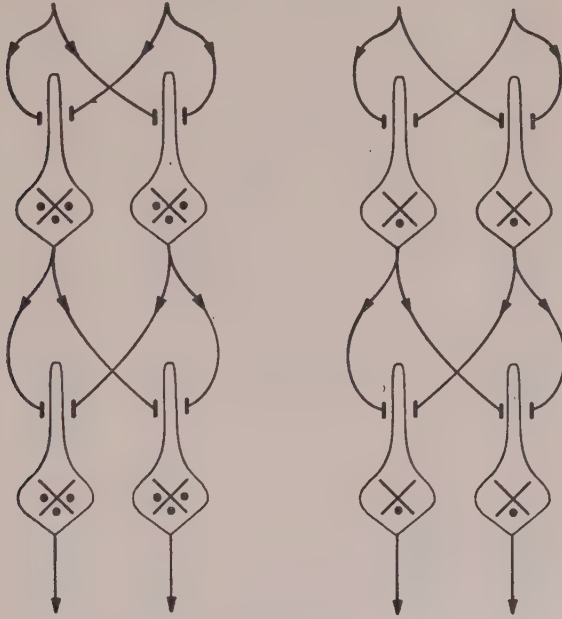


FIG. 22.

In terms of our many-valued logical systems, these realize the logical matrices

$$\begin{bmatrix} 1 \\ 4 \\ 4 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ 4 \end{bmatrix}$$

which are just the Lewis functions  $J_1(X) = \neg \diamond \neg X$  and  $\diamond X$  respectively. (We shall work henceforth with  $x \downarrow y$ , and we lose no

generality by so doing.) Similarly for  $n = 3$ , we have for  $\psi^2$ :

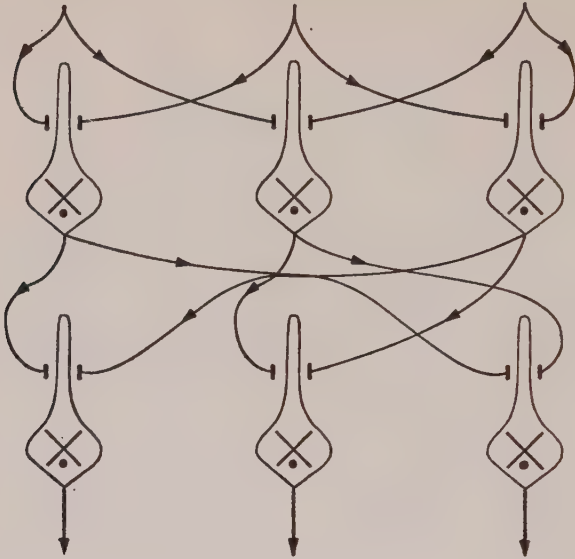


FIG. 23.

which realizes the logical function

1
1
1
7
1
4
6
8

under the correspondence

$$(111, 110, 101, 100, 011, 010, 001, 000) \rightarrow (1, 2, 3, 4, 5, 6, 7, 8).$$

Similar functions for  $n = 4$  and  $5$  may be constructed.



All of these functions may be shown to be Lewis functions, and we may characterize the operation of these functions as follows:

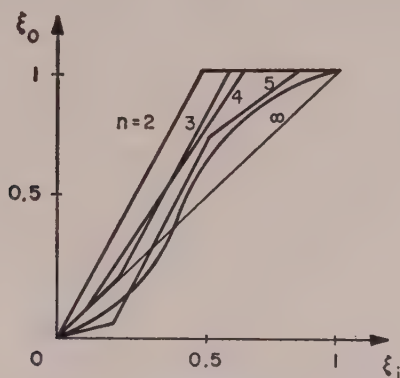


FIG. 24.

We note that these functions (for  $n = 2, 3, 4, 5, \dots$ ) are all isotone or *order-preserving*. In addition, as  $n$  increases, these functions become *order-continuous*.<sup>(19)</sup> It has been shown that this set of isotone order-continuous functions is a subset of the set of all possible Lewis functions.<sup>(27)</sup> It follows that *all restoring organ ranks in VNC are not Boole-Schröder functions, but are more general Lewis functions.*

Finally, we may characterize the randomizing operation on a bundle, by its property of preserving weight. This is clearly order-preserving, and is also order-continuous. Thus this operation is also a Lewis function, similar to restoring functions. *Hence all automata realized by VNC may be described by means of Lewis many-valued logical functions.*

We should now use this fact to investigate the reason for randomizing (which we have purposely avoided until now), and the nature of the degradation previously discussed.

Consider the generalized Sheffer and Pierce functions, for  $n = 2$ ; and let us choose a fiduciary level  $\Delta < \frac{1}{2}$ , so that  $(1 \rightarrow 1,$

$2 \rightarrow i, 3 \rightarrow i, 4 \rightarrow 0$ ).  $\Delta = \frac{1}{4}$ , will suffice. Thus we have

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 4 & 3 & 2 & 1 \\ 2 & 3 & 3 & 1 & 1 \\ 3 & 2 & 1 & 2 & 1 \\ 4 & 1 & 1 & 1 & 1 \end{array} \rightarrow \begin{array}{c|cccc} & 1 & i & i & 0 \\ \hline 1 & 0 & i & i & 1 \\ i & i & i & 1 & 1 \\ i & i & 1 & i & 1 \\ 0 & 1 & 1 & 1 & 1 \end{array}$$

Identifying the  $i$ 's, we obtain

$$\begin{array}{c|ccc} & 1 & i & 0 \\ \hline 1 & 0 & i & 1 \\ i & i & \boxed{\begin{array}{cc} i & 1 \\ 1 & i \end{array}} & 1 \\ 0 & 1 & 1 & 1 \end{array}$$

What can be done with the ambiguous  $i|i$ , which sometimes maps into 1, sometimes into  $i$ ? Let us take the mean value of this,  $(1+i)/2$ , so that we obtain

$$\begin{array}{c|ccc} & 1 & i & 0 \\ \hline 1 & 0 & i & 1 \\ i & i & \frac{1+i}{2} & 1 \\ 0 & 1 & 1 & 1 \end{array}$$

Let us now map back from  $\{1, i, 0\}$  into relative excitations, i.e.  $\{1 \rightarrow 1, i \rightarrow \frac{1}{2}, 0 \rightarrow 0\}$ , obtaining

$$\begin{array}{c|ccc} \zeta & 1 & \frac{1}{2} & 0 : \eta \\ \hline 1 & 0 & \frac{1}{2} & 1 \\ \xi : \frac{1}{2} & \frac{1}{2} & \frac{3}{4} & 1 \\ 0 & 1 & 1 & 1 \end{array}$$

but this is just  $\zeta = 1 - \xi\eta$ , which von Neumann obtained as the mean value of  $\zeta$ , when  $\xi$  and  $\eta$  were statistically independent random variables.

Similarly

1	4	4	4	4
2	4	3	4	3
3	4	4	2	2
4	4	3	2	1

 $\longrightarrow$ 

1	0	0	0	0
<i>i</i>	0	<i>i</i>	0	<i>i</i>
<i>i</i>	0	0	<i>i</i>	<i>i</i>
0	0	<i>i</i>	<i>i</i>	1

 $\longrightarrow$ 

1	0	0	0
<i>i</i>	0	<i>i/2</i>	<i>i</i>
0	0	<i>i</i>	1

leading to

$\zeta$	1	$\frac{1}{2}$	0	$:\eta$
1	0	0	0	
$\xi: \frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{2}$	
0	0	$\frac{1}{2}$	1	

which is just  $\zeta = (1-\xi)(1-\eta)$ . Had von Neumann performed a statistical analysis with this organ, he would have obtained a mean value  $\bar{\zeta} = (1-\xi)(1-\eta)$ , corresponding to the above result.

Generalized even-weight functions can also be treated in this manner. For example:

•	•
---	---

<sup>(2)</sup>
 $\equiv$ 

1	4	3	2	1
2	3	4	1	2
3	2	1	4	3
4	1	2	3	4

 $\longrightarrow$ 

1	0	<i>i</i>	<i>i</i>	1
<i>i</i>	<i>i</i>	0	1	<i>i</i>
<i>i</i>	<i>i</i>	1	0	<i>i</i>
0	1	<i>i</i>	<i>i</i>	0

leading to

1	0	<i>i</i>	1	$\zeta$	1	$\frac{1}{2}$	0	$:\eta$
<i>i</i>	<i>i</i>	$\frac{1}{2}$	<i>i</i>	$\xi:$	1	0	$\frac{1}{2}$	1
0	1	<i>i</i>	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
				0	1	$\frac{1}{2}$	0	

that is,

$$\zeta = (1-\xi)\eta + (1-\eta)\xi = \xi + \eta - 2\xi\eta.$$

Clearly, for large  $n$ , these same formulae will hold; and we see the nature of von Neumann's approximation. Given independent

random variables for  $\xi$  and  $\eta$ , the most probable value, and best approximation to  $f(i, i)$  is certainly the mean value of  $\zeta$ . The odd-weight functions, however,  $f(i, i)$  sometimes maps into 1 or 0, or vice versa. For even-weight functions  $f(i, i)$  always maps into 1 or 0, and  $f(1, 0)$  into 1 or 0. Thus in some cases, odd-weight functions degrade:  $1 \cup 0 \rightarrow i$ , whereas even-weight functions always restore. Hence we do not need any additional restoring organ for even-weight functions, but we do need such organs for odd-weight functions. We may consider these restoring organs not only as decoders of the output of one rank, but as encoders for the next rank.

In this last analysis we have been concerned only with malfunctions caused by that degradation of information which occurs in the absence of computing noise. That is, we considered only *coding malfunctions*, and ignored *computing malfunctions*. When the latter occur, since restoring organs operate on malfunctions independently of their origin, it follows that they are processed exactly as are the former. Thus all malfunctions are restored by these organs, which we conceive of as combined encoders and decoders.

#### TOWARD MORE EFFICIENT SCHEMES

A number of questions regarding odd-weight functions can now be raised: (a) wherein exactly does coding malfunction lie? (b) Is it possible to eliminate coding malfunctions? (c) Can restoration of malfunctions be made more efficient?

The answer to (a) is easily found, via our many-valued logical picture. From what has been discussed, it is clear that intermediate stages of VNC minus restoring organs, etc., can be represented as follows:

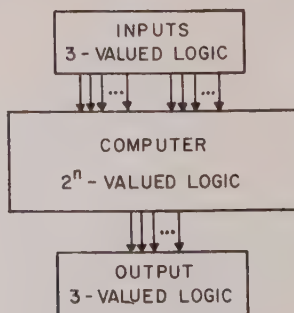


FIG. 25.

The desired 3-valued input-output mapping (for a generalized Sheffer-stroke system) is

	1	<i>i</i>	0
1	0	<i>i</i>	1
<i>i</i>	<i>i</i>	<i>i</i>	1
0	1	1	1

but as we have seen, because of the structure of the computer, all we can obtain are Boole-Schröder or Lewis functions, e.g.

<sub>g</sub>	1	2	3	4
1	4	3	2	1
2	3	3	1	1
3	2	1	2	1
4	1	1	1	1

which is always incompatible with the desired 3-valued function. This is true for all generalized  $2^n$ -valued Boole-Schröder functions. Thus, in general, *the coding logic is incompatible with the computing logic.*

An answer to (b) is now indicated. We must either change the computing logic to fit the coding logic, or the coding logic to fit the computing logic or change both, and devise a completely new scheme. Let us consider the first of these alternatives. It is clear that if we order  $\{1, i, 0\}$  in linear fashion, then we can express the desired 3-valued input-output function as

$$f(x, y) = \min(x^c, y^c)$$

where  $c$  denotes the following operation:

	$c$
1	0
<i>i</i>	<i>i</i>
0	1

But this is just the *Post-Lukasiewicz many-valued negation* (see Appendix) and

$$\min(x^c, y^c) = x|_P y = x^c \cdot y^c$$

where  $|_P$  is the Post-Lukasiewicz generalization of the Sheffer-stroke function, and  $\cdot$  is the Post-Lukasiewicz disjunction. Evidently if we replace the Boole-Schröder functions by the Post-Lukasiewicz functions:

	1	2	3
1	3	2	1
2	$i$	2	1
3	1	1	1

	1	2	3	4
1	4	3	2	1
2	3	3	2	1
3	2	2	2	1
4	1	1	1	1

	1	2	3	4	5
1	5	4	3	2	1
2	4	4	3	2	1
3	3	3	3	2	1
4	2	2	2	2	1
5	1	1	1	1	1

then we can obtain the required compatibility of computing and coding logics.

Two realizations of these schemes are possible, one involving bundling, the other multistate components. The latter realization is obtained by replacing the binary scheme of Fig. 7 by the following  $n$ -ary scheme:

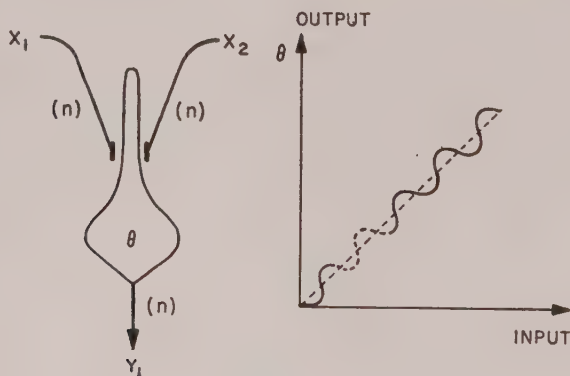


FIG. 26.

Encoding is simply a matter of *amplification*, error restoration is the same as before, and in fact there are more possibilities of performing this than hitherto.<sup>(20)</sup> However, only when the output amplitude range increases with  $n$ , do we obtain arbitrarily low



error probabilities.\* Since we are dealing with essentially band-limited components we shall consider henceforth, only schemes with small fixed amplitude ranges producing only two states.

To realize Post-Lukasiewicz functions with these schemes, we require to note the nature of the constraints present in the various many-valued logics. Consider the following scheme:

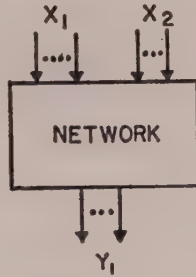


FIG. 27.

Suppose there are  $n$  lines per bundle, so that there are  $2^{2n}$  possible input configurations, and  $2^n$  possible outputs. In general, then, there are  $(2^n)^{2^{(2n)}}$  possible mappings, provided no constraints exist on the structure of the network. Letting  $m = 2^n$ , we have  $m^{m^2}$  possible functions, which is just the number of Post functions in 2 variables. We may represent this as follows, letting  $n = 2$ :

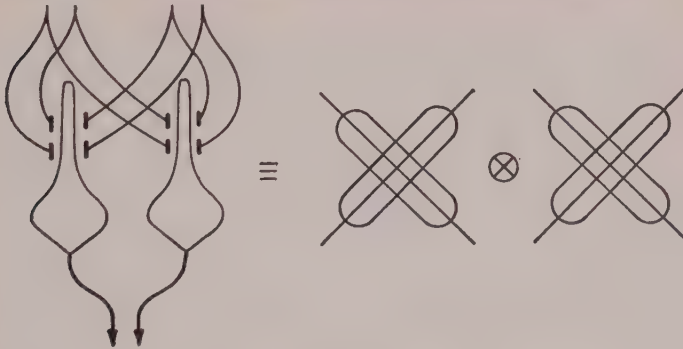
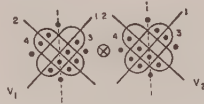


FIG. 28.

\* This will be noted in another paper<sup>(21)</sup> and is a special case of the signal to noise ratio vs. bandwidth analysis of B. M. Oliver, J. R. Pierce and C. E. Shannon.<sup>(22)</sup>

Provided each formal neuron computes independently of the other, we obtain the required Post functions.

To realize Post-Lukasiewicz functions, of which there are  $2^{(2^n)} \cdot m^{(m^n - 2^n)}$  we have only to introduce the constraint  $f(1, m) \in 1 \cup m$ . We do this by insisting on *vertical isomorphism* of all Venns. Thus we realize the Post-Lukasiewicz Sheffer-stroke function (*loc. cit.*) as follows:



and we note that whenever a jot appears along the vertical axis of  $V_1$ , it appears in  $V_2$ , and vice versa.

The other many-valued logics considered, Lewis and Boole-Schröder may be realized in similar fashion, by putting in additional symmetry constraints *in each Venn*. Thus we realize the Lewis or Boole-Schröder Sheffer-stroke function (*loc. cit.*) as follows:



where symmetry in  $V_1$  with respect to the (1, 3) axes becomes symmetry  $V_2$  in with respect to the (2, 4) axes.

Obviously this reduces to



which is the original VNC.

Finally, we obtain 2-valued Boolean logic, by demanding identity of all



There are in fact various other ways of obtaining other than Boole-Schröder functions, e.g.:

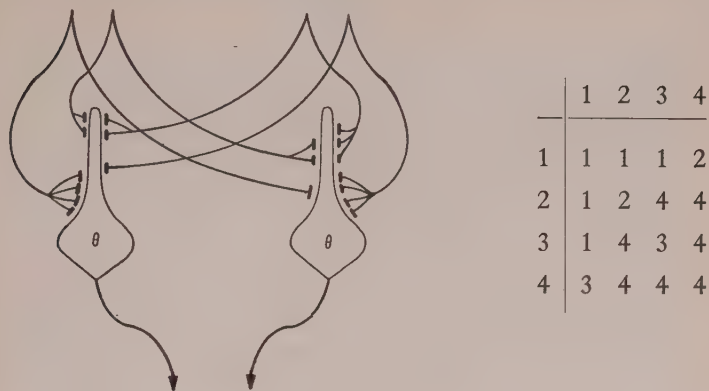


FIG. 29.

which realizes a general Post function, but we shall not go into these here.

At any rate, we see how to realize Post-Lukasiewicz functions, and so eliminate coding malfunctions, by using formal neurons, each with  $2n$  possible inputs, instead of 2 or 3 as in VNC, and allowing these neurons to compute different functions, even in the same rank.

The many-input element has in fact been used in a different context by Allanson<sup>(9)</sup> who argued that von Neumann's location of error as output error was incorrect, as far as neurophysiology goes, and that input errors were far more frequent, and important. Thus he considered the problem of functions of probabilistic arguments, similarly to Moore and Shannon.<sup>(7)</sup> Noting that neurons usually have many afferents synapsing on them, instead of 2 or 3, he investigated many-input neurons, with errors at synapses, and obtained the results that increasing the number of afferents per neuron, in general, reduces the effects of errors in synaptic transmission. (This solution is in fact qualitatively similar to the Moore-Shannon solution, since a many-input neuron can be considered as the equivalent of a redundant relay net.) It turns out that such formal neurons are useful as far as the output noise problem is concerned. McCulloch, Blum, Verbeek<sup>(13-15)</sup> and

Cowan, have shown that a much more effective error reducing scheme is obtained by use of the following construction:

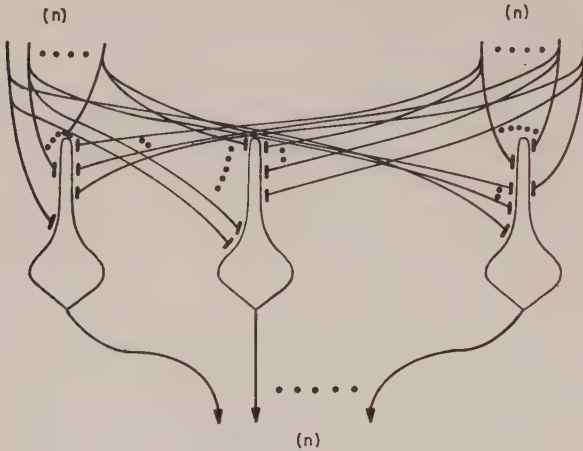


FIG. 30.

in which each line is connected to all  $n$  neurons, so that each neuron has  $2n$ -inputs, and in which each neuron computes the same function. For this scheme, a probability of malfunction  $P_e \sim 2^{-\gamma n}$ ,  $\gamma = \frac{2}{3}$  is obtained, for  $\Delta \sim \frac{1}{3}$  and  $\epsilon = 0.05$ , and input error of probability  $\eta = 0.05$ ; which compares very favorably with the  $P_e = 2^{-\beta n}$ ,  $\beta = 3.5 \times 10^{-3}$  for an  $\epsilon = 0.005$ ,  $\eta = 0$  obtained by VNC. Thus this scheme not only reduces output malfunctions, but also does the same for input malfunctions. In terms of  $2n$ -input components, compared with 2-input components, the comparative information processing rates are as  $\gamma/\beta \doteq 2 \cdot 10^2$ .\*

\* We note that this is approximately the same ratio to VNC, as is the Moore-Shannon construction for relays (*loc. cit.*). It might be argued that to realize a  $2n$ -input neuron, say for  $n = 40$ , to obtain  $P_e \sim 10^{-7}$ , involves about 80 or so relays, so that judged in terms of 2-input components, the rates are of the same order of magnitude; and moreover, a certain measure of redundancy would be required to obtain an  $\epsilon = 0.05$  for this device, given an  $\epsilon = 0.005$  for the 2-inputters. This is a good argument for *not* using relays or any comparable 2-input device as a basic building block, and for using  $2n$ -input neurons, with all their nice properties, as building blocks for reliable automata. Technology, however, has not yet produced such a device, only Nature.

We shall not be concerned with the exact details of this work, and refer the reader to Verbeek (*loc. cit.*). What we wish to comment on, is the logical structure of these schemes.

By virtue of the "all-to-all" connection, in the absence of computing noise, the output is 2-valued, consisting of all 1's or all 0's. So the structure is that of a Boolean logic,<sup>†</sup> and this scheme, like the even-weight functions, restores all input malfunction during the process of computation. For example:



realizes the function

	1	2	3	4
1	4	4	4	1
2	4	1	1	1
3	4	1	1	1
4	1	1	1	1

so that, if  $(2, 3) \equiv i$  then  $i \rightarrow 1$ .

We can thus eliminate coding malfunctions using all-to-all schemes. These differ from the Post-Lukasiewicz schemes in that the latter merely eliminate degradation, while the former, not only eliminate this, but restore *all* malfunctions at the same time. Thus the Post-Lukasiewicz schemes provide an answer to (b), but the all-to-all schemes, not only answer (b), they provide also an answer to (c).

For computations of arbitrary logical depth, VNC introduced restoring organs to control error propagation. It turns out that similar restoring organs, with all-to-all connections however, are much more efficient than those of VNC. For majority organs (or

---

<sup>†</sup> We must distinguish between this scheme, which is a Boolean logic, *only in the absence of noise*, and a Post logic in the presence of noise; and the *nonredundant* schemes of Fig. 7, which are Boolean both in the absence and presence of noise.

polypchecks), Verbeek has obtained the following results:

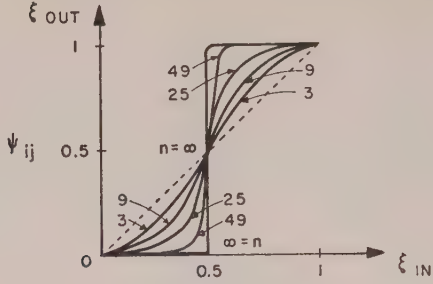


FIG. 31.

Thus for  $n \sim 40$ , almost everything is restored immediately after only one application. So with this  $n$ , we can obtain an error probability of  $P_e \sim 10^{-7}$  for nets of arbitrary logical depth, up to the last rank of restoring organs, and then we are bound only by the noise in the last rank, as before. Thus we obtain a much more efficient error reducing scheme than VNC, and we have answered question (c). It is clear that we obtain such improvement, because of the fact that we make full use of the information at each rank, and process it maximally. Moreover, because of this, no randomizing is necessary, and we obtain a much simpler structure.

Compared with Fig. 19, we have the following:

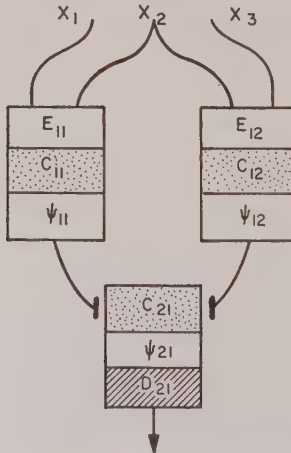


FIG. 32.



where  $E_{ij}$  is line-splitting as before,  $\psi_{ij}$  are restoring organs with all-to-all connections, and the  $C_{ij}$  may be all-to-all with each computing the same function, or all-to-all with each computing a different function. In fact this latter distinction is academic, since the presence of  $\psi_{ij}$  makes it immaterial, and all malfunctions will be efficiently corrected.

It appears possible, however, to obtain an even more elegant system, by noting that  $C_{ij}$  effectively reduces the effects *both of input and output errors*, and thus it may be possible to dispense with restoring organs altogether.

Thus  $P_e \sim 2^{-\gamma n}$  is obtained for  $\eta = \epsilon = 0.05$ , leading to the conclusion that for sufficiently large  $n$ , errors will not propagate through the net, even without restoring organs. In a sense, the all-to-all nets may be said to be *self-restoring*. In this case, Boolean functions must be used, rather than all-to-all Post-Lukasiewicz functions, to eliminate coding malfunction, which would propagate, so that the following scheme:

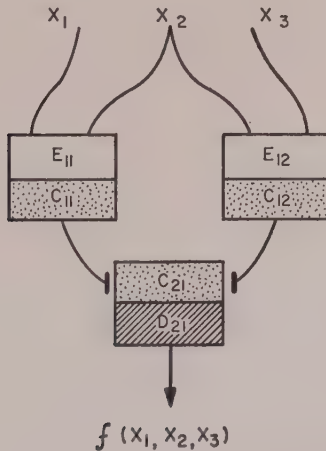


FIG. 33.

in which  $E_{ij}$  is again line-splitting,  $C_{ij}$  are all-to-all Boolean functions, and  $D_{21}$  is a relatively infallible decoder; malfunctions with

a probability of error  $P_e \sim 2^{-\gamma n}$ . We note that all the complexity now resides in the  $C_{ij}$ , which in a sense are encoders and decoders, as well as computing devices, since they are self-restoring.

### CONCLUSIONS

We have now completed a rather devious path from and back to our original discussion of information-rates and error-probabilities, without even really making direct contact with it. The problem of obtaining non-zero rates for reliable computation (especially for schemes of arbitrary logical depth) is still unsolved, since even in the later schemes we considered, the probability of malfunction clearly goes to zero with the rate. What we did was to obtain lower probabilities of malfunction, using fewer but sloppier components. Two points are worth noting. Firstly, for computations of arbitrary logical depth, it is not very meaningful to permit the insertion of noiseless decoders after each computation, and some kind of approach, other than the conventional information-theoretic one appears necessary. In this respect, the all-to-all Boolean scheme discussed, appears promising, in that each rank of computing elements essentially combines the encoding and decoding functions into one, with the computing function. This scheme is such that all the previous complexity of structure-restoring organs, randomizers—has been eliminated, at the cost of more complexity of individual computing ranks. Since errors are independent of function, this is not troublesome.

Secondly, the question of coding logic has not been properly investigated. We have stuck to the “bundling” schemes of EC and VNC, which are clearly limited to  $R \leq 2/n$ . It may be that somewhat different coding logics, together with more general Post functions—along the lines of Post–Lukasiewicz schemes—will be more fruitful, and we are at present investigating this. In any case, we have shown that redundant automata can be described by various systems of many-valued logics, in which not all truth-values are informationally significant, and that these logics have varying structure, corresponding to different methods of coding and processing information. We consider that the search for ways of achieving reliable computations at non-zero rates in the presence of noise is facilitated by the use of these logics.

APPENDIX—MANY-VALUED LOGICS

Many-valued logics (or non-Aristotelian logics, as they are sometimes called) were first constructed by Post<sup>(23)</sup> and independently by Lukasiewicz.<sup>(24)</sup> Post's scheme was a straightforward generalization of Boolean logic. Compared with Boolean functions, which take values in the set  $\{1, 0\}$ , or equivalently  $\{1, 2\}$ , Post functions take values in the set  $\{1, (m-1/m), \dots, 1/m, 0\}$  or equivalently  $\{1, 2, \dots, m-1, m\}$ . At most, there are  $m^{m^n}$  Post functions of  $n$  variables, compared with the  $2^{2^n}$  possible  $n$ -variable Boolean functions. Many Post functions are natural generalizations of Boolean functions. For example, the Boolean disjunction and conjunction can be represented in truth-value matrix form by

•	1	2
1	1	2
2	2	2

∨	1	2
1	1	1
2	1	2

respectively while the corresponding Post functions are represented by the matrices:

•	1	2	3	4
1	1	2	3	4
2	2	2	3	4
3	3	3	3	4
4	4	4	4	4

∨	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	2	3	3
4	1	2	3	4

Both Boolean and Post disjunction and conjunction can be represented by the truth-value functions:

$$x \cdot y = \text{df } \max(x, y)$$

$$x \vee y = \text{df } \min(x, y).$$

Similarly, Boolean and Post negation can be represented by the truth-value function:

$$\sim x = \text{df } (m+1-x).$$

We note, however, that in an  $m$ -valued Post-logic, there are  $m^m$  unitary functions, compared with  $2^2$  unitary Boolean functions.

The most useful of these Post functions is the set of functions  $J_i(x_\kappa)$ , ( $i = 1, 2, \dots, m$ ) defined such that

$$J_i(x_\kappa) = \begin{cases} 1, & x_\kappa = i \\ m, & x_\kappa \neq i \end{cases}$$

There also exists an analog of the fundamental theorem of Boolean logic, which states that every Post-function can be expressed as the conjunction of disjunctions of  $J$ -functions, i.e.:

$$f(x_1, \dots, x_n) = \bigvee_{i_1=1}^m \bigvee_{i_2=1}^m \dots \bigvee_{i_n=1}^m f(i_1, i_2, \dots, i_n).$$

$$J_{i_1}(x_1) \cdot J_{i_2}(x_2) \cdot \dots \cdot J_{i_n}(x_n).$$

Since there are  $m^{m^n}$  Post functions of  $n$ -variables, this implies that every truth-value matrix is a Post function of two-variables, and so on. For this reason, Post logic is said to be *functionally complete*.

The many-valued logics of Lukasiewicz have the same disjunctions, conjunctions and  $J$ -functions as Post logic, but are *functionally incomplete*. That is, not every Post function is a Lukasiewicz function. Lukasiewicz functions, in fact, satisfy the constraint that

$$f(1, m) \in \{1, m\},$$

i.e. the truth-value set  $A = \{1, m\}$  under any composition law corresponding to a Lukasiewicz function, is *closed*. Because of this constraint, there are only  $2^{2^n} \cdot m^{m^n - 2^n}$  possible Lukasiewicz functions of  $n$ -variables.

A further form of many-valued logic was constructed by Lewis.<sup>(25)</sup> Lewis logic is essentially a *Boolean product logic*, whose structure is isomorphic to that of a Boolean algebra of order  $2^\delta$ .<sup>(26)</sup> Thus a  $2^\delta$ -value Lewis logic in  $n$  variables, has a structure isomorphic to that of a Boolean logic in  $\delta n$  variables. There are thus  $2^{2^{\delta n}} = 2^{m^n}$  possible  $m$ -valued Lewis functions in  $n$  variables. Lewis logic is functionally incomplete, i.e.  $f(1, m) \in \{1, m\}$ . In addition, it possesses other *symmetry conditions*, due to its aforementioned structure. Thus for  $m = 4$ ,  $f(2) = f(3)$  and for  $m = 8$ ,  $f(2) = f(3) = f(5)$ ;  $f(4) = f(6) = f(7)$ . The exact nature of these symmetry conditions will be made clear in a later section.

Examples of Lewis functions  $m = 4$ , are as follows:

$\&$	1	2	3	4	$\vee$	1	2	3	4
1	1	2	3	4	1	1	1	1	1
2	2	2	4	4	2	1	2	1	2
3	3	4	3	4	3	1	1	3	3
4	4	4	4	4	4	1	2	3	4

The differences between these functions, and Post-Lukasiewicz functions have been marked.

Important unitary functions are the Lewis  $J$ -functions defined by the following truth-tables:

$x$	$J_1$	$J_2$	$J_3$	$J_4$
1	1	4	4	4
2	4	2	3	4
3	4	3	2	4
4	4	4	4	1

and also the Lewis modal function "Possibly  $x$ " ( $\diamond x$ ):

$x$	$\diamond x$
1	1
2	1
3	1
4	4

i.e.

$$\diamond x = \text{df} \begin{cases} 1, & x \neq 4, \\ 4, & x = 4 \end{cases} = 1$$

The fundamental theorem for Lewis logic is similar to that for Post, with the additional constraints imposed on  $f(i_1, i_2, \dots, i_n)$ . Alternatively, any Lewis function can be repressed by some combination of the triple ( $\&$ ,  $\vee$ ,  $\diamond$ ). A more extensive discussion of these logics can be found in 25, 26 and 27.

A lattice-theoretic characterization of many-valued logics can be made.<sup>(28,29)</sup> All the logics mentioned so far can be represented as complemented, distributive, modal lattices (cf. Birkhoff<sup>(28)</sup>). Thus the 27 unitary functions of 3-valued Post logic can be characterized by the following lattice:

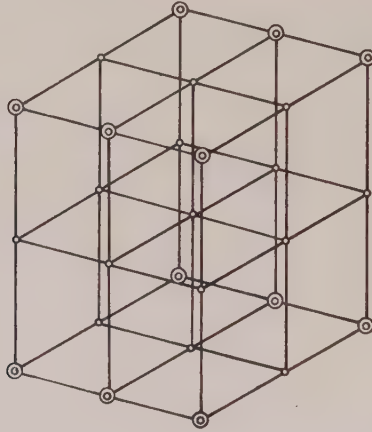


FIG. 34.

The 12-unitary functions of Lukasiewicz logic are also shown in this lattice, or as follows:

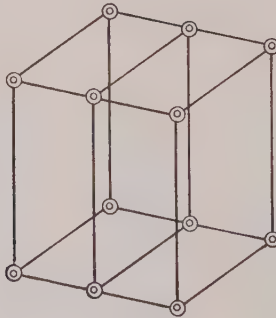


FIG. 35.

Clearly this lattice is also a sublattice of the Post lattice.



The lattice of  $2^2$ -valued unitary functions, 16 in all, can be represented as follows:

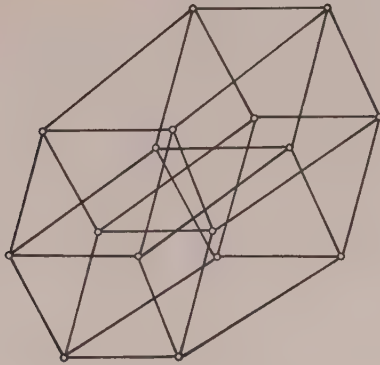


FIG. 36.

which is also a representation of the Boolean lattice of 2 variables. This lattice is a sublattice of the lattice of 64-Lukasiewicz 4-valued unitary functions, which is itself a sublattice of the 256-element Post lattice of 4-valued unitary functions.

Another lattice theoretic characterization of these many-valued logics can be given (which is most useful for our purposes). Functions are now regarded as mappings for a lattice of variables onto or into a *lattice of truth-values*.<sup>(27)</sup> Thus our unitary functions are represented by the following schemes:



Post-Lukasiewicz functions

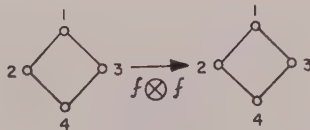


FIG. 37.

Lewis functions.

The extra symmetry conditions possessed by Lewis logic are now evident. The lattice of truth-values in Lewis logic forms a partly ordered set,<sup>(28)</sup> and of course a lattice. The lattices of truth-values in Post-Lukasiewicz logic are all *chains*, and lack these symmetry conditions.

Similarly, the many-valued binary functions can be represented as follows:

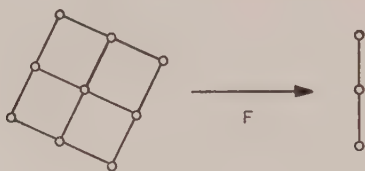


FIG. 38. Post-Lukasiewicz functions.

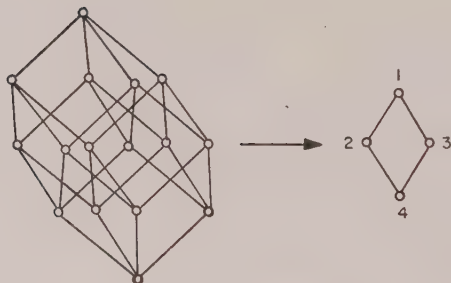


FIG. 39. 4-valued Lewis functions.

Finally, although the Post-Lukasiewicz schemes of truth-values grow in monotonic fashion with  $m$ :

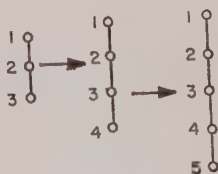


FIG. 40.

the Lewis schemes grow as follows:

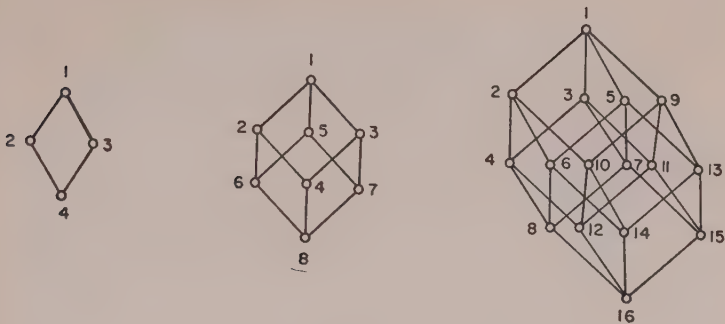


FIG. 41.

ACKNOWLEDGMENTS

This paper is based on a thesis submitted in partial fulfilment of the requirements for the S.M. Degree, Department of Electrical Engineering, M.I.T., June (1960). I should like to acknowledge the substantial help and support of my thesis supervisor Professor Peter Elias, and of my colleagues Dr. Warren S. McCulloch, Dr. L. A. M. Verbeek and Mr. Manuel Blum of the Massachusetts Institute of Technology. In particular Dr. McCulloch's encouragement and aid has been invaluable.

Financial support was provided by a Fellowship from International Computers and Tabulators Ltd., London, England and by the Research Laboratory of Electronics M.I.T., whose work is supported in part by the U.S. Army Signal Corps, the Air Force Office of Scientific Research and the Office of Naval Research.

REFERENCES

1. C. E. SHANNON and W. WEAVER, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois (1948).
2. D. M. MACKAY, The place of meaning in the theory of information. *3rd London Symposium of Information Theory* (ed. E. C. CHERRY) (1956).
3. D. M. MACKAY, The informational analysis of questions and commands. *4th Symposium on Information Theory* (ed. E. C. CHERRY), London (1961).
4. M. EDEN, A note on error detection in noisy logical computers. *J. Inf. and Control* **2** p. 310 (1959).

5. S. W. McCULLOCH, The reliability of biological systems. *Self-organizing Systems* (ed. YOVITTS and CAMERON), Pergamon, London (1958).
6. W. S. McCULLOCH. *The Stability of Biological Systems. Homeostatic Mechanisms*, Brookhaven Symposia in Biology, No. 10) Off. Tech. Serv. U.S. Dept. of Comm. Washington (1957).
7. E. MOORE and C. E. SHANNON, Reliable circuits using less reliable relays. *J. Franklin Inst.* **262**, pp. 191-208 (1956); **262**, pp. 281-97 (1956).
8. M. KOCHEN, Extension of Moore-Shannon model for relay circuits. *IBM J. Res. Devel.* **3**, p. 2 (1959).
9. J. ALLANSON, The reliability of neurons, *Proc. First International Congress on Cybernetics*, Namur (1956).
10. J. VON NEUMANN, Probabilistic logics, and the synthesis of reliable organisms from unreliable components. *Automata Studies* (eds. C. E. SHANNON and J. MCCARTHY) Princeton (1956).
11. P. ELIAS, Computation in the presence of noise. *IBM J. Res. Devel.* **2**, p. 4 (1958).
12. W. PETERSEN, On codes for checking logical operations. *IBM J. Res. Devel.* **3**, p. 2 (1959).
13. W. S. McCULLOCH, Symbolical representation of the neuron. This vol., p. 91.
14. M. BLUM, Properties of a neuron with many inputs. *Bionics Symposium* (WRIGHT-PATTERSON AIR FORCE BASE). Off. Tech. Serv., U.S. Dept. of Comm. Washington (1960).
15. L. A. M. VERBEEK, Reliable computation with unreliable circuitry. *Bionics Symposium* (WRIGHT-PATTERSON AIR FORCE BASE). Off. Tech. Serv., U.S. Dept. of Comm. Washington (1960).
16. R. HAMMING, Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, pp. 147-60 (1950).
17. P. ELIAS, Cooling for noisy channels. *IRE Convention Record*, Pt. 4, pp. 37-44 (1955).
18. L. LÖFGREN, Automata of high complexity and methods of increasing their reliability by redundancy. *Inf. and Control*, **1**, p. 127 (1958).
19. E. J. McSHANE, *Order-preserving Maps and Integration Processes* (Ann. Math. Studies No. 31) Princeton (1942).
20. O. LOWENSCHUSS, Restoring organs in redundant automata. *Inf. and Control*, **2**, p. 113 (1959).
21. J. D. COWAN, Towards a proper logic for parallel computation in the presence of noise. *Bionics Symposium* (WRIGHT-PATTERSON AIR FORCE BASE) Off. Tech. Serv., U.S. Dept. of Comm. Washington (1960).
22. B. M. OLIVER, J. R. PIERCE, and C. E. SHANNON, The philosophy of PCM. *Proc. IRE*, **38**, p. 3577 (1950).
23. E. L. POST, Introduction to a general theory of elementary propositions. *Amer. J. Math.* **43**, pp. 163-85 (1921).
24. J. LUKASIEWICZ, Philosophische Bemerkungen zu mahrwertigen Systemen des Aussagenkalkues. *Compt. rend. Soc. Sci. et Lettres Varsovie* **23**, Classe III, pp. 51-77 (1930).
25. C. I. LEWIS and C. E. LANGFORD, *Symbolic Logic*, Century, New York and London (1932); reprint, Dover Publications, New York (1959).
26. S. I. KISA, *Transformations on Lattices and Structures of Logic*, New York (1947).
27. J. D. COWAN, Many-valued logics and problem solving mechanisms. D.I.C. Thesis, Imperial College, London (1959).

28. G. BIRKHOFF, *Lattice Theory*, American Mathematical Society Colloquium Publications, Vol. 25, American Mathematical Society, New York, rev. ed. (1948).
29. A. ROSE, Systems of logic whose truth-values form lattices. *Math. Ann.* **123**, pp. 152-65 (1951).





## LARS LOFGREN

*University of Illinois, Urbana, Illinois*

# SELF-REPAIR AS THE LIMIT FOR AUTOMATIC ERROR CORRECTION\*

## 1. INTRODUCTION

Let us consider for a moment the interaction between man and nature in general terms. We observe that many of our activities could be done as well by machines. We start to construct such machines, i.e. we try to direct some events in nature according to our wishes. It is not possible, however, for us to isolate these events completely from other events in nature, and we find that our machines do not do precisely what we want them to do. We then speak about errors in the machines. For short intervals of time, however, their probability of correct behavior may be satisfactorily large. What we really want is to have the machines do as much as possible of our work, that is, we want to isolate ourselves as much as possible from our machines, except for the interaction that corresponds to our use of them. For instance, we want to diminish the time we spend repairing the machines.

Furthermore, we want to diminish our interaction with the machines during the period of their construction. It is obvious that our wish for self-repairing, self-constructing machines can never be fulfilled, when we use the prefix "self" in the strict sense that the machines are completely isolated from us except for our desired use of them. That is, if we except the services that nature today offers to mankind. We can speak with a certain confidence, for example, of the fruit of a wild apple tree as self-constructing food. The services we want done are, however, many more. So there must be an end somewhere to these wishes, and we shall find

---

\* This work was supported by the U.S. Office of Naval Research Contract Nonr 1834(21).

that the concept of self-repair is not meaningful, except when qualified.

Only within classical physics can we speak in principle of some deterministic events as being completely isolated from other events. We can also speak in principle of experimental arrangements which follow strictly some physical laws, i.e. the outcome, or output, is strictly determined by causes, the inputs. We can therefore speak in principle of an error-free behavior of the arrangement, the machine.

According to quantum physics, however, such an isolation is not possible even in principle. Hence an error-free self-repairing machine must be an idealization, except for the trivial cases where the normal behavior of the machine is unspecified, or where it is specified to be what it will be, again with no room for errors.

An error is a deviation from "truth". But "truth" is in general not known or knowable. It has to be hypothesized. We will make the hypothesis that "truth" is deterministic behavior (the behavior of a deterministic system). Deterministic systems will be defined later in Section 2.

Quantities related by a deterministic system must, as a matter of principle, be thought of as quantized variables, i.e. variables defined only for certain stable states,  $s_i$ , separated by unstable transition regions,  $u_j$ . This statement is a consequence of the fact that a physical quantity, regarded as a continuous variable, cannot be determined with infinite accuracy within a finite time, without disturbing the quantity. For when we measure a quantity  $Q$ , assumed to be a continuous variable, we must specify the accuracy of the measurement; for instance, if the measurement gives the value  $q$ , we may hypothesize that  $Q$  had a value in the closed interval  $[q - \delta_1, q + \delta_2]$ . If a measurement of this sort is to be regarded as deterministic, we must be able to give a precise answer to the question: if two measurements of  $Q$  give the values  $q'$  and  $q''$ , were then the two corresponding  $Q$ -values equal or unequal? But the only precise answer we can give, if  $Q$  is regarded as a continuous variable, is that if  $|q' - q''| > (\delta_1 + \delta_2)$ , then the two  $Q$ -values were unequal. But if  $|q' - q''| \leq (\delta_1 + \delta_2)$ , then we cannot make any decisions. We can, however, talk about the probabilities of each of the two answers, but this is not in agreement with a deterministic view. So in principle we must forbid

the ambiguity in the situation  $|q' - q''| \leq (\delta_1 + \delta_2)$ , which means that we must consider  $Q$  quantized into certain stable states,  $s_i$ , separated by unstable transition regions,  $u_i$ , each of a width larger than  $(\delta_1 + \delta_2)$ . Only in this case we can say with certainty that if  $|q' - q''| \leq (\delta_1 + \delta_2)$ , then the two corresponding  $Q$ -values are equal, in the sense that they belong to the same stable state. We are in principle not allowed to take conclusions from measurements during a transition period, and we can in principle identify the transition periods by their stipulated unstable character. The stable states may be defined over intervals:

$$s_i = \{q | q_i' \leq q \leq q_i''\} \quad (1)$$

where

$$q_i' \leq q_i'', \quad q_{i-1}' < q_i', \quad q_{i-1}'' < q_i'', \quad q_i' - q_{i-1}'' > (\delta_1 + \delta_2) > 0.$$

A general description of a (quantized) finite, deterministic event can be formalized as a mapping:

$$\begin{aligned} F : A &\rightarrow B \\ A &= S_1 \times S_2 \times \dots \times S_k \\ B &= S_i \times S_{i+1} \times \dots \times S_{i+n-1}. \end{aligned} \quad (2)$$

The domain  $A$  and the range  $B$  of the mapping  $F$  are formed as cartesian products of the sets  $S_j$ . These sets are defined with elements  $s_i$  of type (1). To each specified mapping  $F$ , we can synthesize a switching net of  $k$  inputs and  $n$  outputs with a deterministic behavior defined by the mapping. A *switch* or a *switching net* is a device whose output states are determined by its input states. The behavior of a *sequential circuit*, on the other hand, is described in terms of input and output time sequences. These sequences imply an ordering with respect to time. Each finite deterministic behavior of a sequential circuit can be simulated by a switching net, so that the time-ordering is mapped onto an input-ordering. There are only two different structures for uniform switching nets, the *g-structure* and the *c-structure*.<sup>(10)</sup> The states of a *c-net* are two- or three-valued, i.e. the index  $i$  of (1) can have only two or three values. There is no such restriction for *g-nets*.

From the definition (i.e. an implication in both directions) of a deterministic system given in Section 2, we obtain, by denial, classification of errors (Section 3). Assumptions are made

concerning the statistical nature of the errors. We will treat both temporary and stationary errors.

A *temporary error* has a probability of occurrence which is independent of the actuations. If such an error occurs at a component during an actuation with probability  $p$ , the same component is error-free at the next actuation, with probability  $(1-p)$ . The components themselves can therefore be said to be self-repairing. That arbitrary reliability (i.e. self-repair) can be obtained in a system with self-repairing components is not too surprising. We shall, however, pay attention to these temporary errors in connection with the question of the minimum amount of redundancy of components which is necessary for a given total reliability (Sections 6 and 7). Some measures of reliability are suggested in Section 4.

When a *stationary error* occurs at a component, it will be present at all following actuations of the component. In Section 5 it will be shown that ordinary redundancy design cannot increase the lifetime (reliability) of a system above the lifetime of a component, in the case of stationary errors. It can, however, increase the lifetime in the case of temporary errors. Hence, in order to improve the system reliability, a component replacement mechanism must be involved. The determinant of the action of the replacement mechanism is a system computing the location of errors. Stationary errors are allowed to occur in the components of this system, too.

In Section 8 the question of error-location computability with erroneous components is investigated. Affirmative answers are obtained both for  $c$ -nets and  $g$ -nets. A certain class of errors, transfer-errors, in  $c$ -nets, permit a location of errors of arbitrary weight. In  $g$ -nets, however, only all single errors can be located.

On the basis of these results of error location computability, the concept of self-repair will be treated in Section 9. In particular, it will be shown that a self-repairing  $c$ -net system can be designed to have a maximum lifetime  $T(t)$ , depending on the lifetime,  $t$ , of its components.  $T(t)$  is larger than  $t$  but not infinite when  $t$  is finite

## 2. DETERMINISTIC SYSTEMS

We shall, with motivation from Section 1, introduce the following definitions.

*To determine a physical quantity,  $q$ , is to specify the physical*

dimension of the quantity, to partition the set of all possible  $q$ 's of the dimension in question into disjoint equivalence classes (states)  $s_i$ , and to establish to which equivalence class  $q$  belongs. The notion of equivalence classes implies an equivalence relation  $E$ , i.e. a reflexive, symmetric and transitive relation.  $E$  may be defined over a measure function  $M$ , such that  $M(q) = M_i (M_i \neq M_j \text{ if } i \neq j)$  if and only if  $q$  belongs to the equivalence class  $s_i$ . Hence a determination of  $q$  means a measuring of  $q$  with a measuring function  $M$ . If  $q$  were a continuous quantity, in the sense that any  $q$  belonged to some equivalence class, then there must be two  $q$ 's, in different equivalence classes, infinitely close (over the measuring function  $M$ ) to each other. But, since a quantum of action is involved in a measurement, it is then in principle not possible to decide from two measurements (during a finite time) if the two  $q$ 's were  $E$ -equal or not. Hence, in order to be able to measure (determine)  $q$  in all circumstances,  $q$  must be thought of as quantized, i.e. as having stable states  $s_i$ , well separated by unstable transition regions.

*A deterministic system* is a system with specified output states and input states, such that the output states are determined by the input states (equation (2)). The inputs and outputs are quantized, i.e. have certain stable states ( $s_i$ , equation (1)), for which the mapping  $F$  (equation (2)) is defined. The stable states are disjoint and separated by unstable transition regions. The system is then said to be  $F$ -deterministic.

*Deterministic behavior* is the behavior (mapping equation (2)) of a deterministic system.

*An indeterministic system* is a system with specified inputs and outputs, such that the outputs are not determined under all circumstances by the specified inputs.

*Hypothesis:* Any indeterministic physical system with specified, quantized, inputs and outputs  $A$  and  $B$  (as in equation (2)), contains a system with inputs  $A \times A'$  and outputs  $B$  such that for at least one element of  $A'$  the associated system is deterministic with respect to the sets  $A$  and  $B$  ( $F$ -deterministic).

With these definitions we have a meaning for a deterministic system which is consistent with, for instance, the deterministic aspect of a digital computer. In connection with the measuring of a physical quantity, however, we are inclined to say that we have



determined the quantity when we have obtained a value, and can say with confidence that the value is correct to within a certain specified interval containing the value. With such determinations we cannot answer, in all circumstances, the question of equivalence if the quantity is regarded as continuous. In order to build a theory of deterministic systems, it is of importance to have a well-defined equivalence relation; hence we are led to the definitions above.

In a deterministic system we can regard the input and output quantities as causes and effects, respectively. That is, we regard the causes and effects as events which, with respect to the determinism under consideration, are completely characterized as elements of the input and output sets for which the mapping of the system is defined.

We can, of course, have other aspects of determinism in mind when we study an  $F$ -deterministic physical system. If, for example, an output state, of the physical dimension electric current, is generated by an electric voltage (determined by some inputs) across a resistor, the resistor is evidently one of the causes of the output. We do not ask for the cause of the resistor if it belongs to one and the same equivalence class (is constant) throughout all observations. We may, however, be interested in studying the physical system beyond its  $F$ -determinism, i.e. we would not study the  $F$ -system any more but instead the actual physical system and its history. Then the question of the cause of the resistor may be relevant. We may find that one of its causes was a particular technician. And we may proceed and ask for the causes of the technician and so on. In general there is a large number of effect-cause-sequences emanating from each effect. We may chop all these sequences off at certain effects, and speak of the immediately preceding causes as inputs. The number of these inputs is usually very large and in order to investigate eventual determinism with respect to them, we must consider a large class of constructs.

Classically, determinism is described as the doctrine that everything is entirely determined by a sequence of causes. This is compatible with the above definitions if we qualify "everything" by the quantized aspect of "everything". When tracing an effect-cause-sequence with respect to causes of a thing being present, it seems that such a sequence must be infinite. For example, it seems



impossible to give a finite deterministic explanation of the creation of the universe. However, there may be effect-cause sequences of other kinds which are finite. Such sequences form cycles. In order to give a cycle of effect-cause relations a meaning, we will have to qualify the  $F$ -function.

So far, we have not said anything about the time between a cause  $A$  and the effect  $B$ . We may imagine this time to be zero, which is all right if  $F$  does not contain cycles. If, on more realistic grounds, we associate a value  $\delta_i$  to the execution time for the  $i$ th elementary mapping of which  $F$  may be composed, it may well happen that confusion can arise from the different depths (lengths) of the cause-effect paths inside the system, even if  $F$  does not contain cycles.

A way out of this confusion is to quantize both the inputs and outputs with respect to time. We shall stipulate (see Fig. 1) that

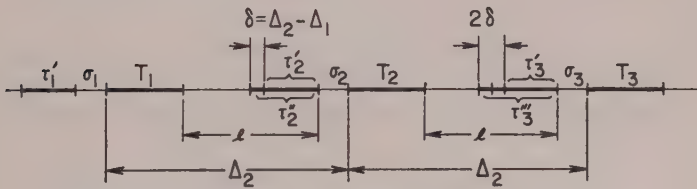


FIG. 1. Time-quantization.

all inputs shall be in stable states (belong to an element of  $A$ ) at time-intervals  $T_i$  (equivalence classes). The lengths  $||T_i||$  of these intervals are independent of the indices  $i$ , which order them according to a natural progression of time. Each  $T_i$ -interval is immediately preceded by an interval  $\sigma_i$  (a synchronization interval, to be explained below).  $||\sigma_i||$  is independent of index  $i$ . Each  $\sigma_i$ -interval is immediately preceded by a transition-interval  $\tau_i'$  in which, and only in which, a transition between two stable input states can occur.  $\tau_i''$  is similarly defined as a transition interval in which, and only in which, a transition between two stable output states can occur.  $\tau_i''$  covers  $\tau_i'$ . Both are immediately prior to the  $\sigma_i$ -interval. Each  $\tau_i$ -interval ( $\tau_i'$  and  $\tau_i''$ ) is preceded, but not immediately preceded, by  $T_{i-1}$  as defined above. The distance

between two consecutive  $T_i$ -intervals (the period) is  $\Delta_2$ , the maximum execution time for the system. The difference between the lengths of the  $\tau_i''$ - and  $\tau_i'$ -intervals is:

$$\|\tau_i''\| - \|\tau_i'\| = \Delta_2 - \Delta_1 = \delta \quad (3)$$

where  $\Delta_1$  is the minimum execution time of the system. We shall use the following notations for the set of ordered pairs  $(\tau_i', T_i)$  and  $(\tau_i'', T_i)$ :

$$T'(i) = \{(\tau_i', T_i), (\tau_{i+1}', T_{i+1}), \dots\} \quad (4)$$

$$T''(i) = \{(\tau_i'', T_i), (\tau_{i+1}'', T_{i+1}), \dots\} \quad (5)$$

$$T^j(i) = \{(\tau_i^j, T_i), (\tau_{i+1}^j, T_{i+1}), \dots\} \quad (6)$$

where:

$$\|\tau_i^j\| = \|\tau_i'\| + (j-1)\delta < l, \quad (7)$$

and  $l$  is the maximum length of the interval in between  $\sigma_{i+1}$  and  $T_i$ . For a given  $l$ , there is a finite number of allowed  $j$ -indices.

Consider the class  $K$  of systems such that each system has the quantized time  $T^j(i)$ . If a system previously described by the mapping  $F: A \rightarrow B$  (equation (2)), belongs to the class  $K$  when the execution time is also taken into account, then we can describe it by a mapping:

$$F: A \times T^j(i) \rightarrow B \times T^{j+1}(i+1). \quad (8)$$

(An input state, belonging to  $A$  at  $T_i$  determines an output state, belonging to  $B$  at  $T_{i+1}$ .) The mapping of equation (8) is formally equivalent to a mapping of equation (2), if, in the sets  $A$  and  $B$ , we also incorporate the quantized time. Hence the formal definition of a deterministic system given above need not be changed when we take the execution times into account.

Consider two systems previously described with the mappings  $F_1: A \rightarrow B$ , and  $F_2: A' \rightarrow C$ , such that  $B$  is contained in  $A'$ . If, when the execution times are taken into account, both belong to class  $K$ , they can be compounded to a system  $[F_1, F_2]$ :

$$[F_1, F_2]: A \times T^j(i) \rightarrow A' \times T^{j+1}(i+1) \rightarrow C \times T^{j+2}(i+2) \quad (9)$$

which again belongs to class  $K$ , provided that  $j' = j+2$  satisfies the inequality (7).

The quantized time  $T_i$  is to be regarded as an input quantity to

the system. A system, however, may have no  $A$ -inputs (from outside the system) and still show an output behavior. The cause-effect path then forms a cycle. With no  $A$ -inputs we have no natural association for a  $T_i$ -input. We can thus foresee some difficulties with the time concept in this case. Consider, for example,

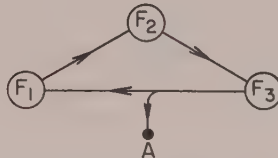


FIG. 2. System with cyclic cause-effect path.

the system  $[F_1, F_2, F_3, F_1]$ , illustrated in Fig. 2. Let each elementary system  $F_j$  be defined as:

$$F_j: \begin{pmatrix} s_1 \rightarrow s_2 \\ s_2 \rightarrow s_1 \end{pmatrix}.$$

If at  $T_1$ ,  $A$  is in state  $s_1$ , the following behavior is obtained:

$$\underline{(s_1, T_1)} \rightarrow (s_2, T_2) \rightarrow (s_1, T_3) \rightarrow \underline{(s_2, T_4)} \rightarrow (s_1, T_5) \rightarrow \dots$$

The underlined pairs represent the sequence of states that will appear at the output. However, this is correct only as long as index  $\nu$  in the mapping:  $A \times T'(1) \rightarrow A \times T''(2) \rightarrow \dots \rightarrow A \times T^\nu(\nu)$  satisfies the inequality (7), i.e. for a finite number of oscillations only.

We can, however, supply the system with an input-time,  $T_i$ , i.e. synchronize the oscillation with input pulses that occur at the  $\sigma_i$ -intervals (see Fig. 1). The effect of the synchronization pulses is such that if a state-transition occurs at  $\tau_i$ , this change of cause to the next elementary system is not transmitted to it directly, but delayed until the synchronization pulse occurs during  $\sigma_i$ .

Consider next the system of Fig. 3. We have here one input,  $A$  (and hence an associated quantized time  $T_i$ ) and one output,  $B$ . The component  $F_4$  is deterministic (two inputs and one output). If the cycle is not synchronized with  $T_i$ , the system is clearly indeterministic, although compounded of deterministic components.

Our introductory hypothesis about indeterministic systems with quantized inputs and outputs is verified for this system. For at least one value of the extra input  $A'$  ( $s_1$ , for instance) the system is deterministic with respect to  $A$  and  $B$ . Or, if we synchronize the oscillation of the cycle with the input time  $T_i$ , the output behavior will be deterministic.

The theory of deterministic systems containing cycles has been developed along different lines.<sup>(16,5,13)</sup> In the following study of errors, we shall restrict the reference systems (error-free systems)

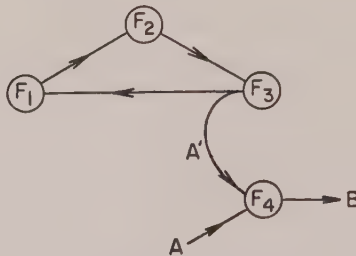


FIG. 3. Indeterministic system.

to deterministic systems with no cycles. It is then sufficient to deal with mappings of type (2) instead of type (8). In doing so, we omit one type of error, namely an error in the execution time. Such an error is equivalent, however, with an error in an ordinary state, and will hence be covered. In a deterministic system with cycles we can deal with input and output time sequences ( $A \times T^j(i)$ ). However, if we restrict ourselves to finite time sequences, we can instead reestablish the time ordering with a parallel ordering: an input- and out put-ordering ( $A \times A' \times A'' \times \dots$ ). In other words, we replace a sequential ordering with a parallel ordering. So, the cycle-free physical system  $F : A \rightarrow B$  can be regarded as a canonical form of a model of any finite, deterministic event.

A behavior of form (2) is the characteristic behavior of a switching net, which we thus will take as a reference frame for the following error study. There are two and only two different structures of uniform, composable switching nets, the  $c$ -structure and the  $g$ -structure.<sup>(10)</sup> Any gate-net is said to have a  $g$ -structure. A

*c*-net is a degenerate case of a *g*-net, namely one where all gates contain only "wires". A wire is a two-terminal element such that  $A=O$  for all  $T$ . The real numbers  $A$  and  $T$  are the across and through quantities<sup>(4,15)</sup> related to the terminals.  $A$ -quantities, measured across a pair of connection points, may be electric voltages, temperature differences, and so on.  $T$ -quantities, measured as flowing through a connection, may be electric currents, rates of heat flow, and so on. A state of a *c*-net can only be determined by measuring both  $A$  and  $T$  of a terminal pair. There are only three possible states of a *c*-net.<sup>(10)</sup> They may be defined over  $A$  and  $T$  relations  $R_1, R_2, R_3$ . A measurement is defined over a fourth relation  $R$  such that  $R_i \cap = (A_i, T_i)$ . For distinct indices:  $A_i \neq A_j, T_i \neq T_j$ . The so measured  $A_i$ 's (or  $T_i$ 's) can be taken as  $q$ 's and define the states  $s_1, s_2, s_3$  according to equation (1).  $s_1$  and  $s_2$  are implied by the relations:

$$\begin{aligned} s_1: R_1 &= \{(A,T) \mid T = O \text{ for all } A\} \\ s_2: R_2 &= \{(A,T) \mid A = O \text{ for all } T\} \end{aligned} \tag{10}$$

Two-valued *c*-nets may thus have states of the dimensions of electric resistance (admittance), magnetic reluctance and so on. The third relation,<sup>(10)</sup> which implies the state  $s_3$ , gives a realizable *c*-net which can be described with the three-valued Post algebra. It is usually required that the input quantities to a *c*-net be  $A$ - or  $T$ -quantities. Then 1-1 switches, which convert these quantities to *c*-state quantities, are required. For the two-valued *c*-net the relay and the cryotron are commonly used 1-1 switches.

A *g*-net can have any number of states. All proper gates are nonlinear with respect to the elementary algebra over the reals.

### 3. COMPONENT ERRORS

An error has been defined as a deviation from "truth". However, "truth" is in general not known and knowable, and has to be hypothesized. We shall make the hypothesis that "truth" is a deterministic behavior.

It should be observed at this point that although we have defined the notion of deterministic behavior, this does not permit us to say with certainty that a physical system is deterministic. We can eventually say that it is not deterministic with respect to



its specified outputs and inputs. But if we cannot say that it is indeterministic, we cannot be sure that it is deterministic until we have tested it for all future actuations. Since this is impossible, we can do no more in this situation than make the hypothesis that it is deterministic.

We shall now investigate how the definitions of determinism can be violated, and thus obtain a classification of errors. We are in particular interested in component errors, i.e. errors in the non-linear components of the canonical model of a deterministic system. These components are the gates in a  $g$ -net and the 1-1 switches in a  $c$ -net. The connecting wires are assumed to be properly connected.

For reasons indicated later in this section, we shall consider nets composed of elementary components, i.e. components with about the same complexity, i.e. with about the same number of inputs, and each with one output.

We are led in the first place to the following classification of component errors:

- $A_1$ : At a certain actuation, i.e. when stimulated with certain stable (error-free) input states, a component gives a stable output state that belongs to the set  $S_m$  of its output states but which is different from the particular state prescribed by the deterministic behavior of the component at this actuation. (Mapping-violation.)
- $A_2$ : A component gives a stable output state which does not belong to the set  $S_m$  of its (deterministic) output states. (Quantization-violation.)
- $A_3$ : A component gives a continuously oscillating (unstable) output state not contained in some elements of  $S_m$ . (Quantization-violation.)

Component errors of types  $A_2$  and  $A_3$  may be observed by measuring the output state of a component alone. An error of type  $A_1$ , however, can only be observed by measuring both the output state and the input states of the component and making a following combinatorial decision. We shall in the following be concerned with errors of type  $A_1$ . Errors of type  $A_2$  and  $A_3$  are easier to deal with because a component output contains here in itself information about an eventual error. Some problems of combinatorial nature concerning the location of an erroneous component can,



however, be anticipated. (An error of type  $A_3$  can be indicated also as  $A_3$ -errors at the outputs of succeeding components.)

Concerning the cause of an error (compare the above hypothesis of Section 2 which asserts a cause to each error) in a component, we will make the following classification of  $A_1$ -errors:

$B_1$ : A component error can be determined at any one actuation by measurements of the input states and the output state of a component.

$B_2$ : A component error can be determined by measurements of the input states and the output state of a component at specific actuations only.

For example, in the case of binary states (where index  $i$  in equation (1) is two-valued), a  $B_1$ -error can be represented by  $s_i'$  (the negation of  $s_i$ ) where  $s_i$  is the error-free output state of the component. Examples of  $B_2$ -errors can be found in threshold components of  $g$ -type.<sup>(11,1)</sup> If the threshold  $\theta$  is 2 in a binary "and"-gate and is changed to  $\theta = 1$  by the error, the gate is converted into an "or"-gate. However, this threshold error cannot be determined (has no effect) at an actuation where both inputs are either stimulated or nonstimulated. For all other actuations this threshold error has effect, however. Errors of type  $B_2$  alone do not represent the behavior of a physical component in a realistic way. They imply that the components are ideal (perfectly error-free) in a certain class of actuations. It is reasonable to assume, however, that the probability of an error in the output of a component of a certain construction depends on the actuation of the component. But this probability should not be precisely zero for any actuation. In what follows we will therefore be mainly concerned with errors of type  $B_1$ . In Section 7 an example is given of a redundancy design for a  $g$ -net insensitive for  $B_2$ -errors. For this net the von Neumann bundle-line trick<sup>17</sup> can be avoided, which might seem surprising. It should be kept in mind, however, that the  $B_2$ -errors represent an idealized situation.

Concerning the statistical nature of the causes of error, we will treat the following cases:

$C_1$ : Temporary errors. A cause of error appears with probability  $p$  at an elementary component upon an actuation. The appearance of error causes at different elementary components upon an actuation are independent of each

other. The appearance of error causes at an elementary component upon different actuations are independent of each other.

$C_2$ : Stationary errors. An error cause appears with probability  $p$  at an elementary component upon an actuation provided that no error cause has appeared at the component upon a previous actuation. The first appearance of error causes at different elementary components upon an actuation are independent of each other. After a first appearance of an error cause at an elementary component, the error cause appears at the component upon all following actuations.

In this  $C_1$ - $C_2$  classification, we have qualified the components as elementary. The idea is that it is reasonable to assume the same probability of error for components of about the same complexity, i.e. with about the same number of error causes. It is thus reasonable to assume about the same probability of error for the elementary "and"-, "or"-, and "not"-components (with two or one inputs) of a  $g$ -net with binary states provided that these components are uniformly constructed with a nonlinear device. An "and"-gate with, say, 100 inputs, again constructed out of the same type of nonlinear device, should, however, have a larger probability of error than the elementary (2-input) "and"-gate.

Gates of threshold type require in this connection a special comment. In a threshold gate an algebraic sum,  $A$ , is formed with linear components. This sum is compared with a threshold,  $\theta$ . If  $A \geq \theta$ , the gate fires. If  $A < \theta$ , the gate does not fire. If we associate errors with the nonlinear part of the gates, the firing mechanism, it would seem that we could add more and more inputs to the gate and still maintain the same reliability. That this is not so can be seen as follows. Consider the following two actuations  $a_1$  and  $a_2$  of a threshold gate with  $m$  inputs:

$$a_1 : s_j, s_1, \dots, s_n : A_1 = \theta,$$

$$a_2 : s_k, s_1, \dots, s_n : A_2 < \theta.$$

$a_1$  and  $a_2$  differ only with respect to one of the  $m$  inputs. At  $a_1$  this input is fed with state  $s_j$ , and at  $a_2$  with the distinct state  $s_k$  (cf. equation (1)). The algebraic sum  $A$  at  $a_1$  is just large enough for the gate to fire. At  $a_2$  the gate does not fire. With reference to

equation (1) we obtain:

$$\begin{aligned} \min A_1 &= \sum q_i' = q_j' + \sum' q_i' \\ \max A_2 &= \sum q_i'' = q_k'' + \sum' q_i'' \\ \min(A_1 - A_2) &= \min A_1 - \max A_2 \\ &= q_j' - q_k'' - \sum'(q_i'' - q_i') \\ &< q_j' - q_k'' - (m-1) \cdot \min(q_i'' - q_i'). \end{aligned}$$

If  $\min(q_i'' - q_i') > 0$ , we see that for a sufficiently large number of inputs,  $m$ ,  $A_1$  can be less than  $A_2$ , which means that the threshold gate will operate in an incorrect way. That the stable states in a threshold gate have to be defined with  $\min(q_i'' - q_i')$  larger than zero, and not precisely equal to zero, is obvious. Considering fluctuation errors in the threshold, the probability of error for a threshold gate will increase with the number of inputs. The  $C_1$ - $C_2$  classification is thus relevant only if the gates are qualified as elementary, in the sense that they are uniformly constructed and have about the same number of inputs.

In the case of  $g$ -nets with multi-valued states such that index  $i$  of equation (1) can assume more than two values, and errors of type  $B_1$ , the expression of the effect of an error on the output-state is more complicated than in the binary case. We will come back to this situation in Section 7. For the moment it is sufficient to state only the probability of the error cause as we have done in the  $C_1$ - $C_2$  classification.

#### 4. RELIABILITY

Let us begin by loosely characterizing the reliability of a system as some measure of the similarity between the (input-output) behavior of the system and its hypothesized deterministic (input-output) behavior. (We have already introduced a probabilistic measure of the reliability of an elementary component,  $q = 1 - p$ , with respect to  $(A_1, B_1, C_1, C_2)$ -errors.)

The main problem of this paper is to investigate in a qualitative way how systems of unreliable elementary components, i.e. deterministic components in which errors are introduced as under  $(A_1, B_1, B_2, C_1, C_2)$ , can be synthesized so that the systems will be more reliable than the elementary components.

As the problem is formulated, it will lead to a study of redundancy techniques, i.e. how a number of elementary components, excessive for the operation of the system in the ideal case of no errors, can be used to diminish the effect of the component errors on the outputs of the system.

Another type of problem would be to investigate how the reliability of a single component can be improved, i.e. to investigate how well a deterministic behavior of the component can be produced by isolation from disturbing events in nature. Again this is an application of redundancy. In its most general form, however, we should allow here for redundancy in objects of kinds which would not be used at all in the ideal situation of no errors. For instance, a shield around the component, protecting it to a certain degree from disturbing events, is a redundancy application of this sort. We will, in what follows, assume that this type of improvement of the component reliability has been made in a uniform way, so that the probability of an error cause in each elementary component is  $p$ , as specified in Section 3.

It is important to emphasize that when redundancy is applied on different complexity levels, quite distinct effects on the system reliability may be obtained.<sup>(6)</sup> As the main problem is formulated, it allows for redundancy applications on the level of elementary components and on higher levels (redundancy of aggregates of elementary components). The elementary components are identified as nonlinear devices as contrasted to the connecting "wires". A system consists of elementary components and "wires" only. With each elementary component (a gate in a proper  $g$ -net and a 1-1 switch in a  $c$ -net) we have associated an error.

However, we can make a milder assumption regarding the errors, namely, that an error is associated with each nonlinear subcomponent of an elementary component, but such that each linear subcomponent is error-free. For example, we may associate an error to each diode in a diode gate, but consider its resistors and capacitors error-free. Then we can apply redundancy on the level of the nonlinear subcomponents (diodes) and obtain a result quite distinct from that obtained with redundancy on the gate level. For  $(A_1, B_1, C_1)$ -errors it is in the latter case (gate-level redundancy) necessary to apply the von Neumann<sup>(17)</sup> bundle-line trick in order to obtain an arbitrarily high system reliability. In the former case

(micro-level redundancy) such a reliability can be obtained without this trick.<sup>(6)</sup> The bundle-line representation is somewhat unsatisfactory because it implies an incomplete determination (computation).

In what follows we will be primarily concerned with the main problem as formulated above, i.e. with redundancy applications on the level of elementary components.

In order to suggest some reliability definitions, let us consider  $A_1$ -errors in a system  $F$ . It can thus be characterized with a mapping:

$$F = A \times E \rightarrow B. \quad (11)$$

$E$  is the set of error causes. Each element  $e$  of  $E$  is an  $n$ -tuple

$$e = (e_1, e_2, \dots, e_n) \quad (12)$$

such that  $e_i = 1$ , if there is an error cause in the  $i$ th elementary component, and for no error cause in the  $i$ th component,  $e_i = 0$ . The system consists of  $n$  elementary components.

Let  $w$  be the weight of an  $e$ -element, i.e.  $w$  is the number of 1's in  $e$ . Let  $E_w$  be a subset of  $E$ , such that  $E_w$  contains all  $e$ -elements of  $E$  with weight less than or equal to  $w$ . Consider a mapping:

$$F_w : A \times E_w \rightarrow B \quad (13)$$

which characterizes a function  $F_w(A)$ , such that the function  $F$ , defined by equation (11), is an extension of  $F_w(A)$  for each  $w$ . We say that the system  $F$  is  $w$ -error insensitive if  $F_0(A) = F_w(A) \neq F_{w+1}(A)$ .

In the case of  $C_1$ -errors, the probability of an  $e$ -error of weight  $\nu$ :  $p^\nu(1-p)^{n-\nu}$ , is independent of the actuation  $a_i$  (element of  $A$ ). Depending on the actuation and the structure of the system, this  $e$ -error may or may not result in an error in the overall behavior of the system at this actuation. The probability of output error at actuation  $a_i$  is thus:

$$P(a_i) = \sum_{\nu=1}^n n_{i,\nu} p^\nu (1-p)^{n-\nu}, \quad (14)$$

where  $n_{i,\nu}$  is the number of  $e$ -errors of weight  $\nu$  which give an output error at this actuation.



The probability of an output error at actuation  $a_i$  in a  $w$ -error insensitive system is:

$$\begin{aligned}
 P_w(a_i) &= \sum_{\nu=w+1}^n n_{i,\nu} p^\nu (1-p)^{n-\nu} \\
 &= n_{i,w+1} p^{w+1} + O(p^{w+2}).
 \end{aligned}
 \tag{15}$$

We are in particular interested in the asymptotic probability of error where  $p$  goes to zero. Here the  $p^{w+1}$ -term of equation (15) will dominate. In this asymptotic sense we can say that any  $w$ -error insensitive system is more reliable than a  $(w-1)$ -error insensitive system. However, if we want to compare two  $w$ -error insensitive systems with the same  $F_0(A)$ -function, we must introduce a measure of the coefficient  $n_{i,w+1}$  of equation (15). One possibility is the measure:

$$M_{1\frac{1}{2}} = \max_i (n_{i,w+1}). \tag{16}$$

We are thus led to the following reliability measures:

- |                                     |  |                          |
|-------------------------------------|--|--------------------------|
| (i) $w$ -error insensitivity        |  | ( $C_1$ -errors)         |
| (ii) $1 - M_1 p^{w+1}$              | (cf. equation (16))  | ( $C_1$ -errors)         |
| (iii) $1 - \max_i P(a_i)$           | (cf. equation (14))  | ( $C_1$ -errors)         |
| (iv) Average lifetime of the system | (mean free path, or average number of actuations, between two output errors) | ( $C_1$ - $C_2$ -errors) |

The first three reliability measures do not depend on any specific assumption regarding the statistical distribution of the actuation  $a_i$ . We will in the next section also qualify the fourth measure in this way.

##### 5. TEMPORARY AND STATIONARY COMPONENT ERRORS IN ORDINARY REDUNDANCY DESIGNED SYSTEMS

We shall here investigate the reliability (lifetime) of an ordinary redundancy designed system in the cases of ( $A_1, B_1, C_1, C_2$ )-errors.

We assume that the system is  $w$ -error insensitive. At a certain



actuation, we will have the probability  $Q$  of correct behavior:

$$Q = Q(q^*, p^*) = 1 - P = 1 - M_1(p^*)^{w+1} + 0((p^*)^{w+2}). \quad (17)$$

In the temporary case,  $p^*$  stands for the probability  $p$  of a component error upon an actuation ( $q^* = 1 - p^*$ ). For stationary errors, the probability that a component is in an error-free state on the  $\nu$ th actuation is  $q^* = (1 - p)^\nu$ , because we must require here that no error has occurred during the previous  $\nu - 1$  actuations ( $p^* = 1 - q^*$ ). Thus, the probability of at least one output error during  $\nu$  actuations,  $P_t(\nu)$  and  $P_{st}(\nu)$ , respectively, for temporary and stationary errors, will be:

$$P_t(\nu) = 1 - Q_t(\nu) = 1 - [Q(1 - p, p)]^\nu \quad (18)$$

$$P_{st}(\nu) = 1 - Q_{st}(\nu) = 1 - Q[(1 - p)^\nu, 1 - (1 - p)^\nu]. \quad (19)$$

In equation (19) we should not raise  $Q$  to any power other than 1, because if there is no output error at the  $\nu$ th actuation, for which  $Q(q^*, p^*)$  is the probability, then there cannot have been any output errors at previous actuations in this case of stationary errors. This argument becomes binding when using  $M_1$ -measures for  $Q$  (equation (16)), i.e. regarding all  $\nu$  actuations as switchings between the most sensitive input-states. Expanding  $Q$  according to equation (17) we obtain:

$$P_t(\nu) \sim M_1 \cdot \nu \cdot p^{w+1} \quad (20)$$

$$P_{st}(\nu) \sim M_1 \cdot (\nu \cdot p)^{w+1}. \quad (21)$$

We see that for small  $\nu$ , the error probabilities will be about the same for the two types of error. But very soon, i.e. for a slight increase in  $\nu$ ,  $P_{st}(\nu)$  will dominate over  $P_t(\nu)$ . For temporary errors we can in fact increase without limit the mean free path between output errors.<sup>(17,12)</sup> It is not so for stationary errors. The lifetime,  $\bar{\nu}$ , of the system will be about the same as for an elementary component, i.e.  $\bar{\nu} \approx \frac{1}{2p}$ . For if in equation (19),  $q^* = (1 - p)^\nu = \frac{1}{2} = p^*$ , this corresponds to complete uncertainty.

Hence, in order to increase the reliability of a system with stationary errors, a component replacement mechanism must be involved. This will lead to the ideas of error-location computation and self-repair, to be treated in Sections 8 and 9.

6. MINIMUM REDUNDANCY  $c$ -NETS  
INSENSITIVE TO TEMPORARY  $w$ -ERRORS

We will deal primarily with binary nets. It is then convenient to express the behavior of the mapping  $F : A \rightarrow B$  as a system of Boolean functions. Let us first consider a system with only one output, i.e. a single Boolean function. Let  $m$  be the number of irredundant literals of the function (a letter, or variable,  $x$ , corresponds to one literal when affirmed and to another literal when negated). Let  $n$  be the number of components (1-1 switches) in the  $c$ -net. Each literal of the Boolean function  $B$  corresponds to one or more components.

The redundancy,  $r$ , of the net is:

$$r = \frac{n-m}{n} = 1 - \frac{m}{n}. \quad (22)$$

The redundancy is zero for  $n = m$ , i.e. for irredundant nets. Not all Boolean functions can be realized with irredundant nets. We have investigated this realizability question in Refs. 7 and 8.

For functions which do not have irredundant nets,  $n_{\min} > m$ . The corresponding residual redundancy or "internal redundancy" is due to realizability restrictions.

The question of this section is: which is the least amount of redundancy (equation (22)) which permits  $w$ -error insensitivity?

We have in Ref. 9 answered the question in terms of bounds on  $r_{\min}(w)$ :

$$1 - \frac{1}{(w+1)^2} \leq r_{\min}(w) \leq 1 - \frac{m}{M} \frac{1}{(w+1)^2}. \quad (23)$$

$M$  is the minimum number of components in a net generating the prescribed Boolean function in the ideal case of no errors.

For a given function  $B$ , i.e. for a specified  $m$  and  $M$ , the minimum number of components,  $n(w)$ , that permit  $w$ -error insensitivity is bounded by:

$$m(w+1)^2 \leq n(w) \leq M(w+1)^2. \quad (24)$$

Then let the quantity  $R(w)$  (maximum "rate" of computation for

$w$ -error insensitivity) be defined as:

$$R(w) = \frac{m}{n(w)} \leq \frac{1}{(w+1)^2} \quad (25)$$

such that  $r_{\min}(w) = 1 - R(w)$ .

We see from equation (24) that  $n(w) > w$ . Hence there must be a  $(w+1)$ -error among these  $n(w)$  components in the net for which an output error will occur. Complete reliability can thus never be obtained for a finite  $w$ .

However, when  $w \rightarrow \infty$ , we can reach:<sup>(12)</sup>

- (i) arbitrarily high reliability
- (ii, iii) a reliability arbitrarily close to 1
- (iv) arbitrarily long lifetime

with respect to which reliability measure we prefer (cf. Section 4).

We see from equations (23), (24), (25) that for each of these reliability limits,  $r_{\min}(w) = 1$ ,  $n(w) = \infty$ ,  $R(w) = 0$ , independently of the specified Boolean function.

In the case of a finite reliability (finite  $w$ ), we see from equation (23) that for Boolean functions which have irredundant nets ( $M = m$ ),  $r_{\min}(w)$  is uniquely determined by its bounds. For other Boolean functions it is more difficult to determine  $r_{\min}(w)$  precisely.

One method, as outlined in Refs. 7 and 9, is to start with a specified redundancy distribution. This specifies, together with the Boolean function and the  $w$ -error insensitivity, part of the truth-table for the redundancy net. By investigating for realizability those loop-matrices which fulfil the truth-table specification (two necessary and sufficient conditions: the  $c$ -criterion and the sub-rearrangement theorem), an eventual sufficiency of the assumed redundancy is established. If the redundancy  $r_1 = (n-m)/n$  is sufficient, but  $r_2 = (n-m-1)/n-1$  is not, then  $r_1$  is the minimum redundancy. Even if the number of possibilities for an investigation of this sort can be restricted somewhat more than indicated above, the method is effective only for small values of  $m$  and  $w$ .

Concerning the closeness of  $R(w)$  to its upper bound  $R_{u.b.}(w)$  as defined in equation (25), we can make the following statement. There exists a denumerably infinite set of increasing  $w$ -integers for which the quotient  $\kappa(w)$  ( $\leq 1$ ) between  $R(w)$  and  $R_{u.b.}(w)$  is a nondecreasing function. This is easily seen by replacing each component of a  $w_1$ -error insensitive net by a  $w_2$ -error insensitive

net for the identity  $B(a) = a$  (containing  $(w_2 + 1)^2$  components). The resulting net will be  $(w_1 w_2 + w_1 + w_2)$ -error insensitive and has the same  $\kappa$  as the first net.

As an illustration, let us consider the majority function of three variables:

$$B = ab \cup bc \cup ca. \quad (26)$$

The minimum number of components,  $M$ , needed to realize  $B$  is five. In order to make the net single-error insensitive, we could replace each component with a single-error insensitive net for the

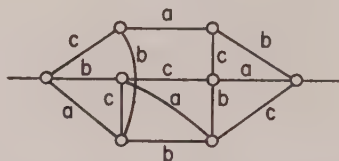


FIG. 4. Minimum redundant, single-error insensitive  $c$ -net for the majority function of equation (26).

identity function (containing four components: for their connection, see Ref. 7). This would give  $n = 20$ , corresponding to the upper bound of equation (24). The lower bound of equation (24) gives  $n = 12$ . The actual minimum number  $n(1)$  is 14. A corresponding net is shown in Fig. 4. This example (and further iterations) shows explicitly that, in general, the method of applying redundancy on a structure generating a desired  $B$ , such that this basic structure is maintained in the redundancy net, may not be an optimal method.

Let us next turn over to a system  $F : A \rightarrow B$  with more than one output (a system of Boolean functions). Again let the number of irredundant literals be  $m$  (each of the  $m$  literals is irredundant with respect to at least one of the functions of the system). Suppose that the system can be generated in a common terminal net<sup>(8)</sup> containing  $m$  components, i.e. an irredundant net. The net will be  $w$ -error insensitive with respect to each output if each component is individually  $w$ -error protected. This requires a number  $n(w) = m(w + 1)^2$  of components. This is the minimal complexity,

since the system has an irredundant net. Hence  $R(w) = 1/(w+1)^2$  (see equation (25)). In case the system has no irredundant net,  $R(w)$  will be less than or equal to this value. So also in this case of a system of Boolean functions we can obtain arbitrarily high reliability ( $w \rightarrow \infty$ ) only if  $R \rightarrow 0$ . This result is independent of the particular system of functions under consideration.

It is interesting to compare this result with the Shannon capacity theorem<sup>(14,3)</sup> for reliable communication over a noisy binary symmetric channel. The communication is usually thought of as a time-sequence of transmissions, but can as well be regarded as a parallel operation  $F: A \rightarrow B$ . The sets  $A$  and  $B$  (cartesian products) are equal, and an arbitrarily reliable communication means that an element  $a$  of  $A$  is mapped onto the same element  $a$  of  $B$  with a probability that is arbitrarily close to 1. The details of the communication system are as follows.  $A$  and  $B$  are the inputs and outputs of the whole system.  $A$  is first encoded to  $A'$  over an error-free encoder.  $A'$  is transmitted over the noisy channel and  $B'$  is received.  $B'$  is finally decoded to  $B$  over an error-free decoder. If the message to be communicated is a sequence of  $m$  binary symbols, the corresponding element of  $A$  is an  $m$ -tuple ( $A$  and  $B$  are sets of  $m$ -tuples).  $A'$  and  $B'$  are sets of  $n$ -tuples ( $n > m$ ). The capacity theorem now says that for a rate  $R = m/n$  which is less than the capacity  $C$  of the channel, an arbitrarily reliable communication can be obtained, provided that  $m$  is large enough ( $m \rightarrow \infty$ ).  $R$  can thus be a positive quantity, different from zero (i.e. the redundancy  $r$  is less than and not equal to one), and arbitrarily reliable communication is still possible.

This is in sharp contrast to the computation case, where we have just seen that  $R = 0$  is necessary for an arbitrary high reliability. The channel capacity  $C$  is defined as the maximum rate for which arbitrary reliable communication is possible. With an analogous definition of the capacity of computation for  $c$ -nets, we have thus reached the conclusion that this capacity must be zero.

These different results in the computation and communication cases are, however, not too surprising, considering the different error-distributions. In the communication case, only the channel is exposed to noise. The encoder and decoder are assumed perfectly error-free. In the computation case, of course, all computing components are exposed to noise.



7. MINIMUM REDUNDANCY  $g$ -NETS  
INSENSITIVE TO TEMPORARY  $w$ -ERRORS

Consider a  $g$ -net (not necessarily binary) with one output. This output must be the output-state of a component (gate). Thus, in the case of  $(A_1, B_1, C_1)$ -errors, the probability of an output-error at an actuation must be at least  $p$ . No redundancy design can give even a single-error correction.

However, for certain  $(A_1, B_2, C_1)$ -errors,  $w$ -error insensitivity can be obtained in a  $g$ -net with one output. An example is the net\* of Fig. 5. Each of the three "or"-gates is supposed to be of

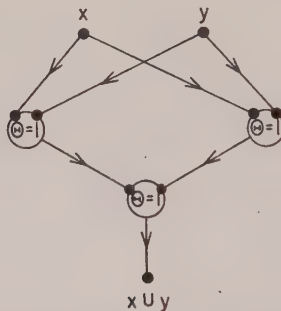


FIG. 5.  $g$ -net of threshold-type, insensitive for single  $B_2$ -errors (an error is supposed to increase an individual threshold by one unit).

threshold type with  $\theta = 1$ . The over-all behavior of the net is that of an "or"-gate. Suppose now that the  $B_2$ -errors are such that  $\theta$  changes (independently) from 1 to 2 with probability  $p$ , with the effect that the gate in which the error occurred is converted to an "and"-gate. The whole net is easily checked to be single-error insensitive. Again if the net is iterated once more, so that each gate of the net of Fig. 5 is replaced by a triplet, then the net will be 3-error insensitive. And for  $\nu$  iterations a  $(2^\nu - 1)$ -error insensitive net is obtained.

If the error-free threshold of the gates of Fig. 5 were 2, i.e. the gates were "and"-gates, the over-all behavior would be that of an

\* The net was suggested by S. Amarel<sup>(1)</sup> as a special case of W. S. McCulloch's concept of logical stability<sup>(11)</sup>.



“and”-gate. If the  $B_2$ -errors were such that  $\theta$  changes (independently) to 1, the net would be single-error insensitive. And we could proceed with further iterations.

As we have pointed out before, the  $B_2$ -errors represent a somewhat unrealistic situation, because the gates are assumed perfectly error-free for certain actuations. We will therefore turn to ( $A_1, B_1, C_1$ )-errors.

In order to remove the residual error (the error from the final output-gate), von Neumann<sup>(17)</sup> suggested a statistical state-representation, the bundle-line representation. Each output of the  $g$ -system is here represented by a bundle of lines (wires, outputs), say  $N$ .

In the case of binary states, which von Neumann deals with, a positive number  $\Delta < \frac{1}{2}$  is chosen. If  $(1 - \Delta)N$  or more lines of the bundle are in state  $s_1$ , the bundle is said to be in state  $s_1$ . If  $\Delta N$  or less lines are in state  $s_1$ , the bundle is said to be in state  $s_2$ . If any other number of lines are in state  $s_1$ , the bundle is said to be in error. In this way, no output (bundle-state) is the effect of a single gate, and the reliability of the system can be made arbitrarily high (arbitrarily close to one).<sup>(17)</sup>

However, with this bundle-line representation of a state, the system cannot be said to be complete. The behavior of the system is of the nature of combinatorial computation. And the part of the system which is left out, if the bundle-representation is used, is a majority-taking device (again a combinatorial computation). So if we really want to have a complete system which is  $w$ -error insensitive, we can connect each bundle to a  $w$ -error correcting majority organ of  $c$ -type (compare the net of Fig. 4). Then the resulting net will be of mixed  $g$ - and  $c$ -type. Or we can connect the bundles to a majority organ of  $g$ -type and make the assumption that the corresponding gates are sufficiently error-free. Some of the gates must then be perfectly error-free, which contradicts our previous assumption that errors (noise) are uniformly introduced to all the components of the system.

Let us however proceed with the idea of a bundle-representation. Even if this is a deviation from our main line of thought, it can be defended by saying that the (error-free) observer, who is going to use the outputs of the system, may have in him a majority organ of some sort. There should be some limit, however, on this line

of reasoning, for otherwise we could once and for all assume that the observer has in him the whole combinatorial system and only observes the (error-free) inputs to it.

So let us consider a system with binary gates and one output. Let the output state be represented by a bundle-state, such that the output of the  $i$ th line of the bundle is  $B_i$ . We require that each  $B_i$ , in the case of no error, carries full information about the correct output  $B$ . We have then only two possibilities for each  $B_i$ . It must equal either  $B$  or  $B'$  (the negation of  $B$ ). We will assume that all  $B_i$  equal  $B$ . We then ask the question: what is the minimum number of lines  $N(w)$  in the bundle for  $w$ -error insensitivity? Evidently  $N(w) = 2w + 1$ . If it were less, for instance,  $2w$ , we could have a  $w$ -error on the  $B_i$ -outputs that at the same time is a  $w$ -error on the negated  $B_i$ -outputs. In other words, we cannot decide for all  $w$ -errors whether a given set of  $B_i$ 's corresponds to  $B = s_1$  or  $B = s_2$ . But if the number of lines is  $2w + 1$  or larger, we can, in all circumstances, i.e. for any  $w$ -error among the  $N(w)$  final gates, decide whether a given set of  $B_i$ 's corresponds to  $B = s_1$  or  $B = s_2$ .

If the given system behavior is the Boolean function  $B(x_1, x_2, \dots, x_\mu)$ , the simplest way of generating a redundant system with the desired  $(2w + 1)$ -output bundle, such that the whole system will be  $w$ -error insensitive, is to make  $2w + 1$  copies of a nonredundant net for  $B$ .

It should be remembered, however, that the  $w$ -error insensitivity is an asymptotic measure of the reliability (for small values of  $p$ ). If the probability  $p$  of a gate-error is large, and the number of gates,  $m$ , in a nonredundant net for the generation of  $B$  is large, the above redundancy design may be inefficient.

But if  $mp$  is small compared to  $\frac{1}{2}$ , we can estimate the probability of an output error as:

$$\frac{(2w+1)!}{w!(w+1)!} (mp)^{w+1}.$$

(The probability of an output error in any one of the  $2w + 1$  net copies is  $mp$ .) Von Neumann<sup>(17)</sup> suggested a redundancy scheme (with a redundancy far above  $2w + 1$ ) with which arbitrary reliability can be obtained if  $p$  is less than a value which is independent of  $m$ .

Let us, however, assume that  $mp$  is small compared to  $\frac{1}{2}$ , and investigate whether  $w$ -error insensitivity can be obtained with a redundancy less than  $2w+1$  (or with  $r$  less than  $1-1/(2w+1)$ ; compare equation (22)). So far, we have shown only that the redundant system must have an output-bundle of  $2w+1$  lines.

Consider a system for a single-error insensitive generation of a function  $B$ . The system has three outputs,  $B_1, B_2, B_3$ , each generating  $B$  if no error occurs. Let the minimum complexity (the number of gates) of a single net generating  $B$  be  $m$ . We know that a system consisting of three such nets, i.e. with complexity  $3m$ , is single-error insensitive. We want to investigate whether we can reduce the complexity further, i.e. if the nets can have some gates in common, and still be single-error insensitive.

Let us assume that the nets for  $B_i$  and  $B_j$  have a sub-network, generating the function  $\phi'$ , in common. Then  $B_i = B_i(x_1, x_2, \dots, \phi')$ ,  $B_j = B_j(x_1, x_2, \dots, \phi')$ . Suppose that  $\phi'$ , but not  $\phi$ , is a state (an input state or a gate-output) inside the network for  $B_i$  and  $B_j$ . Then, if an error in  $\phi'$  results in an error in  $B_i$ , it must also result in an error in  $B_j$ , i.e. two of the output lines will be in error. Such a total error must occur for at least one actuation, for otherwise the sub-network generating  $\phi'$  would be redundant, which contradicts our assumption that the  $(B_i, B_j)$ -net is of minimum complexity. A situation where an error in  $\phi'$  results in an error in  $B_i$  but not in  $B_j$  cannot occur. For under the assumption that  $\phi'$  but not  $\phi$  is a state, we must have one of the following three forms of the output functions:

$$\begin{aligned} B_i &= B_j = \phi'R_1 \cup R_0 \\ B_i &= B_j = \phi R_2 \cup R_0 \\ B_i &= B_j = \phi'R_1 \cup \phi R_2. \end{aligned} \tag{27}$$

$R_0, R_1, R_2$  are functions of  $x_1, x_2, \dots$ , not decomposable in  $\phi$  or  $\phi'$ . (An error in  $\phi'$  implies an error because  $\phi'$ , but not  $\phi$ , is a state.)

If, however, not only  $\phi'$  but also  $\phi$  is a state, it can happen that an error in  $\phi'$ , but not in  $\phi$ , results in an error in  $B_i$  but not in  $B_j$ . We can, however, at once conclude, for this situation, that  $\phi$  must be an input state, i.e. one of the  $x_i$ 's. The  $\phi'$ -state must be generated from the  $\phi$ -state over a succeeding negation gate (or vice versa).

But if the  $\phi$ -state is a proper gate-output, it can be in error, and then the generated  $\phi'$ -state will also be in error. So  $\phi$  must be an (error-free) input state, and then  $\phi'$  can be in error alone. Let us write  $B_i$  and  $B_j$  as:

$$B_i = \phi R_{i,1} \cup \phi' R_{i,2} \quad (28)$$

$$B_j = ((\phi R_{j,1} \cup \phi' R_{j,2})')' = (\phi R_{j,1}' \cup \phi' R_{j,2}')' \quad (29)$$

where  $R_{i,1}$ ,  $R_{i,2}$ ,  $R_{j,1}$ ,  $R_{j,2}$  are functions of  $x_1, x_2, \dots$  but not of  $\phi$ , which according to the above is an input state, say  $x_\mu$ . If there are no errors, we have:

$$R_{i,1} = R_{j,1} \quad (30)$$

$$R_{i,2} = R_{j,2}. \quad (31)$$

If there is error in  $\phi'$  (but not in  $\phi$ ), i.e.  $\phi' \rightarrow \phi$ , we require that at most one of  $B_i$  and  $B_j$  can be in error, i.e.:

$$\phi R_{i,1} \cup \phi R_{i,2} = \phi R_{j,1}' \cup \phi R_{j,2}'$$

i.e.:

$$R_{i,1} \cup R_{i,2} = R_{j,1}' \cup R_{j,2}'. \quad (32)$$

Equations (30), (31), (32) have the unique solution:

$$R_{i,1} = R_{i,2}' = R_{j,1} = R_{j,2}' \quad (33)$$

i.e.:

$$\begin{aligned} B_i = B_j &= \phi R_{i,2}' \cup \phi' R_{i,2} \\ &= \phi \oplus R_{i,2}, \quad (\text{addition mod } 2) \end{aligned} \quad (34)$$

is the only type of function which eventually permits a reduction of the complexity  $3m$ . If  $R_{i,2}$  is not an input state, and not the mod 2 sum of an input state and a function of input states, the complexity will again be  $3m$  for this redundancy design where the  $\phi'$ -gate is shared between the  $B_i$ - and  $B_j$ -nets. If however  $R_{i,2}$  is an input state,  $\psi$ , or is of the form  $\psi \oplus \rho_{i,2}$  then a reduction of the complexity with one negation gate will result, because both a  $\phi'$ -gate and a  $\psi'$ -gate may be shared. The two gates may be shared between the  $B_i$ - and  $B_j$ -nets, or one shared between the  $B_i$ - and  $B_j$ -nets and the other between the  $B_j$ - and  $B_k$ -nets. The last alternative is shown in Fig. 6 for the function  $B = x \oplus y$ . If the  $x \oplus y$  net had been triplicated, the complexity would have

been  $3 \cdot 5 = 15$  gates. We have now demonstrated that we can go below this complexity with one gate and still obtain single-error insensitivity.

For each further decomposition of a residue in a mod 2 sum of an input state and a residue, a gain of one gate is obtained. The minimum complexity of a single-error insensitive net for the most favorable function:

$$B = \sum_{i=0}^{\nu} x_i \quad (\text{addition mod } 2) \quad (35)$$

is thus  $(3 \cdot 5 \cdot \nu - \nu)$  gates. The minimum complexity for a net generating  $B$  without any error insensitivity is  $5 \cdot \nu$  gates. The

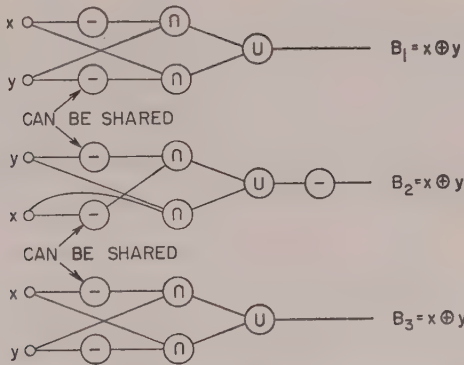


FIG. 6. Single-error insensitive  $g$ -net with indicated component-sharing possibility.

redundancy  $r$  (compare equation (22)) is thus  $9/14$ , which is less than  $r = 2/3$ , corresponding to a pure triplication.

Let us next turn over to  $w$ -error insensitivity in a  $g$ -net with one output. If here a gate is shared between two of the  $2w + 1$   $B_i$ -nets so that a single error in it will give an output error in both nets, then an additional  $(w - 1)$ -error can give output errors in  $w - 1$  other nets. The net is then not  $w$ -error insensitive. A sub-network  $\phi'$  can thus only be shared between two  $B_i$ -nets if an error in  $\phi'$  results in an error in either but not both of the  $B_i$ -nets. Hence, the most favorable function for a minimum complexity is the mod 2 sum of equation (35), as it was also for single-error insensitivity.



We have, for nets generating the function of equation (35), negation gates (which are necessary in order to make the sharing possible) as well as "and"- and "or"-gates. In order not to restrict ourselves to the specific choice of "and"- "or"-gates, let us denote by  $C_2$  the complexity of the subnet generating  $B(x, y, z, w) = xy \cup zw$ . The mod 2 sums are generated with such subnets together with negation gates.  $C_1$  is the complexity of a negation gate.

The minimum complexity of a single net generating  $B$  (equation (35)) is  $(2C_1 + C_2) \cdot \nu$ . In order to make the sharing possible,  $w + 1$  nets can have this complexity and the remaining  $w$  nets the complexity  $(2C_1 + C_2) \cdot \nu + C_1$ . The total complexity of the  $2w + 1$  nets is thus  $(2w + 1)(2C_1 + C_2) \cdot \nu + wC_1$ . We have here  $\nu + 1$  negation gates in each net, which can be shared with allowed negation gates in other nets. One allowed gate can be shared between only two nets. The above complexity can thus at most be reduced by  $(\nu + 1)wC_1$ , i.e. the minimum complexity will be

$$(2w + 1)(2C_1 + C_2)\nu - wC_1\nu.$$

Hence for any Boolean function, the maximum rate (compare equation (25)) will be bounded by:

$$R(w) \leq \frac{1}{\max (2w + 1) - w/(2 + C_2/C_1)}. \quad (36)$$

This bound is a lower upper bound. It contains the complexities  $C_1$  and  $C_2$ . By putting  $C_2 = 0$ , we obtain the following upper bound:

$$R(w) < \frac{1}{\max \frac{3}{2}w + 1} \quad (37)$$

which is independent of any specific choice of gates.

The minimum complexity (the minimum number of elementary gates) in a  $w$ -error insensitive net is thus larger than  $w$ . Hence, there must be a  $(w + 1)$ -error for which the net will give an output error, and arbitrary reliability is possible only if  $w \rightarrow \infty$ . But then the maximum rate of computation is zero, i.e. the capacity of computation is zero for any binary function generated by a  $g$ -net.

In the case of  $g$ -nets generating a system of binary functions, it is not possible to have a complete set of elementary gates which permit an arbitrary system of binary functions to be computed



with a nonzero capacity. For this statement it is sufficient to consider the system:

$$\left. \begin{array}{ll}
 B_1 = \sum_{i=0}^{\nu} x_i & \text{addition mod 2} \\
 B_2 = \sum_{i=0}^{\nu-1} x_i & \text{addition mod 2} \\
 | & \\
 | & \\
 | & \\
 B_\nu = \sum_{i=0}^1 x_i & \text{addition mod 2}
 \end{array} \right\} \quad (38)$$

The system is  $w$ -error insensitive if each  $B_i$  is  $w$ -error insensitive. We know that the maximum rate  $R_{\max}(w)$  of computing  $B_1$  alone is that of equation (36). And a net for  $B_1$ , with this rate, contains also all the outputs at which  $B_2, B_3, \dots, B_\nu$  are generated in a  $w$ -error-insensitive way. Hence the maximum rate of computation for the whole system is  $R_{\max}(w)$ . So we have in equation (38) an example of a system for which the capacity of computation must be zero.

Let us consider finally a  $g$ -net which is not binary. Suppose that each elementary gate of a functionally complete set have  $\nu$ -valued states:  $s_1, s_2, s_3, s_4, \dots, s_\nu$ . Let there be a probability  $p_{ij}$  ( $\neq 0$ ) associated with an error which transfers a correct state  $s_1$  to the incorrect state  $s_j$ . In the case  $\gamma = 2^\mu$ , we can associate with each gate (with one output and two inputs) a network of binary gates with one output bundle and two input bundles, each bundle having  $\mu$  lines. The  $2^\mu$  distinct states of a bundle correspond to the  $s_i$  states. With the probability  $p$  associated with an error in a single line of a gate-output bundle, the probability of a transition from one state to another will be within the bounds  $p(1-p)^{\mu-1}$  and  $p^\mu$ . If we choose  $p$  such that  $p(1-p)^{\mu-1}$  is equal to the smallest value of  $p_{ij}$  for the given  $\nu$ -valued gates, a redundancy-designed  $g$ -net of  $\nu$ -valued gates cannot be more reliable than a structurally equivalent binary  $g$ -net. If the  $\nu$ -valued net is designed with a minimal redundancy for  $w$ -error insensitivity, the corresponding redundancy structure for the binary net is not necessarily a minimum

design. So by changing this bundle structure to a minimum redundancy structure for the binary net, we can obtain an even better result with respect to minimal redundancy. But from our previous results we know that the capacity of computation for an arbitrary system of binary functions cannot be different from zero, and hence this is (even more) true also for  $\nu (= 2^u)$ -valued  $g$ -nets. Multivalued nets are treated in some detail in Ref. 2.

### 8. ERROR LOCATION COMPUTABILITY

We shall first investigate error-location computability for  $c$ -nets. The problem is: under what circumstances is it possible to design a  $c$ -net for the generation of a Boolean function, insensitive for  $W$ -errors, which also computes the location of  $w$ -errors without interrupting the normal computation of the Boolean function?

Let us take a simple example, a  $c$ -net for  $B(a) = a$  with  $W = w = 1$ . A single-error insensitive function for the identity operation is (cf. Ref. 7):

$$B = a_1a_2 \cup a_3a_4 \quad (39)$$

where  $a_i = a$  if no errors occur. In order to locate an error among these four components we must have at least five possibilities; no error, and an error in any one of the four components. This requires in the binary case three error-indication outputs. One system of error-indication functions is:

$$\left. \begin{aligned} B_1 &= a_1 + a_2 \\ B_2 &= a_2 + a_3 \\ B_3 &= a_3 + a_4 \end{aligned} \right\} \quad (40)$$

where  $+$  means addition modulus 2. If there is no error, all  $B_i$  will be zero independent of the value of  $a$ . All single errors will give different responses, still independent of the value of  $a$ .

However, it is impossible to construct a multi-terminal net with a 1-1 correspondence between its components and the variables  $a_i$  (compare the realizability conditions given in Ref. 8). Also we want the error-indication to be completely decoded. The system (40) gives information about where a single error ha

occurred, but in order to utilize this information for a component-replacement it has to be decoded, which means a further combinatorial computation. And we want all combinatorial computations performed with components which are uniformly exposed to noise.

If we change our desire to compute the location of an error, and ask only for an indication of a subset of components which contains the erroneous component, then and only then can we obtain a solution to both problems raised by the previous example, but only by a further restriction on the errors.

Let us consider one of the decoded error-indicating outputs. The functions  $B_i$  to be generated here can be specified by the truth-table:

	$a_1$	$a_2$	$a_3$	...	$a_r$	$B_i$
$\alpha$	1	1	1	...	1	0
$\beta$	0	0	0	...	0	0
$\epsilon$	1	0	1	...	0	1

The two variable combination numbers  $\alpha$  and  $\beta$  correspond to error-free cases for which  $B_i$  should be zero. Let the number  $\epsilon$  correspond to an error of weight less than or equal to  $w$ . Hence  $B_i(\epsilon) = 1$  ( $\epsilon$  is called an implicant number).  $\epsilon$  must contain at least one 1.

The number  $\alpha$ , which is not an implicant number, covers  $\epsilon$ , which is an implicant number. Hence (cf. Ref. 9)  $B_i$  must contain negations of the  $a_i$ -variables, i.e. the  $B_i$ -net must contain components corresponding to these negated variables. But this contradicts the idea of errors we have had so far. For if we have an  $a_i$ -component and an  $a_i'$ -component, an error in one of them means that we no longer have the relations  $a_i a_i' = 0$ ,  $a_i \cup a_i' = 1$ .

So in order to get a solution to our problem we must assume that the stationary errors in which we are interested do not affect the complementation relations between  $a_i$  and  $a_i'$ . If we work throughout with relays with transfer contacts, i.e. with both an  $a$ - and an  $a'$ -contact on the same relay such that the  $a$  and  $a'$ -contacts have one vertex in common, then this error restriction is not very serious. A stationary error usually has its source in the coil of the relay. For instance a short circuit in the coil would

affect both  $a$  and  $a'$  and hence be a permissible error. A dust-particle affecting only one of the contacts has, on the other hand, the character of a temporary error, which needs no repair, i.e. no error location. The prescribed redundancy design makes the net insensitive to such errors.

A transfer component of  $c$ -type has two normal states, either a transfer to the left or to the right. The errors we will consider are of type  $A_1$ , i.e. they will give states which are again contained in the set of normal states. An error on the left-transfer will be a right-transfer and vice versa.

Nets of transfer components satisfying the above truth-table specification, i.e. indicating any  $w$ -error, are of the types illustrated in Fig. 7.

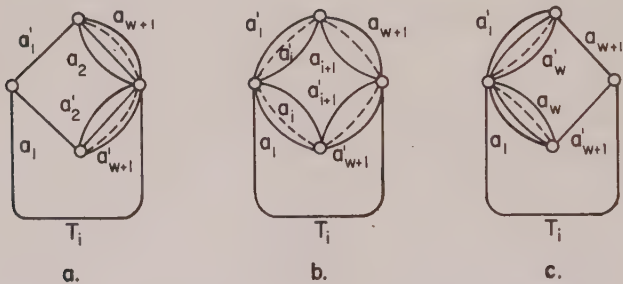


FIG. 7. Transfer-contact nets for  $w$ -error indication.

We shall now see that it is uniformly possible to construct for any Boolean function  $W$ -error-insensitive  $c$ -nets which also locate any  $w$ -error. We must here use separate redundancy protection of each component, i.e. maintain the basic structure of the net for  $W = 0$ . For the redundancy protection, each component of the basic structure is replaced by a net of length  $l$  and width  $v$  (cf. Ref. 12) for instance, according to Fig. 8(a) or 8(b) (where each branch represents only one of the two transfer-contacts, say, the affirmed). Each net is surely  $W$ -error insensitive if  $l \geq W + 1$  and  $v \geq W + 1$ .

The next step in the construction is to replace each of the components of the nets of Fig. 8 with a corresponding component of a net of Fig. 7, i.e. to imbue the error-locating nets in the basic

computing structure. The most efficient way of doing this, in the case of  $W = w$ , is to use  $(w+1)$ -error-locating nets of type Fig. 7(c) for a basic redundancy protection of the net of Fig. 8(b) with  $l = w+2$ ,  $v = w+1$  (see Fig. 9(a)). Notice that the desired

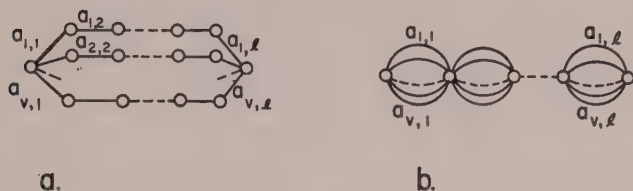


FIG. 8. Redundancy protection of the basic components for error-insensitive, error-locating nets.

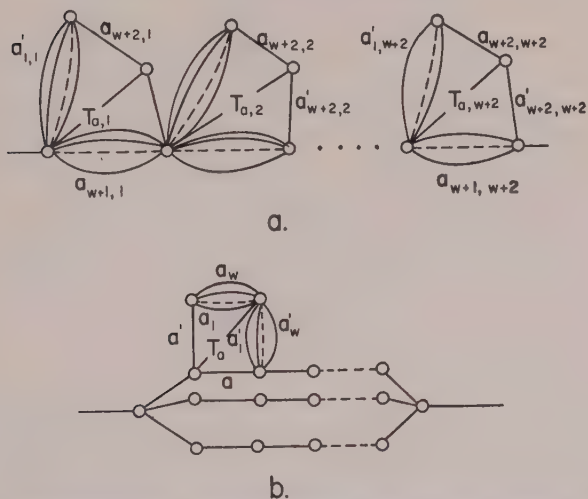


FIG. 9.  $w$ -error-insensitive,  $w$ -error-locating net for the identity  $B(a) = a$ . In Fig. 9b, only one locating net is shown.

properties of the locating nets of Fig. 7 have been established only when these nets operate alone. By adjoining a locating net over two vertices to the basic structure, we must be sure that no disturbing conducting path between the two vertices can arise outside the locating net. We see from Fig 9(a) that if  $a = 0$ , (the state  $S_2$  of equation 10), a conducting path outside the adjoining vertices

of any of the locating nets can occur only for a  $(w+1)$ -error (for example if  $w+1$  of the  $w+2$  branches,  $a_{w+1,i}$  are conducting). The measuring branches  $T_{a,i}$  are supposed to be nonconducting (cf. Section 9).

If  $M$  is the minimum number of single contacts in a network for  $B$ , this construction requires a total of  $M(w+2)^2$  transfer components. It locates a single error in a set of  $w+2$  components, and hence in the most severe case, a  $w$ -error in  $w(w+2)$  components. Since in the next section we want to let the case  $B_i = 1$  direct a replacement of all the  $w+2$  components of the corresponding subset, and still have an undisturbed normal computation during the replacement, we must use a larger amount of redundancy. With the construction of Fig. 9(a), a single error will be equivalent to a  $(w+1)$  error during the replacement operation.

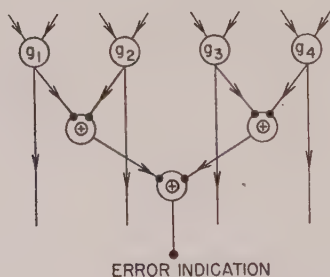


FIG. 10. Single-error locating  $g$ -net.

We will therefore use locating nets according to Fig. 7(a). They can be imbedded in a network of length  $l$  and width  $W+1$ , for instance, as in Fig. 9(b). Hence, the total number of transfer components will be  $M(W+1)(w+1)l$ , where  $l \geq (w+2)$ ,  $l \geq W+1$ .

Let us next investigate the error-location computability for  $g$ -nets. We will first look at a subnet which indicates, in decoded form, the presence of a single error in itself. Let the subnet contain four gates  $g_1, g_2, g_3, g_4$  of the ordinary net (Fig. 10). By the redundancy design in the ordinary net, the four gates should always have equal outputs in the case of no error. The subnet of Fig. 10 indicates any single error in itself (including a single error in any



of the (mod 2)-gates). A locating net operating on only two  $g_i$ -gates consists of one (mod 2)-gate. A gate operating on three  $g_i$ -gates consists of a different gate type.

A decoded location of single errors is the best we can do. Double errors are impossible to locate in decoded form. This is easily seen with reference to Fig. 11, showing the output gate of the

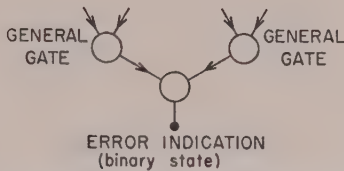


FIG. 11. Output part of a general error-locating  $g$ -net (for the proof of the impossibility of locating double-errors in decoded form).

indicating net and the immediately preceding gates. The output of the output gate is binary (no error or error). The immediately preceding gates do not necessarily have to be binary. A single error in one of the preceding gates must be indicated as an error at the output. But then a double error, consisting of an error in the output gate also, will be falsely indicated as no error. Hence the conclusion above.

## 9. ON THE CONCEPT OF SELF-REPAIR

When we repair a system displaying a fault, this usually means that we localize the source of the fault within a part of the system, and replace this part by a new one, or perhaps add a new part without taking the old one away. The parts may not necessarily mean physical objects, but may be relations as well. For instance a broken string may be repaired by tying a knot in the string, i.e. by re-establishing the connectivity relation.

We may look upon a self-repairing system from two points of view. A system of cells with fixed connections, designed with a large amount of redundancy, might be considered self-repairing in the sense that if some cells are acting improperly, the activity of the system may be transferred to other parts of the system,

previously at rest, so that a continuously correct behavior is established. Another example of a self-repairing system may have the ability to detect errors and identify improperly acting cells as to a subset of the set of all cells and, besides, have a mechanism for replacement of the cells of the indicated subset by fresh cells. We will deal mostly with this second type of system, mainly for the reason that we cannot avoid the question: Who repairs the self-repairing system? And it is evidently more in conformity with accepted ideas to have a watchman inspecting the reservoir of fresh cells, and eventually ordering new cells from the cell-producing department, than to put a large amount of effort in the constructing department and require it to build a huge fixed system containing as many cells as would correspond to the desired lifetime of the system.

First let us argue the necessity of some external repair of a self-repairing system, in connection with the first point of view.

We know that we can represent the dynamical action of a lumped physical system by a linear graph.<sup>(15)</sup> A replacement procedure means a transmission between a reservoir element and an (active) element taking part in the normal activity of the system. The transmission should be regulated by error-location indication in the active element. So a repairing action can be represented by a set of branches in the graph of the system, each having two states, a transmission state or a nontransmission state, in other words, with the same type of branches we have for the graph representing any binary combinatorial function. We see that we can deal with an eventual self-repairing combinatorial system as represented by a fixed graph of binary branches. This is so, provided that an error in a replacement is assumed to be no replacement and vice versa.

We shall distinguish between active components (being constantly actuated) and reservoir components which are not active (the activity or actuation of them is switched off by active components). For the probability of stationary error in the active components we make the same assumptions as before, namely that the independent probability of a component error is  $p$  for any actuation, provided that a previous error has been repaired. For the reservoir components which are not actuated, we assume that the probability of occurrence of an error is zero. (A certain fraction

of the reservoir components may be in an error state, however, provided this fraction is assumed to be constant.)

We shall now see that an unrestricted self-repairing binary combinatorial net for a nontrivial normal computation does not exist, provided that an error state of a branch (component) is within the set of its two normal states. It is sufficient to consider the necessary replacement structure around a single 1-error-indicating cell  $C_1$  and a similar reservoir cell  $C_2$  according to Fig. 12. The active cell  $C_1$  should, for the normal computation, be connected between the vertices  $A$  and  $B$  of the network. The actualizations of the components of cell  $C_2$  are normally switched off. Even

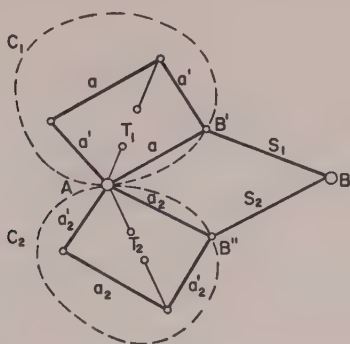


FIG. 12. Graph isomorphic to a replacement action.

if  $C_2$  contains two transfer components, the wires carrying the corresponding  $a_i$ -signals to it could have one wire in common. If this wire is cascaded with  $T_1$ , an error-free operation of  $C_1$  will ensure that  $C_2$  is in rest as a reservoir cell. But this disconnection is not enough.  $C_1$  and  $C_2$  are, for the normal computation, two-terminal devices. They may have one vertex (terminal),  $A$ , permanently connected, but the other terminals  $B'$  and  $B''$  must be connected to  $B$  over a transfer component  $S$ . The state of  $S$  should be  $S_1 = 1, S_2 = 0$  for  $T_1 = 0$  and  $S_1 = 0, S_2 = 1$  for  $T_1 = 1$ . It is evidently impossible also to imbed  $S$  (a transfer component with three terminals) in  $T_1$ . Hence we have a cause-and-effect relation to be realized as a transfer component  $S$ , again subject to error. It must therefore be made 1-error indicating and

concerning its connection (together with its reservoir mate) the same argument applies as for the connection of cell  $C_1$ . Hence no finite combinatorial binary switching net for a nontrivial normal computation can be self-repairing if the error states of the components are within the set of normal states.

However if we relax the condition that the normal computation shall be nontrivial, and consider a net designed only to "generate" a short circuit between its output terminals, then the arguments break down to trivialities. First of all, the components have then only one normal state: "transfer to the right" (or "conducting"). Hence an error defined as "transfer to the left" (or "nonconducting") is a state outside the set of normal states. The transfer component  $S(S_1, S_2)$  will not be necessary, for there is no need to switch off an erroneous cell which, by the error, becomes nonconducting. Also, the reservoir cell may in this case be permanently connected. The vertices  $B'$  and  $B''$  may coalesce to  $B$ . This might indicate the existence of a self-repairing system in this case. However since there are no actuations except when an error occurs, there is no need to distinguish between active cells and reservoir cells. The probability of error is zero when there are no actuations, and if all components are correct at a certain instant, they will be correct for all future times.

It seems fruitful to make the following distinctions. A system in which the error states of a component are contained in the set of normal states of the component ( $A_1$ -errors) is called a system of the first kind. If the error states are not contained in the set of normal states ( $A_2$ -errors) we will speak of a system of the second kind. In a repairing system of the second kind, no error-location computation is necessary. A self-tightening tubeless automobile tire or a self-healing electric fuse may be taken as examples.

In the first type of systems for nontrivial computation we have demonstrated the necessity of some kind of ideal behavior to make "self-repair" possible. We will, to begin with, assume that the replacement action is ideal. Then under the restriction that no combinatorial computation is performed by the replacement mechanism, we have, in a sense, a motivation for the notion of "self-repair". Notice that here the location of an error must be computed in a completely decoded form. This means that if there are  $N$  positions for the cells in the active region, there must be  $N$

error-indication outputs with a 1-1 correspondence to the cell positions.

Looking at a system (of the first kind) from the second point of view (with an actual physical replacement of the cells), we may imagine it as follows. A cell consists of  $w + 1$  transfer components internally connected according to Fig. 7(c). All wires supplying the cell with actuation signals, and all its output-terminal and error-indication wires, are drawn to contacts on a side of the cell. The reservoir cells are floating around in a lake. At the bottom of the lake there are a number of positions or holes in which the cells just fit. The stream of liquid (with about the same specific weight as that of the cells) through the holes will cause one cell to be positioned in each hole. In position, the components of the cell make proper contacts with a wiring layer built into the bottom of the lake according to Fig. 9(b). If there is an error in a cell, there will be a contact between its  $T$ -terminals. This contact will generate a current which, by a thermal phenomenon, will release the cell so that it glides out of the hole.

Again the liquid streaming out of the hole will fill it with a new cell.

The only remaining question that need be answered, before we will have demonstrated the existence of a self-repairing system of the first kind is the following. In the previous section we assumed that we could measure the  $T_i$ -states so that the measurement in itself did not affect the state. Is it actually possible to measure simultaneously the  $T_i$ -states in a multi-terminal network so that one measurement does not influence another? The answer is yes, but the solution is a little bit tricky. First of all, it is not possible to insert a battery between each pair of  $T_i$ -terminals, because it has low resistance and one measurement would influence another. Eventually one could insert the batteries between a succession of  $T_i$ -terminals with the polarities reversed, but one would then have to insure that all batteries be alike. The method illustrated in Fig. 13 is based upon a measurement with an a-c current of a frequency which is high compared to the actuation frequency. The capacitors  $C$  short-circuit the measuring current but act as open-circuits for the measurement of the total state of the net (the normal output) at any actuation. If the  $i$ th cell is in an error state a high-frequency current will flow through the resistor  $R$  and by



thermal action release the cell. Notice that the turn-around of every other cell is necessary. Otherwise the normal state  $a = 0$ ,  $a' = 1$  would falsely release every cell. Now under very specific circumstances, it can happen that an error in the  $i$ th cell will also

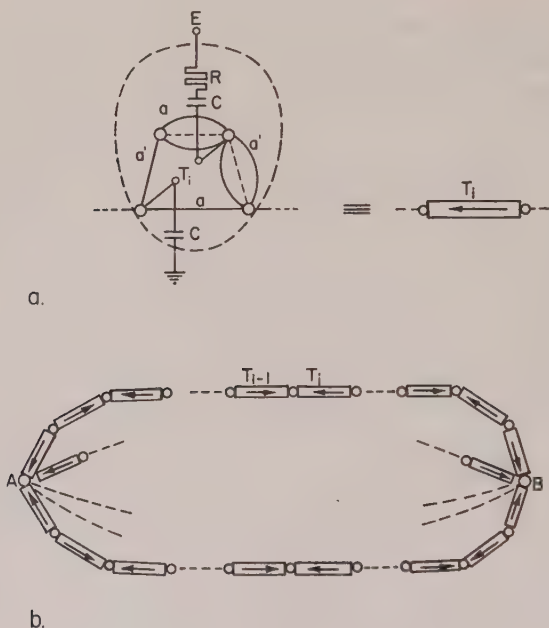


FIG. 13.  
a. Cell in a self-repairing  $c$ -system of the first kind.  
b. Cell connections.

cause a release of one, but only one, of its neighbor cells. If the normal state is  $a = 0$ ,  $a' = 1$ , an error in an  $a$ -,  $a'$ -transfer (Fig. 9b) in cell  $T_i$  will cause cell  $T_{i+1}$  (Fig. 13b) also to be released ( $i$  an odd number). However, even if a  $w$ -error can release  $2w$  cells this does not interrupt the normal computation, provided that the release of one cell does not release a cell in another path of the redundancy structure of Fig. 9(b). This condition is secured by the cell-connection illustrated in Fig. 13, which requires an even number of cells in each of the  $W+1$  paths of Fig. 9(b).



The threshold of the thermal release mechanism should be chosen so that short error current pulses which may arise during the transition between two actuations (the transfer in one component may occur just before the transfer in another component) will cause no release. Also, when a cell is positioned, it may happen that some components in it will obtain actuation signals from the wiring net before other components. A false error current can be prevented either by a mechanical construction which ensures that the contact to the high frequency energy supply always is the last established contact, or by a sufficiently large threshold in the release mechanism.

Let us for a summarizing discussion of the probability of a total failure use the notations:

$t_1$ : cell release time.

$t_2$ : average cell release-replacement time.

( $w$ : error weight for which the location computation is correct.)

( $W$ : error weight for which the normal computation is correct.)

Suppose that the occurrences of errors are not synchronized with the actuation of the components. Let us further assume that all components are ideal, in the sense that their switching time can be made zero. Then we can obtain arbitrary reliability or arbitrarily long lifetime if we also make  $t_1$  go to zero, but in such a way that, in the limit,  $t_1$  is sufficiently larger than the switching time to prevent false releases during normal transfer changes. Then we can safely choose  $w = 1$  and still be sure that the error location is correct. The probability of a double error during the time  $t_1$  is zero in the limit  $t_1 = 0$  (the error currents are assumed not to be synchronized). Hence with a finite  $t_2$  of the same order of magnitude as the actuation time, we have reduced the problem to the temporary case. Each cell now has, however, the equivalent probability  $2p$  of being in an error state, where  $p$  is the probability of error in each transfer component. So with  $2p < \frac{1}{2}$  we can reach arbitrary reliability.

A more realistic assumption is to consider the error events to be synchronized with the normal actuations. Let us further assume that  $t_1$  is large compared with the switching time but small compared with  $t_2$ . If  $t_2$  is of the same magnitude as the actuation time, we can roughly estimate the total probability of error as follows.

The probability  $q^*(\nu)$  that the cells, appearing in a certain position (hole), have no  $(w+1)$ -error during  $\nu$  actuations is:

$$q^*(\nu) = (1-p^{w+1})^\nu. \quad (41)$$

The probability  $p^*(\nu)$  that an error preventing cell replacement occurs during  $\nu$  actuations is thus:

$$p^*(\nu) = 1 - q^*(\nu) \sim \nu p^{w+1}. \quad (42)$$

If, for a moment, we neglect cell errors of weight  $\leq w$ , which, due to the replacement mechanism, occur as temporary errors, we obtain for the probability of total error during  $\nu$  actuations:

$$P(\nu) \sim M_1(\nu p^{w+1})^{w+1} \quad (43)$$

(cf. equations (17) and (21)). Hence an upper bound on the lifetime  $\bar{\nu}$  of the whole self-repairing system is:

$$\bar{\nu} < \frac{1}{2} \cdot 1/p^{w+1}. \quad (44)$$

However, coming back to the temporary cell-replacement events, the probability  $q^*$  that a repairable cell will perform its normal computation task during an actuation is:

$$q^* = (1-p)^{w+1}. \quad (45)$$

Hence, the probability  $p^*$  that the cell is released at an actuation is:

$$p^* = 1 - q^* - p^{w+1} \sim (w+1)p. \quad (46)$$

Since we have assumed that the replacement time is of the same magnitude as the actuation time, the replacement has the same effect as a temporary error occurring with probability  $p^*$ . So in order to obtain any reliability at all, we must (cf. equation (18)) require that:

$$(w+1)p < \frac{1}{2}. \quad (47)$$

The bounds of equations (47) and (44) imply that we cannot obtain arbitrarily long lifetime in self-repairing systems in the case where error events are synchronized with the actuations of the system. For a fixed  $p$ ,  $w$  is bounded (47) and hence also the lifetime (44). For small values of  $p$ , however, we can obtain lifetimes which are very much larger than the lifetime of a single component ( $\bar{\nu} \approx \frac{1}{2}p$ ; cf. Section 5).

A fairly realistic value of the magnitude of the lifetime of a switching component is  $10^7$  actuations. Let us consider a combinatorial net of  $N = 1000$  components, each with  $p = 10^{-7}$ . We may expect that the net will operate without error during  $\nu_0 = 5000$  actuations only ( $N\nu_0p = \frac{1}{2}$ ). Let it be required to design a self-repairing system for the same normal computation, but with a lifetime of about  $10^9$  actuations (greater than the lifetime of a single component). Let us choose  $w = 1$  and  $W = 1$ . The system will contain  $N(W+1)^2 = 4000$  cells, each with  $w+1 = 2$  transfer components.

The probability that a cell-replacement preventing  $(w+1)$ -error occurs at an actuation is  $p^{w+1}$ . So after  $10^9$  actuations there will be about  $10^9(W+1)^2 \cdot Np^{w+1} = 0.04 < 1$  cells in the whole system for which the replacement mechanism does not work. Hence it is sufficient only to consider the errors due to normal replacements, which are temporary in character:

$$P_{sr}(\nu) = M_1 \cdot \nu \cdot [(w+1)p]^{W+1}$$

(cf. equations (46) and (20)). For  $W = 1$ , we have  $M_1 = 4N$ , and  $P_{sr}(10^9) = 0.16 < \frac{1}{2}$ , indicating that the lifetime is in the desired range of  $10^9$  actuations.

The bound of equation (47) allows us, in this example, to increase  $w$  far above 1, and by also increasing  $W$  we can indeed reach lifetimes of astronomic magnitude!

We have found that the maximum lifetime  $T(t)$  of a self-repairing  $c$ -net system of components, each with a lifetime  $t$  ( $= \frac{1}{2}p$ ), is bounded if  $t$  is finite. A crude upper bound of  $T(t)$  is obtained from equations (44) and (47):

$$T(t) < \frac{1}{2}(2t)^t. \quad (48)$$

We have previously assumed that the replacement mechanism is ideal, in the sense that if an error-indication signal is obtained from a cell, the cell is replaced. We need not make such a strong assumption. If a cell is released upon actuation, it obtains only one release actuation. Hence the release-actuation function can be made as reliable as we want with ordinary redundancy design in the release mechanism in each cell.

## 10. CONCLUSIONS

We have found that the maximum lifetime  $T(t)$  of a self-repairing  $c$ -system of transfer-components (each with a lifetime  $t$ ), is in general larger than  $t$ , although limited. This might suggest a super-iteration leading to a self-repairing system whose components are self-repairing systems. Such an iteration is however not possible, for we have required that the components (the self-repairing subsystems) must be of transfer type, such that an error in a specific transfer results in the opposite transfer. This requirement cannot be met except by elementary components.

A corresponding argument prevents super-iteration for a self-repairing  $g$ -net. We have found that there can be at most single-error-locating  $g$ -nets. A self-repairing  $g$ -net therefore has a maximum lifetime  $T(t)$  bounded by  $t^2$ , i.e. in general a much shorter lifetime than  $T_c(t)$  of equation (48). A reliable  $g$ -net must have (for  $B_1$ -errors) an output represented as a bundle-state (a non-decoded output). Hence, if, for a super-iteration, we want to indicate an error in an output bundle-state of a subsystem with a locating  $g$ -net that is part of the subsystem, there must necessarily be an error indication for an error in a single line of the bundle, which is no error in the bundle-state as such (in fact, that is why we use the bundle-representation).

In summary, we have defined errors as deviations from an error-free behavior which we have hypothesized as deterministic. We have defined deterministic behavior such that there is a reasonable chance for it to be displayed by physical systems, and still be consistent with some common aspects of determinism. A classification of errors has been obtained by denial of the definition of determinism. We have arranged component errors in the following sequence: (a) temporary errors distributed over part of the system, (b) temporary errors distributed all over the system, (c) stationary transfer errors distributed over all active parts of the system ( $c$ -net) including the release mechanism, (d) stationary transfer errors distributed over all active parts of the system ( $c$ -net) including the whole replacement mechanism, (e) stationary errors distributed over all active parts of a  $g$ -system including the release mechanism. In this sequence, perfect reliability is possible only for (a) and (b), for (a) with a redundancy less than 1 (the communication

case), for (b) with a redundancy equal to 1 (the computation case). For (c) and (e) only a finite increase of the over-all lifetime of the system is possible. For (d) (cf. Fig. 12) no uniform error correction is possible (a certain increase of the lifetime is possible, however, even in this case, because the probability  $p$  of an error in a component is defined with respect to the actuations of the component, and the replacement actuations are less frequent than the normal actuations).

The stationary errors are the most realistic errors. Hence the result that a system exposed to errors of this kind can have at most a finite lifetime is somewhat of a disappointment. To go beyond this limit, it is necessary to require that some components be ideal, and in reality this means that we let the system interact with ourselves (beyond what corresponds to our desired use of the system).

The question of how to minimize this unwanted interaction is indeed a challenging problem, especially when we extend the interaction to include also the phase of the construction of our systems (machines).

If we relax the requirement that the deterministic system be well localized (has specified inputs and outputs) there are indications that we can go beyond the given lifetime limits if the self-repairing system also has self-reproducing properties. In fact, such a system would tie together the concepts of self-repair and self-reproduction, a connection which was anticipated in Ref. 6.

Let us complete, but not finish, our conclusions by quoting Schiller (Gedichte: *Kassandra*) in the following translation:

“In Errors Only Is There Life,  
and Knowledge Death Must Be.”

## 11. ACKNOWLEDGMENT

The author wishes to acknowledge stimulating discussions on parts of the subject with Professor P. Elias of MIT and with Drs. E. Moore, D. Hagelbarger and T. Crowley of Bell Telephone Laboratories. In particular, the net of Fig. 4 was derived during a discussion with the last three gentlemen. Stimulating discussions with participants of this Symposium are also gratefully acknowledged.



## REFERENCES

1. S. AMAREL, An approach to automatic theory formation. This vol., p. 443.
2. J. COWAN, Many-valued logics and reliable automata. This vol., p. 135.
3. P. ELIAS, Computation in the presence of noise. *IBM J. Res. Devel.* **2**, p. 346 (1958).
4. F. FIRESTONE, A new analogy between mechanical and electrical systems. *J. Acoust. Soc. Amer.* **4**, p. 249 (1932-3).
5. D. HUFFMAN, The synthesis of sequential switching circuits. *J. Franklin Inst.*, pt I, **257**, p. 161 (1954); pt II, p. 275 (1954).
6. L. LÖFGREN, Automata of high complexity and methods of increasing their reliability by redundancy. *Information and Control* **1**, p. 127 (1958).
7. L. LÖFGREN, Irredundant and redundant Boolean branch-networks. *Trans. IRE IT-5*, Special Supplement, p. 158 (May, 1959).
8. L. LÖFGREN, Solution to the realizability problem for irredundant Boolean branch-networks. *J. Franklin Inst.* **268**, p. 352 (1959).
9. L. LÖFGREN, Redundancy bounds for  $w$ -error correcting contact networks for a Boolean function. *Quarterly Progress Report MIT, RLE*, p. 119 (October, 1959).
10. L. LÖFGREN, The structures of switching nets. To be presented at the Fifth Midwest Symposium on Circuit Theory, University of Illinois, May, 1961 (submitted for publication to *Trans. IRE EC*).
11. W. MCCULLOCH, Agathe Tyche of nervous nets—the lucky reckoners. *Mechanization of Thought Processes* (Natl. Phys. Lab. Symposium No. 10) H.M.S.O., London (1959).
12. E. MOORE and C. SHANNON, Reliable circuits using less reliable relays. *J. Franklin Inst.* **262**, p. 191 (1956); Part II **262**, p. 281.
13. D. MULLER and S. BARTKY, A theory of asynchronous circuits. *Proc. International Symp. on the Theory of Switching*, Harvard University Press (1959).
14. C. SHANNON, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, p. 379 (1948).
15. H. TRENT, Isomorphisms between oriented linear graphs and lumped physical systems. *J. Acoust. Soc. Amer.* **27**, p. 500 (1955).
16. A. TURING. On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.* **42**, p. 230 (1936).
17. J. VON NEUMANN, Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata Studies* (eds. C. SHANNON and J. MCCARTHY), Princeton Univ. Press, Princeton, New Jersey (1956).



## GORDON PASK

*Systems Research, Ltd.*

### A PROPOSED EVOLUTIONARY MODEL

Ladies and gentlemen, I hope you will bear with me if I give an informal talk this evening, because I would like to compress four different things into as short a time as possible.

In the first place, I would like to point out why it is we may be impelled to regard some systems as though there were elements in them that made decisions, because I so believe that if, and only if, we do this can we call them self-organizing systems.

Second, I would like to outline very briefly indeed, a general application of the sort of model we get, if we do look at systems just like this.

Third, I would like to go into a more detailed model, though not in detail, and here you must pardon me completely, please, because I am not going to put the arithmetic of it on the board, and I hope you will take it that I can add numbers together, though in fact I cannot; my equations always make something different, but it works.

Finally, I should like to have the presumption to make a few odd comments intended to tie together some loose ends and to establish the community of the subject. I don't hope to do very much here, but I think there are one or two things which could be said in the light of models of this kind, which tend to unify the different approaches which we have heard about today, and no doubt will hear about in the next day.

The first issue then, is the question: why do we think about systems as though they contain decision-making elements?

Now, we are self-organizing systems and we wander around in a world which is full of wonderful black boxes, Dr. Ashby's black boxes. Some of them are turtles; some are turtledoves; some are mocking birds; some of them go "Poop!" and some go "Pop!"; some are computers; this sort of thing. Now these things we tend to categorize in odd ways.

Some black boxes, I go up to and say, "This thing is a chance machine". What do I mean by this? I mean precisely that I know just what sort of inquiry I want to make, being a self-organizing system, about this thing.

I know that it is a chance wheel, it is a roulette wheel, it has got certain positions where it can stop, and I know—I call it a chance machine because I know this—that if I observe it for a long time, I cannot tell where it will stop next. In this sense, I am uncertain about what will happen, but I am not at all uncertain about what sort of things I ought to look at.

Now, again we go up to the poor computer and we say, "Ah, you are a determinate machine", and so he is; we can take him to pieces and find out what happens inside him, and if we think about noise, this is a thing we agree to exclude because it doesn't refer to the kind of question we want to ask about computers at all.

But finally there is a nasty little class of systems that I think are the ones we call self-organizing systems, which includes you, gentlemen. I go up to you and I have a conversation with you. Now, of course, you are an awfully random thing, because you burble out words. On the other hand, if I can establish a conversation with you, this is no longer the case. Why is it no longer the case? Because, of course, I am uncertain about what you will say next, in the same sense that I am uncertain about where the chance wheel will arrive. But, my main uncertainty about you is of a different sort, it's an uncertainty about what sort of inquiries I should make.

Now, it may be the case that this defeats me altogether, and I cannot talk to you at all. On the other hand, it may be that I can so adapt myself as a self-organizing system, to put it mathematically, I can so change my representation, change the sort of inquiries I wish to make about you, that I can make sense of you.

In other words, it is a deliberate expedient, because, for some reason or other, it would be useful if I talked to you. I adopt this particular procedure of changing the representation in order that this consistency in the behavior of the system, which we can express in all kinds of manners by saying there is a group property in the transformatin of the system itself, shall come about.

Now, systems of this kind we tend to call self-organizing systems. When it happens that we *must* adopt this expedient,

whether we like it or not, they are self-organizing systems. But, it is a confusing and blurred one, because there are many cases where we can adopt this expedient if we want to, but need not do so. I don't mind which case we have got. The absolute distinction can be made, I think, on the basis of the Gabor-McKay theory. You will notice that if I had been talking in terms of Shannon information and Shannon communication theory, I would not have made one of the distinctions I have made. I make it only because I separate the metrical and the logical aspects of information. I talk about logons and metrons separately. And I think that this is an important distinction in the present case. And it will be discussed much more ably than I can discuss it, I believe, later. It has already been discussed, incidentally, by Cowan here, and I believe he is going to make some more comments on this subject.

Now there are some funny things about these systems. For one thing, if we say of them that they learn, we cannot really distinguish that statement, because of the peculiar mixup of structural and metrical information, from a statement that they evolve, and this evening I am going to be particularly concerned with the evolution of an apparatus which, in a particular stationary condition of the system, we may then say is a learning mechanism.

In other words, taking the body of this afternoon's talk, the networks which McCulloch's group and Jerry Lettvin, at the physiological level, talk about as logical filters, I take as structures which we can understand *if* we find them. The sort of model I am going to discuss now, doesn't refer in the least bit to how they work, it only refers to how they shall come about in a system which is initially unstructured or moderately structured.

In doing this, I think I assume above all that we drink "Beer"—the pun is no worse than "Torus".

BEER: Touché!

PASK: Now, the kind of system we do have, when we do talk about it as a self-organizing system, is a system in which the elementary particles we are dealing with are not the elementary particles with which a physicist will commonly deal. These are replaced with unitary elements which may be considered to be automata, players, decision makers, "neurons" or the like. They can go "Poop!" and send a signal to another, the implication of

the signal being that the state of some remote element is changed by the fact that this one goes "Poop!"

In order that they shall go "Poop!" they must feed. I do not mind how you represent the feeding, but it is important that they do feed. It is a condition on the measurability of the system rather than on its energetics, but it is convenient to present this for explanation as though it were food or energy coming into the system that is required in order that the signalling activity take place. It is a measure on the system in the sense that this is the way we are going to talk about and find out about the state of the system. It is a conservable quantity.

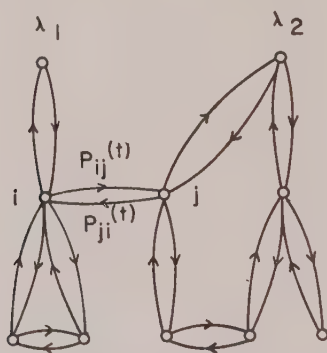


FIG. 1. Formal representation of food distribution network.

We consider that food enters a system and goes through a diffusion network. All I mean by a diffusion network is a system of tubes and basins, say, over which we can define food neighborhoods. A formal representation of a food diffusion network is shown in Fig. 1. It is a directed graph with nodes. The lines connecting nodes have quantities associated with them that represent the food impedance, the amount of resistance to the passage of food between nodes. The nodes at the top represent food sources; those at the bottom represent the nodes accessible to our elements, at which they feed. Such a formal representation is only to insure that we can define food neighborhoods, so that if one element feeds at a node we can say that it will deplete the food available

to another element at a neighboring node, but will not affect the amount available to a more distant zone.

Now in addition to our elements and a food distribution network, we require a material in which the signalling of the elements builds and maintains signal pathways through the expenditure of food. Hence there is a signalwise connectivity among the elements; erected on the stage of every signalling activity is a structure which is determined, made, and maintained by the expenditure of food. This structure cannot exist and persist independently of the activity of the system, it comes into existence as a result of this activity and is maintained by it. Turn the system off, and it all disappears.

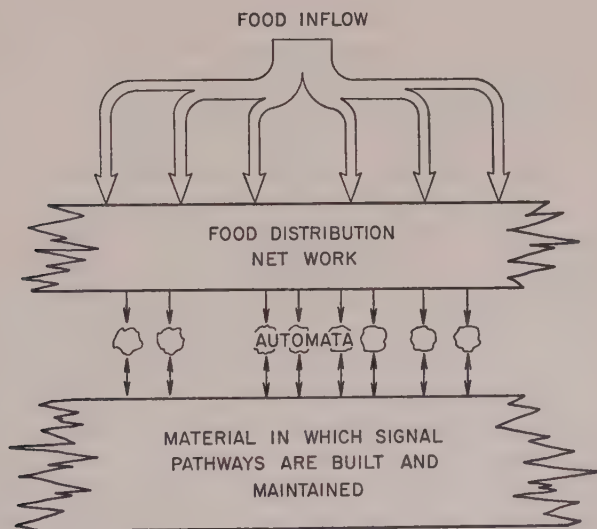


FIG. 2. Diagrams of model system.

Figure 2 is a diagram of the whole system. The important thing to note about such a system is that, as a result of the signalling activity of the elements, one builds up the patterns of structure by which the elements become connected. Those elements which become connected can then be in a position to correlate their strategies.



In particular, I am interested in those connections such that two elements can signal reciprocally. Such a connection would be represented by a cycle in the graph of the connection matrix. In a situation such as this, it is possible for two elements to correlate their strategies completely.

It is important to remember, because it is crucial to the next stage of my development, that we have decided, for reasons based only on our selfish considerations, only on efficiency or utility to our own observations, to call these elements "decision makers". We do not need to inquire what in fact is inside these elements. This is irrelevant to the matter.

Hence if I call such an element a player, and then go on to the alarming statement that such a cyclic connectivity can be a coalition, you will perhaps not take exception to me. The system is closely coupled, and if I regard the elements as players, I will call this a structure which can permit coalition. Please note that a coalition in this system is something which must have a structure associated with it, because it is nonsensical to talk about these things correlating their strategies if they cannot communicate, and they can communicate only if they establish and maintain a communication structure by the expenditure of food. There is a communication cost implicit in their bringing themselves into the same signal neighborhood.

Now, the point I want to make at this level, by considering this first very general model, and before we go on to the more accurate one, is the fact that we can derive some interesting conclusions about what coalition structures can exist for a given payoff function in the food network.

If the payoff function which we derive from the food network, the diffusion network, is the payoff function of a competitive game, so that no advantage is gained by the players cooperating, it is nonsensical to have the idea that this can form a self-organizing system. The self-organizing system is something which occurs when cooperative activity is favored or, to put it concisely, when the payoff function determines an essential game. When it does, there is still a restriction upon the coalition structures which can form; the restriction is introduced by any sensible assignment of the cost of maintaining these coalition structures.

For example, we have, in Fig. 3, examples of coalition structure



involving the coalitions of  $\alpha$ ,  $\beta$  and  $\gamma$ . The second one involves a cycle, a single cycle, and it has the same maintenance cost as four independent elements. The remaining are also structures able to realize these coalitions, plus others, and they cost more.

Looking at Fig. 3 in a little more detail, we can see that there are different maintenance costs along here, assignable to different coalitions of linear elements, for example, there are the trivial coalitions of  $\alpha$  and  $\beta$  alone, for,  $N$ , the number of players, equal to two. This is the only cost for the coalition of  $\alpha$  and  $\beta$ , it is a unique affair.

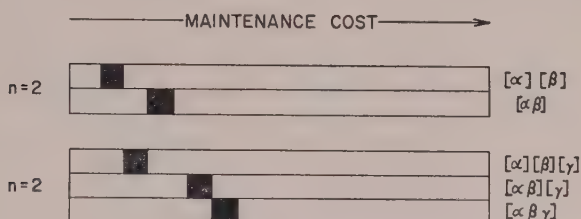


FIG. 3. Maintenance costs of coalitions.

In the case of coalitions of three, we have a cost for  $\alpha$ ,  $\beta$  and  $\gamma$  separately per player. We have a cost for  $\alpha$ ,  $\beta$  and  $\gamma$  separately, and we have a cost for the  $\alpha$ ,  $\beta$ ,  $\gamma$  coalitions of which there are two, and so on for four and five players.

In other words, there are discrete maintenance cost levels which can be realized in this system. We may also add to this further constraints comparable to Luce's Psi function, which determine what possible coalition structures can emerge.

Now, a number of quite intriguing things occur when we consider how, when looking at a system like this, we might be able to make sharp, well-defined observations upon the state of the system. It is intriguing, for example, to plot different levels of maintenance costs at which coalition structures can occur (Fig. 4), and to consider what will happen if we make the surplus amount of food available to the system  $U^*$ , decrease, we can plot on the same time coordinate the probability  $P$  that a given observed structure, that is to say, a given connectivity  $F$ , equals a coalition structure,  $c$ ; in other words, the probability that a given observed structural entity mediates a particular function.

Clearly if there is a level at which, say, coalition  $X$  can occur, then when we get down to the cost at which it just can occur, if it exists at all, then it is certainly being used, for otherwise it would collapse.

Hence, we have a local maximum in  $P$ . As we decrease  $U^*$ , we know that  $X$  cannot exist but  $Y$  perhaps can. Supposing  $Y$  does exist, we will have another maximum when we reach that point. In general, we will get a curve of this kind, as we decrease  $U^*$ , the local maxima of  $P$  will be the sharp-valued observations, the points at which we can make definite statements about the system.

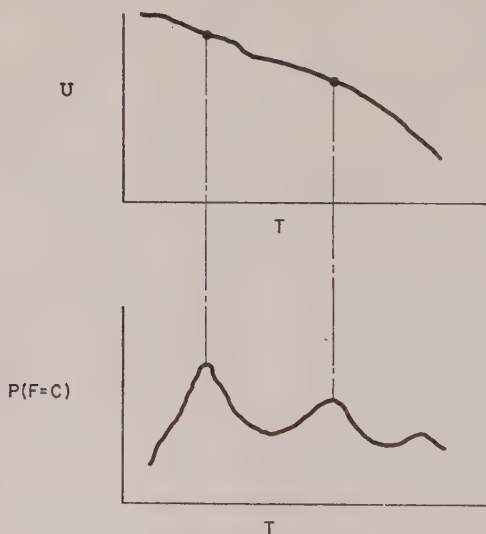


FIG. 4. Observation of coalitions.

This phenomenon is not unobservable in sociology. It is confirmed by anybody who plays around with the sort of network which, for example, Paul Weston makes with ionic resistors and neons; as you decrease the standing potential the thing locks into a stable state.

But this perhaps is a trivial case. The more important cases are in those systems, say, like your systems, Dr. Beurle, where you have critical values at which certain forms, certain modes can be maintained.

Now this is all I really want to say about the very general model, because I think that a certain clarity comes into the discussion when we consider the most primitive possible model we can conceive.

Now, for this purpose I have taken Edwin Abbott's "Flatland" as a universe. If you remember, Edwin Abbott wrote a little scientific fantasy about dimension. In this he was a two-dimensional figure, a square. There was a hierarchy of these figures, ranked according to the number of the sides they had, and they were supposed to have evolved, in some distant time, from one-dimensional creatures, and in fact he made a journey to "Line-land". Also it was possible that there were figures—indeed there was a missionary sphere—which lived in three-space. It is on this sort of format that I made my little model, and you must pardon its simplicity, which comes largely from having no computer and only a desk calculating machine, and a gentleman to work it for me.

What we did was to suppose a food distribution network in which the subset of nodes that are accessible to our automata for feeding lie in a plane. This accessible subset I call the field. It is obvious that we might apply other geometric constraints on the field, have it be the surface of a sphere or a torus, with consequent changes in the kinds of communication structures possible.

Now we will think of primitive little zero-dimensional automata, of which, since we are at least two-dimensional, we can distinguish two species: those that move longitudinally and those that move latitudinally.

But what does an automaton do? Because I drink "Beer", I have defined an automaton as that which is designed to survive in a specified universe. This means that the automaton itself is going to be subject to the same conditions that affect the coalition structures we spoke about. The automaton is not something that exists and persists in its own right; it has to pay for its creation and existence. I suppose its creation to occur by a process I call nucleation. If the food stored at any accessible node accumulates to a critical value, an automaton appears there.

An automaton, when it appears, sucks in food and builds this food into a communication structure, which structure is subject to degradation and must be maintained. The rate of feeding depends

on the local concentration of food; the rate at which it feeds depends on how much food there is.

Here I want to point out an important thing about these automata, about any automaton which can be said to be designed to survive, designed to compete. To make my point, I will distinguish between two classes of automata. The first class includes those we most often come across. They are things which are able to make decisions, moves, signals or whatever. They do so on the basis of accumulating evidence about the activities of other automata and possibly about conditions in their environment engendered by changes other than the activities of their fellows. This evidence is signified by a vector of some sort, and the values of this vector are piped into a decision rule, the output of which is a move, a signal, or whatever.

It is conceivable, and note that it must be conceivable in the universe in which we define competition as the thing we are looking at, that such an automaton can encounter a situation which is undecidable. In such a case, automata of this class, the sort we encounter in "hill-climbing" devices, are given a fresh strategy from outside. They call for independent information from outside, so that they are no longer a closed system. They are invaded, as it were, by a wheel of chance or a table of independent numbers. They just ask for a number, and this decides their undecidable decision for them.

Now this is not the sort of automaton I have in my model. I have one of the other class, which, when presented with this same dilemma, either evolves, or dies. If it has enough substance, it evolves; if it does not, it's had it!

The manner of evolution can be expressed rather precisely in the language used either by Dr. Ashby or by Dr. Rashevsky, one of them in terms of states, the other in terms of biological functions and the graphs of these. The results of evolving will always be that the automaton which found a certain situation undecidable now becomes a larger, more complex automaton which can comprehend a larger world in which the situation may not be undecidable.

In our trivial little universe of these creatures moving around, we have not given a great many facilities to our evolving automata. In Fig. 5 I have shown the possible moves of the primitive automata.

As I mentioned, there are two species, a and b, capable of one-dimensional latitudinal and longitudinal moves, respectively. Each of these primitive automata is capable of just three moves, either up, down, or stay put, or left, right, or stay put. If he stays where he is he sucks up all the food and dies.

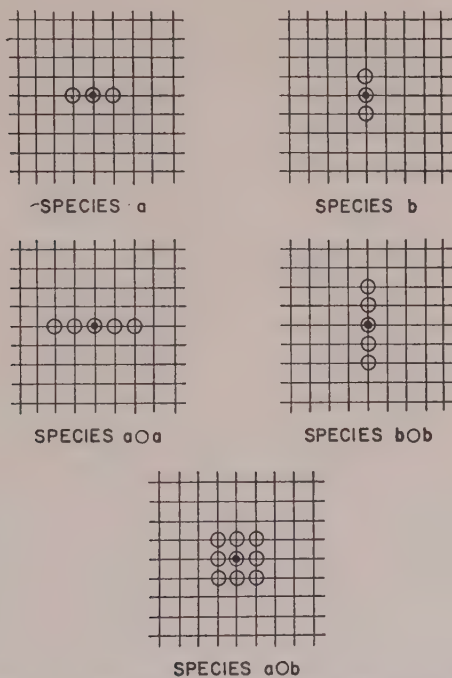


FIG. 5. Moves of primitive and evolved automata.

The evolved automata, of which a few are also shown in Fig. 5, are evolved by composing two or more of the primitive automata. If we compose two of the same species, we get beasts capable of five moves instead of three. More viable are the creatures obtained by composing one from each of the two primitive species. This creature is capable of nine moves, and has the advantage of having a two-dimensional move neighborhood. These compositions can be continued indefinitely, with each new species having all the moves of its predecessors. Although the move neighborhoods can



never be other than squares or rectangles, the individual moves may be quite peculiar.

The evolutionary rule is exercised when automata get into difficulty, and when, having got into difficulty, they come together. The difficulties are engendered by the characteristics of the surroundings, the distribution of automata in the field and the food supply at the accessible nodes of the food distribution network.

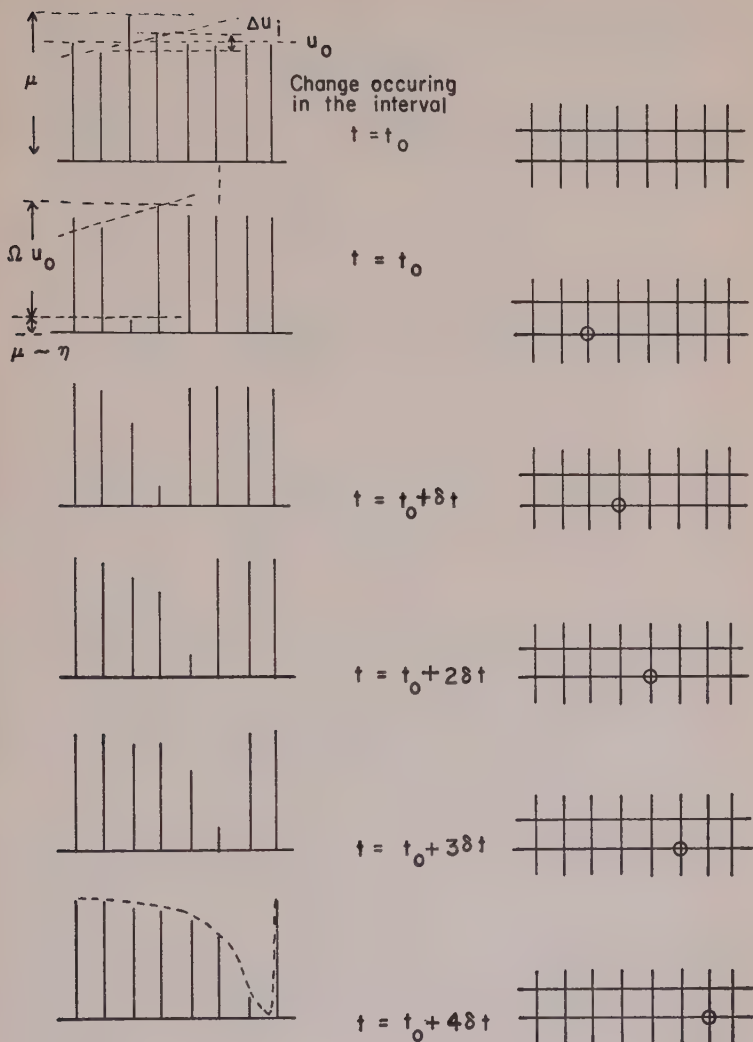
I would now like to talk about the characteristics which we impose upon the food distribution network in order to be sure that out of this model there arises something which, when embodied in a fabric, but not on its own, would be a self-organizing system. The constraints are simply as follows; we say that when an automaton is given play on the accessible nodes of a food distribution network, the food impedance, that is, the impedance between the accessible node at which the automaton feeds and all inaccessible nodes with which that node is connected, is a function of its eating and of time. We make the weight of the connectivity of an accessible node to other nodes change as a function of feeding activity at that node.

Now, I would like to illustrate some of the tricks which these automata get up to when you actually run this model. Consider an automaton of this sort in a plane and consider just a line in this plane (Fig. 6). The food concentration at the nodes along this line are plotted on the left, the position of an automaton on the right.

Supposing we start out with a certain local perturbation of food concentration. Since one value is a little less than an adjacent value, this gives the thing a direction, so that he can sense that the food concentration here is higher than it is there. This determines the creature's move in that direction. Having determined this, it eats, and this determines its subsequent movement. If it is in an indefinitely extending plane, it just goes on until it encounters a boundary or another automaton, and it leaves behind it a wave of food depletion as shown in the lowest diagram of Fig. 6.

If you have rather more automata in the system, you get structures which are chain-like structures, due to the fact that automata tend to become nucleated and move into a region, where they approach each other reflecting each other, so that you get chains of oscillating automata which form coherent structures in the system. This is still in a plane.





Assume invariance for nodes which are not visited

FIG. 6. Element movement and food depletion.

If we modify the topology of the thing a little and make, for example, a cylinder, we come across the possibility of cyclic action. Such cycles can act, rather obviously, as a sort of memory device, but they can also act as filtering devices. Supposing I establish a cycle in a cylindrical field (Fig. 7), and I establish a gradient down this supposedly indefinitely long cylinder, so automata tend to jump with the gradient, then the cycle of an automaton in the

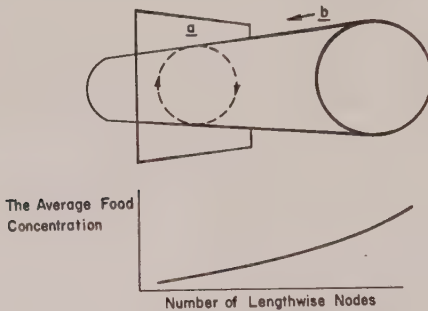


FIG. 7. Cylindrical field model.

transverse plane selectively prohibits or allows other automata to get through. Hence there is a certain sense in which filter-like structure, in terms of automata, are readily built up. For example, an automaton of composition *aoaobob* would be able to jump over this cycle altogether, it would not mind it, or at least a certain percentage of them would not, depending on whether they were on odd or even numbered nodes.

In sum, when you muck around a bit, and in particular when you define two neighborhoods with respect to the supply or source of food, it is possible to establish, in such a system, both cooperative and competitive activity. The dotted line in Fig. 8 is intended to show a source of food and, in a cylindrical field, you will see there are two alternative and energetically equivalent cycles,  $\gamma_1$  and  $\gamma_2$ , which can be established in this system, and which are liable to change into the other. A particularly interesting system, which I will not discuss at all, but which I think is worth notice because of the structures, which may be discussed later, is that of a multiple-genus torus, where you have got the possibility of independent

cycling where different species can come together (Fig. 9). I have a hunch it is no more than this, but so far we have not been able to realize such a field through lack of time or facilities, but it should be interesting to do so. I would just like to make one

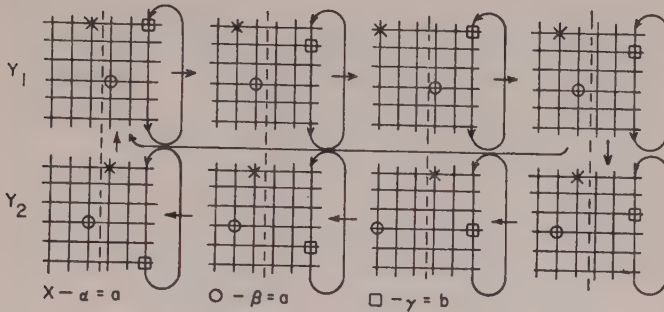


FIG. 8. Oscillatory movement in cylindrical field.

conjecture and please take this as such: that it seems that one way of introducing synthetic *a priori*'s into a system of this sort would be to produce topologies of this kind.

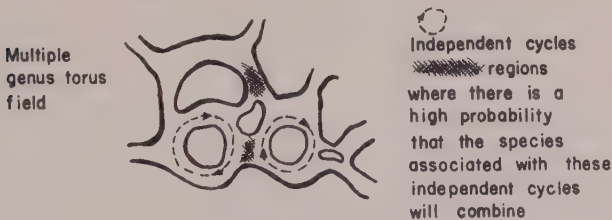


FIG. 9. High genus field.

Now I would like to pass on from this to considering what happens when we have a whole lot of automata interacting with the food supply network on which they live. Clearly in this case, there is a perfectly good sense in which the activity of an automaton, and in particular, a species of automaton, generated by the evolution permitted as a result of their previous feeding, will structure the world around them so that this particular species alone is favored.

Hence a system of this sort, and it can be argued quite rigorously,

is a self-replicating system. Furthermore since these structures, being geometrically bounded, are constrained, there will be a finite size to the structures and things will tend to come apart when they reach this critical limit. What I would like to do is give a special name to this odd kind of structure, which is a close coupling between a lot of automata and the world they live in. I will call it a domain. I will suppose that the domain is an existent in this sort of system, for there is no chance to discuss it adequately at the moment. I am particularly interested in what happens to a domain when, for example, we give the elements a lot of food. I

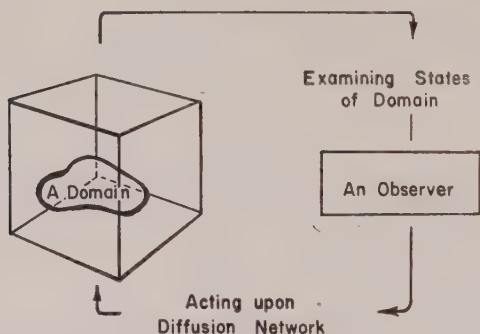


FIG. 10. Domain observation.

am particularly interested in this because it occurs to me this is one of the ways in which a game-theoretic explanation leads to clarity where others do not.

For example, suppose a three-space, in which we have got these creatures wandering around, and we have a domain (Fig. 10). Suppose I, as an observer, can by some miraculous means put my finger on this ephemeral entity and say that it is an organization which is wandering around, it is automata in equilibrium with their environment, which they modify, which is wandering around in a cube. Supposing I could put my finger on it in this way, I would be doing an operation, if I allowed it more food, if I favored this entity, equivalent to a linear transformation of the payoff function of the game. This is an entirely explicit thing to do. I simply add to each entry in the pay-off matrix, a small positive number. The

result of this is that more automata can live within the domain, and they will be the sort of automata which happen to be playing in this region. I will be favoring just those players.

Or, if you like, you can regard the whole system as having a gigantic payoff function, and you can say I am favoring the activities of a given coalition, for indeed, there is a certain identity, a certain similarity rather, between the domain and the notion of coalition which we advanced earlier.

Now, supposing I do this—and I agree that it is not a thing I can easily do—what happens is that the density of automata increases in this region. Here I think we get a very interesting result from our model, which I think has been repeated often enough to assert it. At a certain critical stage it locks solid. And at this stage, we must change our whimsical, though not unnecessary description of what happens. Instead of talking about automata in the region of this highly rewarded domain, we must now talk about chains and structures which exhibit exactly the logical characteristics of a model nervous system.

We have a refractory interval, a partial refractory interval, an impulse which is transmitted with a wave of food depletion. If we “Poop!” at one end of the chain of neurons of the sort I illustrated on the board, we get a result at the other end which is transmitted by this local energy depletion process. It does not much matter where the automata move, because they are not allowed to move very far. They are constrained by their own kind. It does matter a great deal in what order they move when they move.

The domain always locks solid like this, and at the end of the chains there are link violations which quite obviously have temporal and spatial summative characteristics which, hopefully perhaps, I would like to identify with synaptic connections.

It seems to me of interest that the operation of this hypothetical “thumb-putting-on” procedure will lead us, with the choice of parameters which, perhaps, I have taken as a hunch, always to this result: that the center of the domain, regarded as a critter, walking around a world in which it feeds, becomes structured and acquires this rather net-like sort of nervous system.

And I think it is also interesting that I cannot really describe what is happening in terms of putting my thumb on it, but I can

describe quite precisely what is happening in terms of the game-theoretic model.

There are one or two other things which I think we might point out. I did make a model in these terms. I ran a small program of these terms, to describe the development of a population of small social amoebae, cellular slime molds, which seemed to turn out quite successfully. I set up conditions whereby we had these

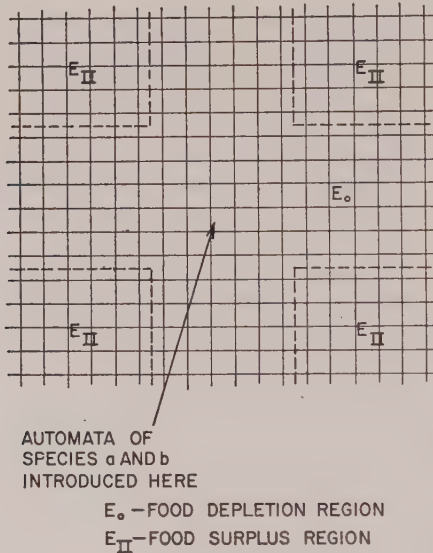


FIG. 11. Maze experiment.

creatures wandering around in a regionally depleted environment, and I added to the system the possibility of specific signalling. I gave them some "acrasin". And the result of this experiment will be perhaps interesting, although I cannot describe it in detail for lack of time. Figure 11 shows the field. We found that without any signal potentialities, given an experimental setup in which we have a region where there is food depletion and a region where there is plenty of food, and a structured network, then to get from one to the other you had to be a coalition, you had to combine. We do not know how many get out when you grant the



possibility of signalling, that is to say, when you vary the signal possibility given to each individual one so that they can track their kind. This sort of model is interesting, and interesting perhaps in the same way that the structuring one is.

Finally I want to comment on the payoff, because we have gladly supposed it is food, but really in some senses this can be ridiculous. Alternatively, it is perfectly possible for a domain to feed on automata. There is no real reason why we should take, in the gigantic cube of Fig. 10, only the supply of food as being that

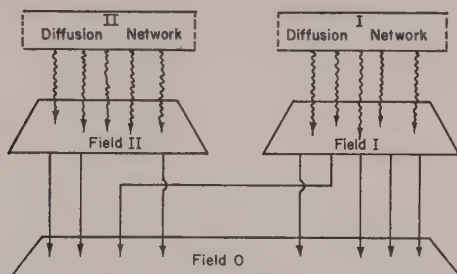


FIG. 12. Domain migration diagram.

which limits and causes competition. Equally, we may regard the migration of other domains of particular families of automata as being a source of food, or if we like, as initially a kind of catalytic action, since the specificity of these automata with respect to the domain into which they migrate can induce an amplifying action, which is, I think, rather easy to conceive.

The sort of thing I am thinking of in broad outline is shown in Fig. 12. Here, I suppose, a diffusion network and another diffusion network, two fields, and automata which can migrate down in a cubic lattice, for example, onto another field where there are other domains, where the migration of these would affect the survival or extinction of the domains that exist below.

So we have got into the realm of payoff functions which are defined in several ways. And I think this is again a very interesting region.

Now, finally, what is this all about, why should we play with these automata? I think the reason why we play with them is in

order to find how the structures which have been described as the structures engendering logical stability, redundancy of potential command, and redundancy of computation, occur, in our mushes, our threads, our Dr. Beurle's networks, and anywhere else we please, all over the place.

How is it that systems like this occur, systems which are essentially characterized by a non-zero sum partially competitive game in which, in addition to the usual concepts, structure must be paid for?

How is it that these domains acquire, as indeed they do, as is obvious from the discussions, just these characteristics? I do not know how it is, but I do hope by means of a much more sophisticated and a much better presented model we may find out.

## DISCUSSION

VON FOERSTER: Thank you, Gordon, for a most delightful paper, and I think if somebody was asleep at the beginning he is now wide awake. And I would like to ask first one question, Gordon, if you may, perhaps, give a little explanation of the particular experiment with the amoeba, so that everybody really can follow what kind of a setup this was, and so that we really see what happens in this particular instant.

PASK: The conditions are as follows: we wanted to simulate a cellular slime mold population; if you recall, the life cycle of a slime mold is a crazy mixed-up thing. The creatures start out as amoebae, and they live anywhere on the place and they just look for food. These amoebae have a rather remarkable signalling system. They produce a substance called acrasin, which is simulated by certain steroids. This diffuses away from them and the appearance of acrasin, at the boundary of another amoeba, causes this amoeba to move towards the source, in a chemotactic manner. So if you look at the culture, you will eventually see aggregates and there are streams of amoeba moving into them. When they have collected sufficiently, they form an organized whole, in which cells appear to take different functional parts. There is a definite hind and middle and fore end of the thing.

Ultimately, and this is the part of the model I haven't realized, because I would have to put in a duplication rule which I haven't mentioned, the thing forms a spore body on the end of some sort of structure, depending upon the species; they form all sorts of peculiar kinds of cellular structures. A spore body is formed which leads back to further amoeba, which are disseminated by wind and water.

Now, this is the life cycle I was anxious to induce, and I was anxious to induce it given simply the constraint that it was possible for a certain sort of structure to exist, in a food-depleted environment, where the original members of the population could not exist. And I wanted to see how their existence was determined by giving them the acrasin signalling system which gets them from here to there.

The experiment was set up in the following manner. I took a field as shown in Fig. 11. Food concentration at EII regions is high enough to sustain anything. Elsewhere we have a depleted region; the amount of food therein is dropping all the time. There are automata in the depleted region, they are eating and taking away the possibility of living as simple automata, but they can only survive if they are automata that can move to the EII regions.

So in order to get through the depleted regions, they have got to combine. The experiment was a very simple one indeed, namely, to plot the number that escaped as a function of the amount of acrasin given them.

RAPOPORT: Why did they have to combine in order to get through?

PASK: The automata introduced into the depleted region of the maze were only of species *a* and *b*. To reach the EII regions they had to combine as *aob* automata.

RAPOPORT: These are real flesh and blood amoeba?

PASK: I am simulating the amoeba's life cycle. The matter is simply that in order to get out of the central depleted region, it is necessary for an automaton to turn at least one corner. In order to be able to turn a corner, it must be a combined sort of an automaton.

YOVITS: Is this a geometrical combination?

PASK: It is a geometrical combination.

YOVITS: It is sufficiently large so that part of it remains in a food-giving region, is that what you are trying to say?

PASK: No, there is a food-depleted region in which it is just viable.

YOVITS: If the whole automaton is within the region which has no food, will it die?

PASK: Well, you start out with simple automata. Now they would all die, because they eat all the food. So that in order to get out they must get through this region where they would die before they got through it. So in order to get through the region they must turn at least one corner.

YOVITS: After they combine, how do they get through the region?

PASK: The boundary here actually is in terms of having to turn the corner. They have got to be combined automata in order to get out.

NOVIKOFF: Isn't that their food demands are less when combined?

PASK: No, no, definitely. As a matter of fact, I have another model in which this is the case, but that particular one I was describing here is one that has to turn a corner.

ROSENBLATT: In this problem, this model actually produces a depletion of food there around it?

PASK: I am modelling in terms of my stupid little automata. But why do I model in them? Because it is the simplest system I know of in which there is a definite cooperative action which increases, in other words, cooperation occurs in conditions of depletion, and a definite cooperating organism is formed under conditions of plenty of food.

YOVITS: This is a hypothesis that if they turn a corner they will get out?

MCCULLOCH: May I speak in answer to that? The point as I see it, there is such a beast as an amoeba, but he is not talking about that beast, he is talking about a model composed of straight lines.

ROSENBLATT: But how does the real beast get out itself? Does it have to organize to get out too?

PASK: Yes, it has to actually, it gets out by sticking together. The act of turning a corner in my model is equivalent to the act of sticking together.

VON FOERSTER: I have a particular question. Now we have talked at the

moment of actual living systems, but what I was thinking, in terms of Pask's model where one could really assign a mathematical functional way of formulation, or, if you would like, to use the second law of thermodynamics for open systems, one could assign to a model the precise values which would be necessary to carry out these confirmations. From this point of view, you see, we could make, a real one-to-one correspondence.

In this simple model, I think it might not be too complicated to begin a simple correlation between these two fields, and then develop or try to apply it to the actual system.

PASK: In other words, do it without hope that it may be a testable model.

BOWMAN: Just a very brief word. I speak now not on the subject as presented, but from purely a biological standpoint. I have at first hand observed the movement of a slime mold\* in the plasmodium stage across a dusty dirt road. Now what particular gradient could have induced that I have not yet found out. Yet the animal has no eyes and apparently no structural organization at all, and yet it was deliberately crossing that dirt road.

ADDISON: Does it ever go down the road?

BOWMAN: My research was incomplete.

SHERWOOD: Did you try to deviate it?

BOWMAN: I did, it went around an obstacle and continued on the path.

SHERWOOD: 180 degrees?

BOWMAN: No, that I did not do, I put a board in its path, a dry board which it did not go over, but went around, kept on the move.

YOVITS: What does it do after it gets across the road?

BOWMAN: I followed this for about ten days. It finally found an old, rotten log and formed its spore stalks.

ADDISON: Did you try digging the soil it was on and turn it?

BOWMAN: My research isn't that complete.

BEER: How big was that colony?

BOWMAN: The size of my hand, about that size and shape.

PASK: That is a big one.

ROSEN: Just a little word, in a very trivial sense, on the growth of domains in ferromagnetics and ferro-electrics. They have many of the properties that you outlined except it is quite simple, and there the equivalent of your food would be the energy of the whole system.

PASK: Yes.

ROSEN: What is real interesting in the terms of the geometry you described that is required for this to get out of the food depletion area. In the ferro-electric domain, if you have one domain neighboring another domain which is oriented ninety degrees to it, this is a very stable condition, nothing happens.

On the other hand, if there are two domains that are oriented 180 degrees from each other, it is quite possible under the action of a gradient, which in this context is a field, for the larger domain to swallow up the smaller domain, and it does this in a peculiar fashion, putting out wedges which grow.

YOVITS: But is this the same phenomenon, isn't the essential action of these domains one to form minimum energy? Now these domains you are talking about do not tend to form minimum energy.

ROSEN: There is one more remark I want to make, nothing happens until

---

\* *Editor's note*—Mr. Pask has commented that Dr. Bowman appears to be describing one of the acellular slime molds. Mr. Pask's model has reference only to cellular slime molds with an "acrasin" signalling system.



you get a lot of cooperation among a bunch of dipoles. An individual does not do a thing. As I say, this is very simple.

YOVITS: Because it happens not to be a minimum configuration.

ROSEN: It happens that in physics you have a nice principle to go back on. Perhaps one can find such a nice simple unified principle here.

PASK: Yes, look, I think the point I will perhaps bring out, and I think that your comment, Dr. Rosen, is not at all without relevance. I think it is a highly relevant thing, but I think there is a distinction to be made between these two concepts, although they are related.

ROSEN: One is very much simpler than the other.

RAPOPORT: I would like to ask a couple of questions. First with regard to your calling this a game-theoretic model. If I understand it, it has to do with the game-theoretical model inasmuch as it applies to coalition formations, each coalition commanding a certain payoff. Therefore, we are reminded of the game in characteristic function form, in which each coalition commands a certain payoff, but when one asks questions about what coalition you actually form, one finds that immediately one gets into hot water, because this straightforward answer, that that coalition will form which commands the greatest payoff, is by no means the case. Because if such a coalition should form, which commands the greatest payoff, it means that the anti-coalition would also form because that's the best thing that can happen to the other side, that would command the smallest payoff. And then one is faced with the dilemma, if the coalition which commands the greatest forms, why does the other coalition form, since it commands the least?

The whole meat of the coalition formation theory is that there is competition for the members. So that it is not at all true that the coalition which commands the greatest payoff will form, and indeed, the theory of games in characteristic function form, as you well know, in its original formulation, didn't have any answer whatsoever. Except for von Neumann's solution of games, they are a laugh. They are ridiculous, they tell you absolutely nothing. In most cases every coalition is a solution. Every coalition is an imputation, is a solution. So that one cannot even say anything about what will happen.

But what I want to ask you is whether, aside from these general natural selections which you were talking about, there is anything else that can maybe shed light upon this respect? That is my first question.

My second question has to do with something that you started with, and I thought that you would elaborate it further, I would very much like to hear your opinion on this.

You said the self-organizing systems are characterized by this peculiar trait, that when one looks at them one doesn't know what questions to ask. Is that so?

PASK: I said that there is a class of systems where we do not know what questions to ask.

RAPOPORT: Right. And I thought for a moment you were talking about a situation similar to the following one, and wondered if it has any connection.

There was a man that came from Mars, and he began watching a chess game as they were playing it on earth, and he decided that having watched 10,000 chess games he had a mathematical theory of chess, and he went ahead and developed it, and that theory was simply wonderful. It predicted with great accuracy what the rate of depletion will be of the chess pieces, and it is replicable; if you take the first 10,000 and the next 10,000 games, the rate of depletion will be exactly the same.

And he even developed differential equations from which these rates of depletions were deduced. He developed equations which told him the distribution of the length of the game, of the probability of white winning over black in every kind of game played, and vice versa. In fact, he developed every possible statistics of chess you can think of, and he developed a very good set of fundamental axioms from which all this mathematics was developed.

And then he brought his theory to an earthly chess player, and he told him that it wasn't chess, he told him he had asked the wrong questions.

Now, does this have anything to do with what you are thinking of?

PASK: It does have something to do with it.

RAPOPORT: Would you please comment then further, and also my first question, please?

PASK: I agree that your comments about the von Neuman  $n$ -person game and the coalition formation in it were entirely valid, and I think that the best way to answer the question, and to expose the possible utility of our model, is to develop slightly the conditions which arise within it.

In the first place, we are not thinking so much of the von Neumann model, as of the kind of model proposed by Luce, in which there is a thing called a  $\Psi$  function which is absolutely central.

Let  $B$  represent the coalition structure, that is the aggregate of coalitions, of the whole set of coalitions, of the game as the time  $T$  equals  $T_{\text{zero}}$ ; let  $\rho$  represent the correlated strategies adopted, in equilibrium, which will be some solution by these numbers of these coalitions. Then  $\Psi(B\rho)$  is the set of coalition structure, common strategy pairs, which are admissible at  $T_0 + \delta T$ .

Now, we are interested in this in the following sense, that in a sociological situation we commonly have to guess at  $\Psi$ , which represents the social inertia. In the present kind of automaton, we are in a much happier situation, because even if we make it out of threads or goo or semiconductors, we know the  $\Psi$  function; we know the admissible coalition structures which can occur. These arise from the inertial parameters of the system.

RAPOPORT: There is a permissible transition in the system.

PASK: It is a permissible transition, given those in a given stage. The same is true here, surely. In the sense that: given I have a certain coalition structure, I can, for example, take a  $K$  game and just add one, or whatever it may be, whatever rule exists for this possibility. But in the  $K$  it won't be quite as simple as just adding one or subtracting one or something of this kind. It will, in fact, be more complicated, and it will be determined by the characteristics of the funny material at the bottom, and by the time constants of this and so forth. And so a certain amount of structure can come in this way.

The second thing is that we surely can make some assertions, although I agree that they are rather poverty stricken, about the payoff functions which can be in equilibrium with a given common strategy, a coalition pair.

So that looking at the thing again a little crudely, we can talk about equilibria which consist of a sequence of payoff functions, induced by the existence of a coalition structure here, and another one which arises because of this, and then the arrival of a coalition structure which is induced by the payoff function.

If it happens that transformation  $T$  of  $G_1$  and  $G_2$  form a group and it is a cyclic group, we have a stable condition. And these kinds of stable conditions are analogous, perhaps, to resonance hybrids, because it always happens that food maintenance cost of one coalition structure must map into another if the food maintenance cost, the average payoff, is constant over the set of  $G$ 's.

Question number two, the man from Mars, yes, I think he is looking at a



and the people of the United States. It is a great and noble work, and one which will be read and valued by all who are interested in the history of our country. The author has done his best to make it as complete and as accurate as possible, and I am sure that it will be found to contain all the facts and details which are necessary for a full and correct knowledge of the subject. It is a work which will be read and valued by all who are interested in the history of our country.

The author has done his best to make it as complete and as accurate as possible, and I am sure that it will be found to contain all the facts and details which are necessary for a full and correct knowledge of the subject. It is a work which will be read and valued by all who are interested in the history of our country.

The author has done his best to make it as complete and as accurate as possible, and I am sure that it will be found to contain all the facts and details which are necessary for a full and correct knowledge of the subject. It is a work which will be read and valued by all who are interested in the history of our country.



**W. ROSS ASHBY**

*University of Illinois*

## PRINCIPLES OF THE SELF-ORGANIZING SYSTEM\*

Questions of principle are sometimes regarded as too unpractical to be important, but I suggest that that is certainly not the case in *our* subject. The range of phenomena that we have to deal with is so broad that, were it to be dealt with wholly at the technological or practical level, we would be defeated by the sheer quantity and complexity of it. The total range can be handled only piecemeal; among the pieces are those homomorphisms of the complex whole that we call "abstract theory" or "general principles". They alone give the bird's-eye view that enables us to move about in this vast field without losing our bearings. I propose, then, to attempt such a bird's-eye survey.

### WHAT IS "ORGANIZATION"?

At the heart of our work lies the fundamental concept of "organization". What do we mean by it? As it is used in biology it is a somewhat complex concept, built up from several more primitive concepts. Because of this richness it is not readily defined, and it is interesting to notice that while March and Simon (1958) use the word "Organizations" as title for their book, they do not give a formal definition. Here I think they are right, for the word covers a multiplicity of meanings. I think that in future we shall hear the *word* less frequently, though the *operations* to which it corresponds, in the world of computers and brain-like mechanisms, will become of increasing daily importance.

The hard core of the concept is, in my opinion, that of "conditionality". As soon as the relation between two entities *A* and *B*

---

\* The work on which this paper is based was supported by ONR Contract N 049-149.

becomes conditional on  $C$ 's value or state then a necessary component of "organization" is present. Thus *the theory of organization is partly co-extensive with the theory of functions of more than one variable.*

We can get another angle on the question by asking "what is its converse?" The converse of "conditional on" is "not conditional on", so the converse of "organization" must therefore be, as the mathematical theory shows as clearly, the concept of "reducibility". (It is also called "separability".) This occurs, in mathematical forms, when what looks like a function of several variables (perhaps very many) proves on closer examination to have parts whose actions are *not* conditional on the values of the other parts. It occurs in mechanical forms, in hardware, when what looks like one machine proves to be composed of two (or more) sub-machines, each of which is acting independently of the others.

Questions of "conditionality", and of its converse "reducibility", can, of course, be treated by a number of mathematical and logical methods. I shall say something of such methods later. Here, however, I would like to express the opinion that the method of Uncertainty Analysis, introduced by Garner and McGill (1956), gives us a method for the treatment of conditionality that is not only completely rigorous but is also of extreme generality. Its great generality and suitability for application to complex behavior, lies in the fact that it is applicable to any arbitrarily defined set of states. Its application requires neither linearity, nor continuity, nor a metric, nor even an ordering relation. By this calculus, the *degree* of conditionality can be measured, and analyzed, and apportioned to factors and interactions in a manner exactly parallel to Fisher's method of the analysis of variance; yet it requires no metric in the variables, only the frequencies with which the various combinations of states occur. It seems to me that, just as Fisher's conception of the analysis of variance threw a flood of light on to the complex relations that may exist between variations on a metric, so McGill and Garner's conception of uncertainty analysis may give us an altogether better understanding of how to treat complexities of relation when the variables are non-metric. In psychology and biology such variables occur with great commonness; doubtless they will also occur commonly in the brain-like

processes developing in computers. I look forward to the time when the methods of McGill and Garner will become the accepted language in which such matters are to be thought about and treated quantitatively.

The treatment of "conditionality" (whether by functions of many variables, by correlation analysis, by uncertainty analysis, or by other ways) makes us realize that the essential idea is that there is first a product space—that of the *possibilities*—within which some sub-set of points indicates the actualities. This way of looking at "conditionality" makes us realize that it is related to that of "communication"; and it is, of course, quite plausible that we should define parts as being "organized" when "communication" (in some generalized sense) occurs between them. (Again the natural converse is that of independence, which represents non-communication.)

Now "communication" from  $A$  to  $B$  necessarily implies some constraint, some correlation between what happens at  $A$  and what at  $B$ . If, for given event at  $A$ , all possible events may occur at  $B$ , then there is no communication from  $A$  to  $B$  and no constraint over the possible ( $A, B$ )-couples that can occur. Thus the presence of "organization" between variables is equivalent to the existence of a *constraint* in the product-space of the possibilities. I stress this point because while, in the past, biologists have tended to think of organization as something extra, something *added* to the elementary variables, the modern theory, based on the logic of communication, regards organization as a restriction or constraint. The two points of view are thus diametrically opposed; there is no question of either being exclusively right, for each can be appropriate in its context. But with this opposition in existence we must clearly go carefully, especially when we discuss with others, lest we should fall into complete confusion.

This excursion may seem somewhat complex but it is, I am sure, advisable, for we have to recognize that the discussion of organization theory has a peculiarity not found in the more objective sciences of physics and chemistry. The peculiarity comes in with the product space that I have just referred to. Whence comes this product space? Its chief peculiarity is that *it contains more than actually exists in the real physical world*, for it is the latter that gives us the actual, constrained *subset*.

The real world gives the subset of what *is*; the product space represents the uncertainty of the *observer*. The product space may therefore change if the observer changes; and two observers may legitimately use different product spaces within which to record the same subset of actual events in some actual thing. The "constraint" is thus a *relation* between observer and thing; the properties of any particular constraint will depend on both the real thing and on *the observer*. It follows that a substantial part of the theory of organization will be concerned with *properties that are not intrinsic to the thing but are relational between observer and thing*. We shall see some striking examples of this fact later.

### WHOLE AND PARTS

"If conditionality" is an essential component in the concept of organization, so also is the assumption that we are speaking of a whole composed of parts. This assumption is worth a moment's scrutiny, for research is developing a theory of dynamics that does *not* observe parts and their interactions, but treats the system as an unanalysed whole (Ashby, 1958, a). In physics, of course, we usually start the description of a system by saying "Let the variables be  $x_1, x_2, \dots, x_n$ " and thus start by treating the whole as made of  $n$  functional parts. The other method, however, deals with unanalysed states,  $S_1, S_2, \dots$  of the whole, without explicit mention of any parts that may be contributing to these states. The dynamics of such a system can then be defined and handled mathematically; I have shown elsewhere (Ashby, 1960, a) how such an approach can be useful. What I wish to point out here is that we can have a sophisticated *dynamics*, of a whole as complex and cross-connected as you please, that makes no reference to any parts and that therefore does *not* use the concept of organization. Thus the concepts of dynamics and of organization are essentially independent, in that all four combinations, of their presence and absence, are possible.

This fact exemplifies what I said, that "organization" is partly in the eye of the beholder. Two observers studying the same real material system, a hive of bees say, may find that one of them, thinking of the hive as an interaction of fifty thousand bee-parts, finds the bees "organized", while the other, observing whole states



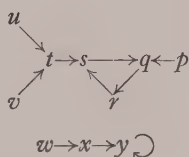
such as activity, dormancy, swarming, etc., may see *no* organization, only trajectories of these (unanalysed) states.

Another example of the independence of "organization" and "dynamics" is given by the fact that whether or not a real system is organized or reducible depends partly on the point of view taken by the observer. It is well known, for instance, that an organized (i.e. interacting) linear system of  $n$  parts, such as a network of pendulums and springs, can be seen from another point of view (that of the so-called "normal" coordinates) in which all the (newly identified) parts are completely separate, so that the whole is reducible. There is therefore nothing perverse about my insistence on the relativity of organization, for advantage of the fact is routinely taken in the study of quite ordinary dynamic systems.

Finally, in order to emphasize how dependent is the organization seen in a system on the observer who sees it, I will state the proposition that: given a whole with arbitrarily given behavior, a great variety of arbitrary "parts" can be seen in it; for all that is necessary, when the arbitrary part is proposed, is that we assume the given part to be coupled to another suitably related part, so that the two together form a whole isomorphic with the whole that was given. For instance, suppose the given whole,  $W$  of 10 states, behaves in accordance with the transformation:

$$W \begin{array}{c} \downarrow \\ \begin{array}{cccccccccc} p & q & r & s & t & u & v & w & x & y \\ q & r & s & q & s & t & t & x & y & y \end{array} \end{array}$$

Its kinematic graph is



and suppose we wish to "see" it as containing the part  $P$ , with internal states  $E$  and input states  $A$ :

	$E$	}
↓	1 2	
	2 1	
$A$	2 1 1	

With a little ingenuity we find that if part  $P$  is coupled to part  $Q$  (with states  $(F, G)$  and input  $B$ ) with transformation  $Q$ :

$$\begin{array}{c|cccccc}
 & \downarrow & 1, 1 & 1, 2 & 1, 3 & 2, 1 & 2, 2 & 2, 3 \\
 B & 1 & 2, 1 & 1, 2 & 1, 2 & 2, 1 & 1, 2 & 1, 2 \\
 & 2 & \cdot & 2, 3 & \cdot & 2, 1 & 2, 2 & 2, 2
 \end{array} \left. \vphantom{\begin{array}{c|cccccc} \right\} Q$$

by putting  $A = F$  and  $B = E$ , then the new whole  $W'$  has transformation

$$W': \quad \downarrow \begin{array}{cccc}
 1, 1, 1 & 1, 1, 2 & 1, 1, 3 & 1, 2, 1, \text{ etc.} \\
 2, 2, 1 & 2, 1, 2 & 2, 1, 2 & 1, 2, 1, \text{ etc.}
 \end{array}$$

which is *isomorphic with*  $W$  under the one-one correspondence

$$\begin{array}{cccc}
 1, 1, 1 & 1, 1, 2 & 1, 1, 3 & 1, 2, 1, \text{ etc.} \\
 \downarrow & & & \\
 w & s & p & y, \text{ etc.}
 \end{array}$$

Thus, subject only to certain requirements (e.g. that equilibria map into equilibria) *any dynamic system can be made to display a variety of arbitrarily assigned "parts"*, simply by a change in the *observer's* view point.

### MACHINES IN GENERAL

I have just used a way of representing two "parts", "coupled" to form a "whole", that anticipates the question: what do we mean by a "machine" in general?

Here we are obviously encroaching on what has been called "general system theory", but this last discipline always seemed to me to be uncertain whether it was dealing with *physical* systems, and therefore tied to whatever the real world provides, or with mathematical systems, in which the sole demand is that the work shall be free from internal contradictions. It is, I think, one of the substantial advances of the last decade that we have at last identified the *essentials* of the "machine in general".

Before the essentials could be seen, we had to realize that two factors must be *excluded as irrelevant*. The first is "materiality"—the idea that a machine must be made of actual matter, of the hundred or so existent elements. This is wrong, for examples can

readily be given (e.g. Ashby, 1958, a) showing that what is essential is whether the system, of angels and ectoplasm if you please, *behaves* in a law-abiding and machine-like way. Also to be excluded as irrelevant is any reference to energy, for any calculating machine shows that what matters is the *regularity* of the behavior—whether energy is gained or lost, or even created, is simply irrelevant.

The fundamental concept of “machine” proves to have a form that was formulated at least a century ago, but this concept has not, so far as I am aware, ever been used and exploited vigorously. A “machine” is that which behaves in a machine-like way, namely, that its internal state, and the state of its surroundings, defines uniquely the next state it will go to.

This definition, formally proposed fifteen years ago (Ashby, 1945) has withstood the passage of time and is now becoming generally accepted (e.g. Jeffrey, 1959). It appears in many forms. When the variables are continuous it corresponds to the description of a dynamic system by giving a set of ordinary differential equations with time as the independent variable. The *fundamental* nature of such a representation (as contrasted with a merely convenient one) has been recognized by many earlier workers such as Poincaré, Lotka (1925), and von Bertalanffy (1950 and earlier).

Such a representation by differential equations is, however, too restricted for the needs of a science that includes biological systems and calculating machines, in which discontinuity is ubiquitous. So arises the modern definition, able to include both the continuous and the discontinuous and even the discrete, without the slightest loss of rigor. The “machine with input” (Ashby, 1958, a) or the “finite automaton” (Jeffrey, 1959) is today defined by a set  $S$  of internal states, a set  $I$  of input or surrounding states, and a mapping,  $f$  say, of the product set  $I \times S$  into  $S$ . Here, in my opinion, we have the very essence of the “machine”; all known types of machine are to be found here; and all interesting deviations from the concept are to be found by the corresponding deviation from the definition.

We are now in a position to say without ambiguity or evasion what we mean by a machine’s “organization”. First we specify which system we are talking about by specifying its states  $S$  and its

conditions  $I$ . If  $S$  is a product set, so that  $S = \prod_i T_i$  say, then the parts  $i$  are each specified by its set of states  $T_i$ . The "organization" between these parts is then specified by the mapping  $f$ . Change  $f$  and the organization changes. In other words, the possible organizations between the parts can be set into one-one correspondence with the set of possible mappings of  $I \times S$  into  $S$ . Thus "organization" and "mapping" are two ways of looking at the same thing—the organization being noticed by the observer of the actual system, and the mapping being recorded by the person who represents the behavior in mathematical or other symbolism.

### "GOOD" ORGANIZATION

At this point some of you, especially the biologists, may be feeling uneasy; for this definition of organization makes no reference to any *usefulness* of the organization. It demands only that there be conditionality between the parts and regularity in behavior. In this I believe the definition to be right, for the question whether a given organization is "good" or "bad" is quite independent of the prior test of whether it is or is not an organization.

I feel inclined to stress this point, for here the engineers and the biologists are likely to think along widely differing lines. The engineer, having put together some electronic hardware and having found the assembled network to be roaring with parasitic oscillations, is quite accustomed to the idea of a "bad" organization; and he knows that the "good" organization has to be searched for. The biologist, however, studies mostly animal species that have survived the long process of natural selection; so almost all the organizations he sees have already been selected to be good ones, and he is apt to think of "organizations" as *necessarily* good. This point of view may often be true in the biological world but it is most emphatically not true in the world in which we people here are working. We *must* accept that

- (1) most organizations are bad ones;
- (2) the good ones have to be sought for; and
- (3) what is meant by "good" must be clearly defined, explicitly if necessary, *in every case*.

What then is meant by "good", in our context of brain-like mechanisms and computers? We must proceed cautiously, for the

word suggests some evaluation whose origin has not yet been considered.

In some cases the distinction between the "good" organization and the "bad" is obvious, in the sense that as everyone in these cases would tend to use the same criterion, it would not need explicit mention. The brain of a living organism, for instance, is usually judged as having a "good" organization if the organization (whether inborn or learned) acts so as to further the organism's survival. This consideration readily generalizes to all those cases in which the organization (whether of a cat or an automatic pilot or an oil refinery) is judged "good" if and only if it acts so as to keep an assigned set of variables, the "essential" variables, within assigned limits. Here are all the mechanisms for homeostasis, both in the original sense of Cannon and in the generalized sense. From this criterion comes the related one that an organization is "good" if it makes the system stable around an assigned equilibrium. Sommerhoff (1950) in particular has given a wealth of examples, drawn from a great range of biological and mechanical phenomena, showing how in all cases the idea of a "good organization" has as its essence the idea of a number of parts so interacting as to achieve some given "focal condition". I would like to say here that I do not consider that Sommerhoff's contribution to our subject has yet been adequately recognized. His identification of *exactly* what is meant by coordination and integration is, in my opinion, on a par with Cauchy's identification of exactly what was meant by convergence. Cauchy's discovery was a real discovery, and was an enormous help to later workers by providing them with a concept, rigorously defined, that could be used again and again, in a vast range of contexts, and always with exactly the same meaning. Sommerhoff's discovery of how to represent *exactly* what is meant by coordination and integration and good organization will, I am sure, eventually play a similarly fundamental part in our work.

His work illustrates, and emphasizes, what I want to say here—*there is no such thing as "good organization" in any absolute sense.* Always it is relative; and an organization that is good in one context or under one criterion may be bad under another.

Sometimes this statement is so obvious as to arouse no opposition. If we have half a dozen lenses, for instance, that can be



assembled this way to make a telescope or that way to make a microscope, the goodness of an assembly obviously depends on whether one wants to look at the moon or a cheese mite.

But the subject is more contentious than that! The thesis implies that there is no such thing as a brain (natural or artificial) that is good in any absolute sense—it all depends on the circumstances and on what is wanted. Every faculty that a brain can show is “good” only conditionally, for there exists at least one environment against which the brain is handicapped by the possession of this faculty. Sommerhoff’s formulation enables us to show this at once: whatever the faculty or organization achieves, let that be *not* in the “focal conditions”.

We know, of course, lots of examples where the thesis is true in a somewhat trivial way. Curiosity tends to be good, but many an antelope has lost its life by stopping to see what the hunter’s hat is. Whether the organization of the antelope’s brain should be of the type that does, or does not, lead to temporary immobility clearly depends on whether hunters with rifles are or are not plentiful in its world.

From a different angle we can notice Pribram’s results (1957), who found that brain-operated monkeys scored higher in a certain test than the normals. (The operated were plodding and patient while the normals were restless and distractible.) Be that as it may, one cannot say which brain (normal or operated) had the “good” organization until one has decided which sort of temperament is wanted.

Do you still find this non-contentious? Then I am prepared to assert that there is not a single mental faculty ascribed to Man that is good in the absolute sense. If any particular faculty is *usually* good, this is solely because our terrestrial environment is so lacking in variety that its usual form makes that faculty usually good. But change the environment, go to really different conditions, and possession of that faculty may be harmful. And “bad”, by implication, is the brain organization that produces it.

I believe that there is not a single faculty or property of the brain, usually regarded as desirable, that does not become *undesirable* in some type of environment. Here are some examples in illustration.

The first is Memory. Is it not good that a brain should have



memory? Not at all, I reply—only when the environment is of a type in which the future often *copies* the past; should the future often be the *inverse* of the past, memory is actually disadvantageous. A well known example is given when the sewer rat faces the environmental system known as “pre-baiting”. The naïve rat is very suspicious, and takes strange food only in small quantities. If, however, wholesome food appears at some place for three days in succession, the sewer rat will learn, and on the fourth day will eat to repletion, and die. The rat without memory, however, is as suspicious on the fourth day as on the first, and lives. Thus, in *this* environment, memory is positively disadvantageous. Prolonged contact with this environment will lead, other things being equal, to evolution in the direction of diminished memory-capacity.

As a second example, consider organization itself in the sense of connectedness. Is it not good that a brain should have its parts in rich functional connection? I say, No—not *in general*; only when the environment is itself richly connected. When the environment’s parts are *not* richly connected (when it is highly reducible, in other words), adaptation will go on faster if the brain is also highly reducible, i.e. if its connectivity is small (Ashby, 1960, d). Thus the *degree* of organization can be too high as well as too low; the degree we humans possess is probably adjusted to be somewhere near the optimum for the usual terrestrial environment. It does not in any way follow that this degree will be optimal or good if the brain is a mechanical one, working against some grossly non-terrestrial environment—one existing only inside a big computer, say.

As another example, what of the “organization” that the biologist always points to with pride—the development in evolution of specialized organs such as brain, intestines, heart and blood vessels. Is not this good? Good or not, it is certainly a specialization made possible only because the earth has an atmosphere; without it, we would be incessantly bombarded by tiny meteorites, any one of which, passing through our chest, might strike a large blood vessel and kill us. Under such conditions a better form for survival would be the slime mould, which specializes in being able to flow through a tangle of twigs without loss of function. Thus the development of organs is not good unconditionally, but is a specialization to a world free from flying particles.

After these actual instances, we can return to theory. It is here that Sommerhoff's formulation gives such helpful clarification. He shows that in all cases there must be given, and specified, first a *set of disturbances* (values of his "coenetic variable") and secondly a goal (his "focal condition"); the disturbances threaten to drive the outcome outside the focal condition. The "good" organization is then of the nature of a *relation* between the set of disturbances and the goal. Change the set of disturbances, and the organization, without itself changing, is evaluated "bad" instead of "good". As I said, there is no property of an organization that is good in any absolute sense; all are relative to some given environment, or to some given set of threats and disturbances, or to some given set of problems.

### SELF-ORGANIZING SYSTEMS

I hope I have not wearied you by belaboring this relativity too much, but it is fundamental, and is only too readily forgotten when one comes to deal with organizations that are either biological in origin or are in imitation of such systems. With this in mind, we can now start to consider the so-called "self-organizing" system. We must proceed with some caution here if we are not to land in confusion, for the adjective is, if used loosely, ambiguous, and, if used precisely, self-contradictory.

To say a system is "self-organizing" leaves open two quite different meanings.

There is a first meaning that is simple and unobjectionable. This refers to the system that starts with its parts separate (so that the behavior of each is independent of the others' states) and whose parts then act so that they change towards forming connections of some type. Such a system is "self-organizing" in the sense that it changes from "parts separated" to "parts joined". An example is the embryo nervous system, which starts with cells having little or no effect on one another, and changes, by the growth of dendrites and formation of synapses, to one in which each part's behavior is very much affected by the other parts. Another example is Pask's system of electrolytic centers, in which the growth of a filament from one electrode is at first little affected by growths at the other electrodes; then the growths become

more and more affected by one another as filaments approach the other electrodes. In general such systems can be more simply characterized as “self-connecting”, for the change from independence between the parts to conditionality can always be seen as some form of “connection”, even if it is as purely functional as that from a radio transmitter to a receiver.

Here, then, is a perfectly straightforward form of self-organizing system; but I must emphasize that there can be no assumption at this point that the organization developed will be a good one. If we wish it to be a “good” one, we must first provide a criterion for distinguishing between the bad and the good, and then we must ensure that the appropriate selection is made.

We are here approaching the second meaning of “self-organizing” (Ashby, 1947). “Organizing” may have the first meaning, just discussed, of “changing from unorganized to organized”. But it may also mean “changing from a bad organization to a good one”, and this is the case I wish to discuss now, and more fully. This is the case of peculiar interest to *us*, for this is the case of the system that changes itself from a bad way of behaving to a good. A well known example is the child that starts with a brain organization that makes it fire-seeking; then a change occurs, and a new brain organization appears that makes the child fire-avoiding. Another example would occur if an automatic pilot and a plane were so coupled, by mistake, that positive feedback made the whole error-aggravating rather than error-correcting. Here the organization is bad. The system would be “self-organizing” if a change were *automatically* made to the feedback, changing it from positive to negative; then the whole would have changed from a bad organization to a good. Clearly, *this* type of “self-organization” is of peculiar interest to us. What is implied by it?

Before the question is answered we must notice, if we are not to be in perpetual danger of confusion, that *no machine can be self-organizing in this sense*. The reasoning is simple. Define the set  $S$  of states so as to specify which machine we are talking about. The “organization” must then, as I said above, be identified with  $f$ , the mapping of  $S$  into  $S$  that the basic drive of the machine (whatever force it may be) imposes. Now the logical relation here is that  $f$  determines the changes of  $S$ :— $f$  is *defined* as the set of

couples  $(s_i, s_j)$  such that the internal drive of the system will force state  $s_i$  to change to  $s_j$ . To allow  $f$  to be a function of the state is to make nonsense of the whole concept.

Since the argument is fundamental in the theory of self-organizing systems, I may help explanation by a parallel example. Newton's law of gravitation says that  $F = M_1M_2/d^2$ , in particular, that the force varies inversely as the distance to power 2. To power 3 would be a different law. But suppose it were suggested that, not the force  $F$  but the *law* changed with the distance, so that the power was not 2 but some function of the distance,  $\phi(d)$ . This suggestion is illogical; for we now have that  $F = M_1M_2/d^{\phi(d)}$ , and this represents not a law that varies with the distance but *one* law covering all distances; that is, were this the case we would *re-define* the law. Analogously, were  $f$  in the machine to be some function of the state  $S$ , we would have to re-define our machine. Let me be quite explicit with an example. Suppose  $S$  had three states:  $a, b, c$ . If  $f$  depended on  $S$  there would be three  $f$ 's:  $f_a, f_b, f_c$  say. Then if they are

↓	$a$	$b$	$c$
$f_a$	<b>b</b>	$a$	$b$
$f_b$	$c$	<b>a</b>	$a$
$f_c$	$b$	$b$	<b>a</b>

then the transform of  $a$  must be under  $f_a$ , and is therefore  $b$ , so the whole set of  $f$ 's would amount to the *single* transformation:

$a$	$b$	$c$
↓	$b$	$a$

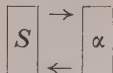
It is clearly illogical to talk of  $f$  as being a function of  $S$ , for such talk would refer to operations, such as  $f_a(b)$ , which cannot in fact occur.

If, then, no machine can properly be said to be self-organizing, how do we regard, say, the Homeostat, that rearranges its own wiring; or the computer that writes out its own program?

The new logic of mechanism enables us to treat the question rigorously. We start with the set  $S$  of states, and assume that  $f$  changes, to  $g$  say. So we really have a *variable*,  $\alpha(t)$  say, a function of the time that had at first the value  $f$  and later the value  $g$ . This

change, as we have just seen, cannot be ascribed to any cause in the set  $S$ ; so it must have come from some outside agent, acting on the system  $S$  as input. If the system is to be in some sense “self-organizing”, the “self” must be enlarged to include this variable  $\alpha$ , and, to keep the whole bounded, the cause of  $\alpha$ 's change must be in  $S$  (or  $\alpha$ ).

Thus the appearance of being “self-organizing” can be given only by the machine  $S$  being coupled to another machine (of one part):



Then the part  $S$  can be “self-organizing” within the whole  $S + \alpha$ .

Only in this partial and strictly qualified sense can we understand that a system is “self-organizing” without being self-contradictory.

Since no system can correctly be said to be self-organizing, and since use of the phrase “self-organizing” tends to perpetuate a fundamentally confused and inconsistent way of looking at the subject, the phrase is probably better allowed to die out.

#### THE SPONTANEOUS GENERATION OF ORGANIZATION

When I say that no system can properly be said to be self-organizing, the listener may not be satisfied. What, he may ask, of those changes that occurred a billion years ago, that led lots of carbon atoms, scattered in little molecules of carbon dioxide, methane, carbonate, etc., to get together until they formed proteins, and then went on to form those large active lumps that today we call “animals”? Was not this process, on an isolated planet, one of “self-organization”? And if it occurred on a planetary surface can it not be made to occur in a computer? I am, of course, now discussing the origin of life. Has modern system theory anything to say on this topic?

It has a great deal to say, and some of it flatly contradictory to what has been said ever since the idea of evolution was first considered. In the past, when a writer discussed the topic, he usually assumed that the generation of life was rare and peculiar,



and he then tried to display some way that would enable this rare and peculiar event to occur. So he tried to display that there is *some* route from, say, carbon dioxide to the amino acid, and thence to the protein, and so, through natural selection and evolution, to intelligent beings. I say that this looking for special conditions is quite wrong. The truth is the opposite—*every* dynamic system generates its own form of intelligent life, is self-organizing in this sense. (I will demonstrate the fact in a moment.) Why we have failed to recognize this fact is that until recently we have had no experience of systems of medium complexity; either they have been like the watch and the pendulum, and we have found their properties few and trivial, or they have been like the dog and the human being, and we have found their properties so rich and remarkable that we have thought them supernatural. Only in the last few years has the general-purpose computer given us a system rich enough to be interesting yet still simple enough to be understandable. With this machine as tutor we can now begin to think about systems that are simple enough to be comprehensible in detail yet also rich enough to be suggestive. With their aid we can see the truth of the statement that *every isolated determinate dynamic system obeying unchanging laws will develop "organisms" that are adapted to their "environments"*.

The argument is simple enough in principle. We start with the fact that systems in general go to equilibrium. Now most of a system's states are non-equilibrial (if we exclude the extreme case of the system in neutral equilibrium). So in going from *any* state to one of the equilibria, the system is going from a larger number of states to a smaller. In this way it is performing a selection, in the purely objective sense that it rejects some states, by leaving them, and retains some other state, by sticking to it. Thus, as every determinate system goes to equilibrium, so does it select. We have heard *ad nauseam* the dictum that a machine cannot select; the truth is just the opposite: every machine, as it goes to equilibrium, performs the corresponding act of selection.

Now, equilibrium in simple systems is usually trivial and uninteresting; it is the pendulum hanging vertically; it is the watch with its main-spring run down; the cube resting flat on one face. Today, however, we know that when the system is more complex and dynamic, equilibrium, and the stability around it, can be



much more interesting. Here we have the automatic pilot successfully combating an eddy; the person redistributing his blood flow after a severe haemorrhage; the business firm restocking after a sudden increase in consumption; the economic system restoring a distribution of supplies after a sudden destruction of a food crop; and it is a man successfully getting at least one meal a day during a lifetime of hardship and unemployment.

What makes the change, from trivial to interesting, is simply the *scale* of the events. "Going to equilibrium" is trivial in the simple pendulum, for the equilibrium is no more than a single point. But when the system is more complex; when, say, a country's economy goes back from wartime to normal methods then the stable region is vast, and much interesting activity can occur within it. The computer is heaven-sent in this context, for it enables us to bridge the enormous conceptual gap from the simple and understandable to the complex and interesting. Thus we can gain a considerable insight into the so-called spontaneous generation of life by just seeing how a somewhat simpler version will appear in a computer.

### COMPETITION

Here is an example of a simpler version. The competition between species is often treated as if it were essentially biological; it is in fact an expression of a process of far greater generality. Suppose we have a computer, for instance, whose stores are filled at random with the digits 0 to 9. Suppose its dynamic law is that the digits are continuously being multiplied in pairs, and the right-hand digit of the product going to replace the first digit taken. Start the machine, and let it "evolve"; what will happen? Now under the laws of this particular world, even times even gives even, and odd times odd gives odd. But even times odd gives even; so after a mixed encounter *the even has the better chance of survival*. So as this system evolves, we shall see the evens favored in the struggle, steadily replacing the odds in the stores and eventually exterminating them.

But the evens are not homogeneous, and among them the zeros are best suited to survive in this particular world; and, as we

watch, we shall see the zeros exterminating their fellow-evens, until eventually they inherit this particular earth.

What we have here is an example of a thesis of extreme generality. From one point of view we have simply a well defined operator (the multiplication and replacement law) which drives on towards equilibrium. In doing so it *automatically* selects those operands that are *specially resistant* to its change-making tendency (for the zeros are uniquely resistant to change by multiplication). This process, of progression towards the specially resistant form, is of extreme generality, demanding only that the operator (or the physical laws of any physical system) be determinate and unchanging. This is the general or abstract point of view. The biologist sees a special case of it when he observes the march of evolution, survival of the fittest, and the inevitable emergence of the highest biological functions and intelligence. Thus, when we ask: What was necessary that life and intelligence should appear? the answer is not carbon, or amino acids or any other special feature but only that the dynamic laws of the process should be *unchanging*, i.e. that the system should be *isolated*. *In any isolated system, life and intelligence inevitably develop* (they may, in degenerate cases, develop to only zero degree).

So the answer to the question: How can we generate intelligence synthetically? is as follows. Take a dynamic system whose laws are unchanging and single-valued, and whose size is so large that after it has gone to an equilibrium that involves only a small fraction of its total states, this small fraction is still large enough to allow room for a good deal of change and behavior. Let it go on for a long enough time to get to such an equilibrium. Then examine the equilibrium in detail. You will find that the states or forms now in being are peculiarly able to survive against the changes induced by the laws. Split the equilibrium in two, call one part "organism" and the other part "environment": you will find that this "organism" is peculiarly able to survive against the disturbances from this "environment". The *degree* of adaptation and complexity that this organism can develop is bounded only by the size of the whole dynamic system and by the time over which it is allowed to progress towards equilibrium. Thus, as I said, every isolated determinate dynamic system will develop organisms that are adapted to their environments. There is thus no difficulty

in principle, in developing synthetic organisms as complex or as intelligent as we please.

In *this* sense, then, *every* machine can be thought of as “self-organizing”, for it will develop, to such degree as its size and complexity allow, some functional structure homologous with an “adapted organism”. But does this give us what we at this Conference are looking for? Only partly; for nothing said so far has any implication about the organization being good or bad; the criterion that would make the distinction has not yet been introduced. It is true, of course, that the developed organism, being stable, will have its own essential variables, and it will show its stability by vigorous reactions that tend to preserve its own existence. To *itself*, its own organization will *always*, by definition, be good. The wasp finds the stinging reflex a good thing, and the leech finds the blood-sucking reflex a good thing. But these criteria come *after* the organization for survival; having seen *what* survives we then see what is “good” for that form. What emerges depends simply on what are the system’s laws and from what state it started; there is no implication that the organization developed will be “good” in any absolute sense, or according to the criterion of any outside body such as ourselves.

To summarize briefly: there is no difficulty, in principle, in developing *synthetic organisms as complex, and as intelligent as we please*. But we must notice two fundamental qualifications; first, their intelligence will be an adaptation to, and a specialization towards, their particular environment, with no implication of validity for any other environment such as ours; and secondly, their intelligence will be directed towards keeping their own essential variables within limits. They will be fundamentally selfish. So we now have to ask: In view of these qualifications, can we yet turn these processes to our advantage?

#### REQUISITE VARIETY

In this matter I do not think enough attention has yet been paid to Shannon’s Tenth Theorem (1949) or to the simpler “law of requisite variety” in which I have expressed the same basic idea (Ashby, 1958, a). Shannon’s theorem says that if a correction-channel has capacity  $H$ , then equivocation of amount  $H$  can be

removed, *but no more*. Shannon stated his theorem in the context of telephone or similar communication, but the formulation is just as true of a biological regulatory channel trying to exert some sort of corrective control. He thought of the case with a lot of message and a little error; the biologist faces the case where the "message" is small but the disturbing errors are many and large. The theorem can then be applied to the brain (or any other regulatory and selective device), when it says that the amount of regulatory or selective action that the brain can achieve is absolutely bounded by its capacity as a channel (Ashby, 1958, b). Another way of expressing the same idea is to say that any quantity  $K$  of appropriate selection demands the transmission or processing of quantity  $K$  of information (Ashby, 1960, b.) *There is no getting of selection for nothing.*

I think that here we have a principle that we shall hear much of in the future, for it dominates all work with complex systems. It enters the subject somewhat as the law of conservation of energy enters power engineering. When that law first came in, about a hundred years ago, many engineers thought of it as a disappointment, for it stopped all hopes of perpetual motion. Nevertheless, it did in fact lead to the great practical engineering triumphs of the nineteenth century, because it made power engineering more realistic.

I suggest that when the full implications of Shannon's Tenth Theorem are grasped we shall be, first sobered, and then helped, for we shall then be able to focus our activities on the problems that are properly realistic, and actually solvable.

#### THE FUTURE

Here I have completed this bird's-eye survey of the principles that govern the self-organizing system. I hope I have given justification for my belief that these principles, based on the logic of mechanism and on information theory, are now essentially *complete*, in the sense that there is now no area that is grossly mysterious.

Before I end, however, I would like to indicate, very briefly, the directions in which future research seems to me to be most likely to be profitable.



One direction in which I believe a great deal to be readily discoverable, is in the discovery of new types of dynamic process. Most of the machine-processes that we know today are very specialized, depending on exactly what parts are used and how they are joined together. But there are systems of more net-like construction in which what happens can only be treated statistically. There are processes here like, for instance, the spread of epidemics, the fluctuations of animal populations over a territory, the spread of wave-like phenomena over a nerve-net. These processes are, in themselves, neither good nor bad, but they exist, with all their curious properties, and doubtless the brain will use them should they be of advantage. What I want to emphasize here is that they often show very surprising and peculiar properties; such as the tendency, in epidemics, for the outbreaks to occur in waves. Such peculiar new properties may be just what some machine designer wants, and that he might otherwise not know how to achieve.

The study of such systems must be essentially statistical, but this does not mean that each system must be individually stochastic. On the contrary, it has recently been shown (Ashby, 1960, c) that no system can have greater efficiency than the determinate when acting as a regulator; so, as regulation is the one function that counts biologically, we can expect that natural selection will have made the brain as determinate as possible. It follows that we can confine our interest to the lesser range in which the sample space is over a set of mechanisms each of which is individually determinate.

As a particular case, a type of system that deserves much more thorough investigation is the large system that is built of parts that have many states of equilibrium. Such systems are extremely common in the terrestrial world; they exist all around us, and in fact, intelligence as we know it would be almost impossible otherwise (Ashby, 1960, d). This is another way of referring to the system whose variables behave largely as part-functions. I have shown elsewhere (Ashby, 1960, a) that such systems tend to show habituation (extinction) and to be able to adapt progressively (Ashby, 1960, d). There is reason to believe that some of the well-known but obscure biological phenomena such as conditioning, association, and Jennings' (1906) law of the resolution of physiological states may be more or less simple and direct expressions

of the multiplicity of equilibrial states. At the moment I am investigating the possibility that the transfer of "structure", such as that of three-dimensional space, into a dynamic system—the sort of learning that Piaget has specially considered—may be an *automatic* process when the input comes to a system with many equilibria. Be that as it may, there can be little doubt that the study of such systems is likely to reveal a variety of new dynamic processes, giving us dynamic resources not at present available.

A particular type of system with many equilibria is the system whose parts have a high "threshold"—those that tend to stay at some "basic" state unless some function of the input exceeds some value. The general properties of such systems is still largely unknown, although Beurle (1956) has made a most interesting start. They deserve extensive investigation; for, with their basic tendency to develop avalanche-like waves of activity, their dynamic properties are likely to prove exciting and even dramatic. The fact that the mammalian brain uses the property extensively suggests that it may have some peculiar, and useful, property not readily obtainable in any other way.

Reference to the system with many equilibria brings me to the second line of investigation that seems to me to be in the highest degree promising—I refer to the discovery of *the living organism's memory store*: the identification of its physical nature.

At the moment, our knowledge of the living brain is grossly out of balance. With regard to what happens from one millisecond to the next we know a great deal, and many laboratories are working to add yet more detail. But when we ask what happens in the brain from one hour to the next, or from one year to the next, practically nothing is known. Yet it is these longer-term changes that are the really significant ones in human behavior.

It seems to me, therefore, that if there is one thing that is crying out to be investigated it is the physical basis of the brain's memory-stores. There was a time when "memory" was a very vague and metaphysical subject; but those days are gone. "Memory", as a *constraint* holding over events of the past and the present, and a *relation* between them, is today firmly grasped by the logic of mechanism. We know exactly what we mean by it behavioristically and operationally. What we need now is the provision of adequate



resources for its investigation. Surely the time has come for the world to be able to find resources for *one* team to go into the matter?

### SUMMARY

Today, the principles of the self-organizing system are known with some completeness, in the sense that no major part of the subject is wholly mysterious.

We have a secure base. Today we know *exactly* what we mean by "machine", by "organization", by "integration", and by "self-organization". We understand these concepts as thoroughly and as rigorously as the mathematician understands "continuity" or "convergence".

In these terms we can see today that the artificial generation of dynamic systems with "life" and "intelligence" is not merely simple—it is unavoidable if only the basic requirements are met. These are not carbon, water, or any other material entities but the persistence, over a long time, of the action of any operator that is both unchanging and single-valued. *Every* such operator forces the development of its own form of life and intelligence.

But will the forms developed be of use to *us*? Here the situation is dominated by the basic law of requisite variety (and Shannon's Tenth Theorem), which says that the achieving of appropriate selection (to a degree better than chance) is absolutely dependent on the processing of at least that quantity of information. Future work must respect this law, or be marked as futile even before it has started.

Finally, I commend as a program for research, the *identification of the physical basis of the brain's memory stores*. Our knowledge of the brain's functioning is today grossly out of balance. A vast amount is known about how the brain goes from state to state at about millisecond intervals; but when we consider our knowledge of the basis of the important long-term changes we find it to amount, practically, to nothing. I suggest it is time that we made some definite attempt to attack this problem. Surely it is time that the world had *one* team active in this direction?

## REFERENCES

1. W. ROSS ASHBY, The physical origin of adaptation by trial and error, *J. Gen. Psychol.* **32**, pp. 13-25 (1945).
2. W. ROSS ASHBY, Principles of the self-organizing dynamic system. *J. Gen. Psychol.* **37**, pp. 125-8 (1947).
3. W. ROSS ASHBY, *An Introduction to Cybernetics*, Wiley, New York, 3rd imp. (1958, a).
4. W. ROSS ASHBY, Requisite variety and its implications for the control of complex systems, *Cybernetica*, **1**, pp. 83-99 (1958, b).
5. W. ROSS ASHBY, The mechanism of habituation. In: *The Mechanization of thought Processes*. (Natl. Phys. Lab. Symposium No. 10) H.M.S.O., London (1960).
6. W. ROSS ASHBY, Computers and decision-making, *New Scientist*, **7**, p. 746 (1960, b).
7. W. ROSS ASHBY, The brain as regulator, *Nature, Lond.* **186**, p. 413 (1960, c).
8. W. ROSS ASHBY, *Design for a Brain; the Origin of Adaptive Behavior*, Wiley, New York, 2nd ed. (1960, d).
9. L. VON BERTALANFFY, An outline of general system theory, *Brit. J. Phil. Sci.* **1**, pp. 134-65 (1950).
10. R. L. BEURLE, Properties of a mass of cells capable of regenerating pulses, *Proc. Roy. Soc.* **B240**, pp. 55-94 (1956).
11. W. R. GARNER and W. J. MCGILL, The relation between information and variance analyses, *Psychometrika* **21**, pp. 219-28 (1956).
12. R. C. JEFFREY, Some recent simplifications of the theory of finite automata. Technical Report 219, Research Laboratory of Electronics, Massachusetts Institute of Technology (27 May 1959)..
13. H. S. JENNINGS, *Behavior of the Lower Organisms*, New York (1906).
14. A. J. LOTKA, *Elements of Physical Biology*, Williams & Wilkins, Baltimore (1925).
15. J. G. MARCH and J. A. SIMON, *Organizations*, Wiley, New York (1958).
16. K. H. PRIBRAM, Fifteenth International Congress of Psychology, Brussels (1957).
17. C. E. SHANNON and W. WEAVER, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1949).
18. G. SOMMERHOFF, *Analytical Biology*, Oxford University Press, London (1950).

**R. W. SPERRY**

*California Institute of Technology*

## ORDERLY FUNCTION WITH DISORDERED STRUCTURE

I shall try to make my comments as brief as possible, for I suspect that anything that I can contribute here is likely to be toward the periphery of our symposium. I was under the misapprehension that a majority of the conferees would be engineers interested in devising ingenious circuits for servomechanisms and the like, who would have a timely concern for the problem of building compact control boxes that would continue to control after being nicked by a meteorite, peppered by radiation, or even pierced by a few bullets. With this in mind, I plan to start by recalling the always remarkable capacity of the cerebral machinery to continue to operate effectively and in relatively orderly manner in the face of extensive damage to the structure.

As most of you know, when large lesions are inflicted in the brain, or even a whole lobe removed, like the temporal or frontal, the cerebral machinery may continue to operate in such good fashion that most of us could not detect any functional change. It often requires very sensitive tests indeed to detect even major unilateral damage, while for some brain areas the tests have yet to be devised that will expose the defect.

Another point in regard to the preservation of function after large lesions is the fact that the nearby cortical networks generally continue to work effectively right to the very edge of the damaged tissue. For example, the outline of a blind spot in the visual field produced by a cortical lesion may be mapped by having the subject follow a moving object as it approaches from different directions to disappear into the blind area.

Remarkable preservation of function is seen also after diffuse damage like that found in approaching senility or after disease

where vast numbers of cells may be eliminated in scattered arrays all through the system. So long as a moderate percentage of the remaining neuronal elements survive and regain a healthy state, organized function is recovered.

In some of our laboratory work we have had occasion to riddle the visual and surrounding cortex of the cat with pins or needles of tantalum wire. Dozens of pins were poked into and through the cortical tissue in pin cushion fashion until the whole visual cortex was saturated and our patience exhausted. The biologically inert inserts were simply left in the brain. In subsequent testing of near-threshold pattern discrimination the visual performance was practically as good as before.

The ability of the cortex to function close to the edge of lesions is most strikingly illustrated perhaps in some experiments that involved multiple intersecting knife cuts. Cross-slicing of cortical areas with numerous criss-cross knife cuts extending vertically through the depth of the cortex and placed only a few millimetres apart was found to have very little effect on organized function in the scored area. Figure 1a shows the brain of a cat in which the visual cortex has been subdivided in this way. The animal appeared quite blind for the first four days after surgery, but as the post-operative edema began to clear, vision reappeared and continued to improve during the next month and a half until the animal's level of performance on pattern discrimination tests was within the normal range. This cat was able to pick out the equilateral triangle shown in the center of Figure 1b when it was paired with any one of the other triangles surrounding it, under conditions controlled for position, odor and other non-visual cues.

In contrast to this preservation of organized function in the living brain, recall what happens in most of our man-made circuits as the result of a single burnt-out tube, a single broken wire, a single short circuit and so on. With an eye to future design principles for incorporating built-in repair, compensating, corrective and other self-organizing devices, it might be apropos to point out some of the design features of the living brain that enable it to continue to function in the face of extensive structural damage.

Although a complete explanation is still far out of reach, of course, there are a few things one can say. In the first place, it probably is not any single construction principle that is responsible.

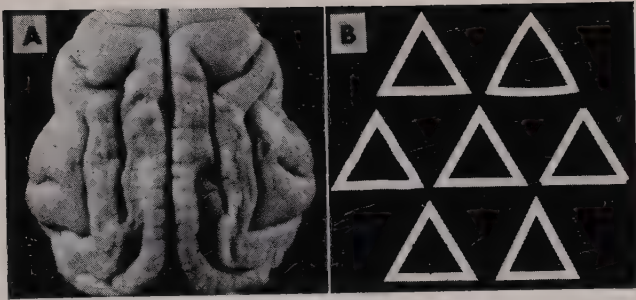


FIG. 1





There would seem to be at least several now recognizable and probably many yet to be worked out. I will run over a few points that come quickly to mind in this connection without particular regard to their order or relative importance and I am sure that others here more directly concerned with these problems will be able to add to the list.

There is, for example, the simple factor of multiple connection. Each fiber of the central nervous system usually terminates not on one terminal connection or two, but on ten, or more typically, hundreds or even thousands of downstream elements. Similarly, each neural fiber picks up from tens or hundreds of terminals at the other end.

A second factor is that of the wide overlap in the connections of nearby elements. For any two fibers running side by side in a given nerve cable or tract, the overlap among their multiple hookups at each end may run roughly on the order of say 60 per cent or higher. With this extensive overlap it is evident that a sizeable number of elements may be damaged or eliminated from the complex and the loss remain unnoticed. In some parts of the central network, or in all parts in some species, the neural elements have regenerative capacity with the result that a break in the central circuit is promptly repaired. The new connections formed in the regenerative process are as orderly and neatly arranged as were the original, because the same orderly forces of growth persist in these adults and operate in regeneration. It is as if each neuronal element has a kind of built-in navigation plan or homing instinct of its own that enables it to recover its appropriate hookups.

This kind of thing is illustrated in the studies on optic nerve regeneration already referred to by Dr. McCulloch earlier in the conference. After the nerve has been completely cut and there has been a thorough scrambling of fibers in the nerve scar, a given fiber approaching the point within the brain where the optic tract divides into medial and lateral branches, will select the correct branch according to whether its locus of origin was in the dorsal or the ventral quadrant of the retina. The medial and lateral bundles skirt around the medial and lateral edges respectively of the optic lobe of the midbrain, giving off fibers along the way as they progress from anterior to posterior poles of the lobe. A given regenerating fiber after choosing the correct branch along

the periphery of the lobe, then chooses correctly the proper point at which to enter the lobe. Fibers from the central retina enter near the center of the lobe, those from the nasal retina enter at posterior levels, and those from the temporal retina at anterior levels.

Once inside of the optic lobe or tectum, the fibers run from the periphery toward the center in a superficial parallel layer. Again, they are faced with numerous alternative possibilities with regard to the point at which they will each leave the parallel layer to dip downward into the underlying plexiform layer. Fibers from the peripheral retina dip quickly near the periphery of the tectum. Those from the central retina remain in the parallel layer to cross the peripheral portions of the tectum and dip downward only after they reach the central tectum. In the plexiform layer each fiber branches and forms multiple contacts with the dendrites and cell bodies of the tectal neurons. Again, we must infer a selective choice in synaptic formation among the numerous possible alternatives. The latter inference is based largely on the fact that the trained discrimination of red, blue, yellow, green, gray and other colors from one another in a large variety of combinations has been shown to be restored in its original form after optic nerve regeneration.<sup>(1)</sup> (It is of further interest here that the newly regenerated pattern of central synapses is adequate for reactivation of the memory for visual discrimination habits learned with the original synaptic connections prior to optic nerve section. And incidentally, the restored color discrimination performance survives ablation of forebrain plus cerebellum.)

The foregoing description is based on studies in fishes. The central nervous system of the mammal, of course, is not capable of much functional regeneration. Even in the mammalian cortex, however, it is conceivable that the detailed pattern of synaptic connections is in a state of continuous flux and that the normal pattern is prevented from drifting into a state of randomized chaos by the constant operation of specific biochemical forces within the individual elements similar to those responsible for orderly central regeneration in the lower vertebrates.

A fourth factor tending to preserve organized activity may be described as the multiple reinforcement of any given function from numerous different sources, any one of which may in itself

be capable of sustaining the activity. As a simple example, take a locomotor gait, such as the running, trotting, or galloping of a horse or dog. It appears that the feedback of any single one of the four limbs in action is sufficient to keep the whole system running. In fact, experiments dealing with the control of the locomotor gait in amphibians show that the sensory nerves of the entire spinal system may be eliminated except for those in just one or two of the dorsal roots and these will be sufficient to keep the locomotor gait going.<sup>(5)</sup> Furthermore, it does not matter which two are left; any two will do. Thus, in the normal condition there is an extensive multiple reinforcement of the same pattern from many different sources. Presumably a given cerebral pattern may also be activated and sustained by numerous "mental associations". Redundancy is one element here, perhaps, but this kind of multiple reinforcement involves, of course, a good deal more than simple redundancy.

Part of the problem of maintaining organized activity lies in controlling and preventing disrupting and disturbing influences from other unrelated functions. In this respect functional control factors like "reciprocal inhibition" and "inhibitory surround" must play important roles.

Another point, a bit less relevant perhaps but the very essence of self-organization, is that the guidance and control of sequential activity in the nervous system in general does not depend upon a central schedule or clock for ticking off each act at its proper time—excepting perhaps in the case of very simple cycles or for extremely rapid sequences like consecutive finger movements in piano playing. The much more general method of cerebral control is to have each act set off and sustained only by those conditions for which the act is appropriate and which usually are a product of the preceding act in the sequence. In this way the sequence runs itself. Further, with this type of control many kinds of disturbances and changes in speed and timing and the like may occur without disrupting the whole pattern.

The circuitry by which all this is achieved may be likened to a vast collection of negative feedback systems, multiply interlocked with one another and broadly organized throughout on a hierarchical plan, that is, a tremendous network of interlocked checks and balances permeated by homeostatic loops within loops.

A minor safety factor may be seen in the arrangement of the cortical circuits along vertical rather than horizontal dimensions. This accounts in part for the ability of small portions of cross-cut areas and/or remnants adjacent to large lesions to continue their orderly function.

Another simple and obvious safety factor is that of right-left duplication. Brain centers, much like kidneys, lungs, gonads and other organs, are furnished in matched pairs. Damage to or complete loss of one member often is not critical because the function involved can be handled by the other member on the opposite side. This would seem to apply to the frontal and temporal lobes of most mammals and to many of the hypothalamic and other homeostatic control centers.

Studies involving hemispherectomy and commissurotomy suggest that the mammalian brain is in many ways essentially two separate half-brains, that is, two whole control systems, each capable of carrying on independently of the other. Similarly, the chances that a black control box sent into space will be put out of commission by a shell, may be reduced by half if it has two independent and separated control circuits properly oriented.

This raises another problem in the circuitry of higher controls. If one has two complete sets of higher level controls, is there any advantage to building cross connections between these? The answer would seem definitely to be "yes" if the example of the mammalian brain means anything. In the mammalian brain the largest single fiber tract by far is the corpus callosum—the system of fibers that interconnects the neocortex of the two cerebral hemispheres.

Which brings us to the "riddle of the corpus callosum", certain aspects of which have definite relevance to our topic of self-organization. First perhaps it should be pointed out that it is no longer the enigmatic riddle that it was a little over ten years ago when Dr. McCulloch somewhat facetiously but not without good justification stated that the only known function of this structure was to aid in the transmission of epileptic seizures from one side of the body to the other. About the same time Lashley, in a similar vein, used to suggest that the only apparent function for the corpus callosum seemed to be mainly mechanical in nature, i.e. to prevent the two hemispheres from sagging too far apart.



Today it is probably fair to say that more is known about the anatomy and the physiology of this particular cortico-cortical fiber system than about any other in the brain. Embedded deeply at each end in the cortical networks, any information about its connection plan and function is bound to have implications for the secrets of cortical organization in general. Most of this new information regarding the function of the corpus callosum and other cerebral commissures has been obtained from animal studies of the past six or seven years, which I will not attempt to go into in any detail here since they have only recently been reviewed elsewhere.<sup>(3,4,6)</sup>

For our present purposes it will be enough to point out that in the absence of the cerebral commissure, i.e. following their surgical section in the midline, a cat or monkey is in many respects like an animal with two brains in the place of one. Having two instead of one brain seems to make little difference under most ordinary conditions. In fact, the cat or monkey even after deep bisection of the brain through the quadrigeminal plate and the cerebellum is hardly distinguishable from its normal cagemates under ordinary circumstances. With the proper testing conditions, however, wherein one can stimulate and train each hemisphere independently, it is possible to show that in the absence of the cerebral commissures each hemisphere has its own perceptual, learning and memory processes, i.e. its own cognitive or psychic system. It is as if neither of the separated hemispheres has any longer any direct awareness of the mental activity of the other, nor any direct memory of anything experienced by the other subsequent to section of the commissure. The control of the animal's behavior under these conditions may be governed predominantly from one of its half brains if one hemisphere is markedly dominant, or the control may shift from one to the other and back in an alternate fugue-like fashion, or the two hemispheres may continue to operate simultaneously in parallel so long as their lower level effects are harmonious. With proper testing conditions it can be shown that the two half brains can operate simultaneously in the learning of separate—even conflicting—discrimination habits.

There are some advantages perhaps in having the two cerebral control systems working independently, but presumably these are

outweighed by the disadvantages. In the separated condition neither hemisphere benefits from the learning and experience of the other. In a sense, the commissures thus serve to keep each hemisphere up to date on what is new in the other. They appear also to facilitate certain sensory-sensory and sensory-motor integrations—as for example volitional visual uses of the hand across the midline of the visual field.

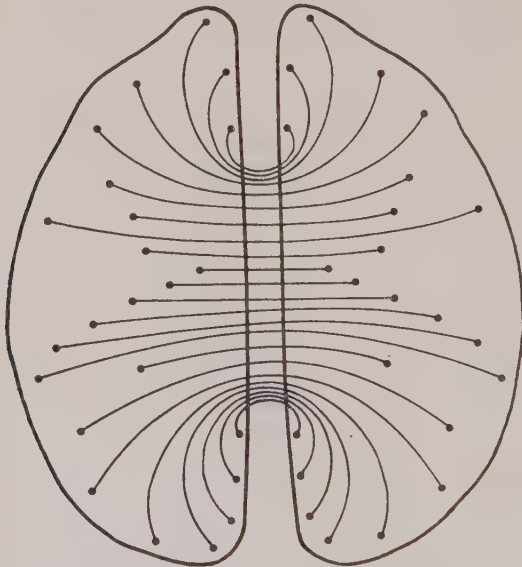


FIG. 2. Homotopic principle of callosal cross-connections between corresponding loci of right and left hemisphere.

There is still one aspect of the corpus callosum that remains something of a real riddle: namely, the problem of the meaning of the bilateral symmetry of its cross connections. Physiological and anatomical studies indicate that the great majority (though not all) of the fibers of the callosum are homotopic, that is, the fibers arising from a given point in one hemisphere project across the midline to the same point in the opposite hemisphere.<sup>(4)</sup> This homotopic principle is illustrated in Figure 2. More than this, it seems that within these symmetrical loci the fiber systems arising



from different layers of the cortex tend to terminate predominantly in the same layer on the opposite side.<sup>(6)</sup>

The question, then, is this: "What good, from an engineering standpoint, is served by having this tremendous system of fiber interconnections linking identical points in the two hemispheres?" It would appear that any activity pattern in one cerebral cortex

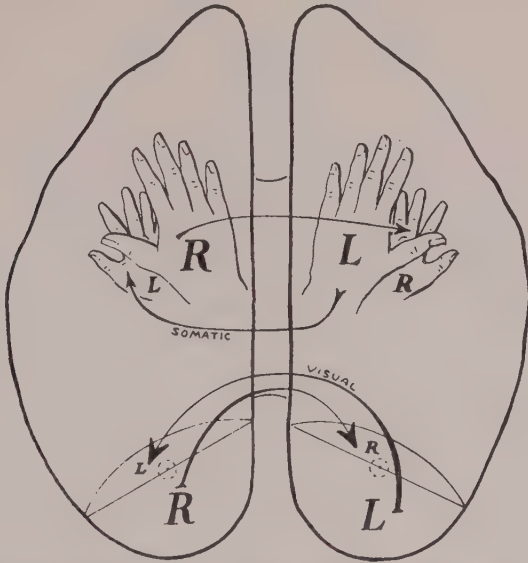


FIG. 3. Hypothetical alternative to homotopic cross-connections, the principle of "supplemental complementarity".

would tend to be duplicated by the corpus callosum in the other—that the corpus callosum would thus act as a symmetrizing influence on all cerebral activity.

What good would it do to have cross connections between the representations of identical points in right and left halves of the visual fields? The same can be asked with respect to corresponding points on right and left sides of the body. Except in rare instances when the two halves of the visual field or the body contacts on right and left sides are mirror images, it would seem that such interaction could only make for confusion.

Instead of symmetric cross connections it would seem to make

more sense offhand to have the cross connections more supplemental or complementary in nature as suggested in Fig. 3. This figure illustrates an older notion I had about the corpus callosum that has been reactivated by this conference. This plan is quite speculative with little evidence to support it. Instead of having the right half of the visual field reciprocally cross connected to the left half, in this scheme each half is cross projected so as to supplement the direct projection. This gives in each hemisphere a whole visual field—or its representation in a subsequent step of the data processing system. The same applies to the representation of the body surface and other sensory fields, and also to deeper association activities of the hemisphere. In an effort to keep up with our modern physicists, I used to call this the “principle of supplemental complementarity”.

Actually the evidence is not yet sufficient to completely rule out some scheme of this kind. There is something puzzling and irregular about the callosal cross connections between the visual areas of the cortex. In the cortical map of the body surface, the contralateral and ipsilateral representations seem to be in register so that a supplemental scheme would be difficult to differentiate from one of symmetrical reciprocity. In the past investigators of the callosum were not searching with such questions in mind, but the time is now ripe for more detailed investigations of the meaning of the cross connection pattern. It is accepted that a minority of the callosal connections are heterotopic in nature and in some instances the fibers from a cortical area project to quite different regions on the other side.

Possibly both symmetric and supplemental as well as other schemes are represented in the corpus callosum. Perhaps the symmetric system serves as a great detector of all-important asymmetry. In any case, you see the problem. It is a circuitry problem essentially, and the kind of thing that may well be answered more quickly in the developing logics of artificial intelligence than in the probing of the neurophysiologist.

#### REFERENCES

1. H. L. ARORA and R. W. SPERRY, Studies on color discrimination following optic nerve regeneration in the cichlid fish, *Astranotus ocellatus*, *Anat. Rec.* **131**, p. 529 (1958).

2. D. ATTARDI and R. W. SPERRY, Central routes taken by regenerating optic fibers, *The Physiologist* 3, p. 12 (1960).
3. F. BREMER, J. BRIHAYE and G. ANDRE-BALISAUX, Physiologie et pathologie du corps calleux, *Arch. Suisses Neurol. Psychiat.* 78, p. 31 (1956).
4. B. Grafstein, Organization of callosal connections in suprasylvian gyrus of cat, *J. Neurophysiol.* 22, pp. 504-15 (1959).
5. J. GRAY and H. W. LISSMAN, Further observations on the effect of deafferentation on the locomotory activity of amphibian limbs, *J. exp. Biol.* 23, pp. 121-32 (1946).
6. R. E. MYERS, Interhemispheric interconnections between occipital poles of the monkey brain, *Anat. Rec.* 136, p. 249 (1960).
7. R. W. SPERRY, Some developments in brain lesion studies of learning, *Fed. Proc.* (in press).
8. R. W. SPERRY, Cerebral organization and behavior, *Science* 133, p. 1749 (1961).

### DISCUSSION

DR. YOVITS: What happens when the individual cortical halves give completely opposite orders to the animal body? Does the animal eventually learn to disregard one order?

DR. SPERRY: Yes, in general one or the other hemisphere tends to dominate the lower centers and motor outflow and these consistently follow the dictates of the dominant control. Any incompatible orders coming down from the other hemisphere tend to be ignored or inhibited. It is one of the important features of brain organization that it does not get confused. It is a this-or-that proposition: either this or that but never "thiast" or other mixtures. There are exceptions, but they are the exceptions that prove the rule. There is some built-in mechanism including perhaps "reciprocal innervation" and "inhibitory surround" as mentioned above, that makes for unified activity. Once an activity pattern gains the ascendancy it wipes out thousands of other possibilities. If you can build that into a machine you can perhaps begin to get some of the "many-machines-in-one" kind of versatility we see in the brain.

DR. SHERWOOD: I think the point should be made here which arises from the surgical operations or experiments in man which someone in England has carried out. He has done a series of hemispherectomies for birth brain damage to children who keep having seizures in one hemisphere. Now the point that arises from this is that while the whole hemisphere is ablated with the exception of the basal ganglia—they are left—now, I have seen some of these children and their two-point discrimination on the decentralized side improves after hemispherectomy, as does their gait. It shows, I think, that the damage of a machine that works wrong, is worse than not having the machine.

CHAIRMAN BOWMAN: Is there any further discussion of Dr. Sperry's paper?

DR. MCCULLOCH: Yes, please. My statement concerning the corpus callosum was not made lightly. I have worked from, oh, the early thirties to the mid-forties on the anatomy and physiologically detectable anatomy of the system and at that time there were a large number of men who had had the corpus callosum, and the anterior commissures in some cases, cut in order to prevent spread of seizures from one hemisphere to the other, and if they cut the anterior commissure as well as the corpus callosum, they generally stopped having seizures that passed all the way over. There were some exceptions. Now the most interesting thing about the corpus callosum, to my

mind, is not this homeo-projection. It is that there exists, let's say, from area 8, from area 6, and from some parts of the post-central cortex, much more widely distributed afferents to the opposite hemisphere. There are more of these heterotopic fibers from areas that are farther removed from direct sensory and direct motor projections, and I think this is probably what you are looking for. My feeling is that it is more sensibly organized. The other thing is also not quite a frivolous remark. When you cut the two hemispheres, you form a caricature of the Post logic. So long as the two are coupled, point for point, you may remove, you may simplify the logical structure of the machine. That is the very fact that you have made symmetrical Venn's out of them, in miniature sections, that is clear. You have got a Post logic.

**R. L. BEURLE**

*English Electric Valve Co. Ltd.*

## FUNCTIONAL ORGANIZATION IN RANDOM NETWORKS

### *Summary*

Some years ago an attempt was made to show how a randomly connected network of neuron cells could perform a useful function as a part of the central nervous system of a living organism.

During the intervening period certain new facts have come to light and various constructive comments have been made on this approach. It now seems opportune to present a résumé of the original argument, revised in some respects in accordance with these facts.

### 1. INTRODUCTION

It was the aim of an earlier analysis<sup>(1)</sup> to show that a randomly connected mass of neuron cells could account for some of the plasticity in behavior in the organism of which it forms a part. Although the initial structure of the mass is determined by chance factors, such as a mass of cells may, by interacting with the environment, gradually build up an internal organization giving it the power to choose behavior having survival value.

The ability to do this depends principally on the properties of the cells of which the mass is composed. It is proposed to take this opportunity first to outline the properties that were originally assumed and then to comment on two aspects of the original theme which appear to call for further development. The first of these is the need for, and the nature of the stabilizing mechanism controlling activity in the cell mass. The second is the possibility of a close inter-relationship between long-term memory and short-term memory.

### 2. THE BASIC CELL PROPERTIES

The original analysis, which has already been referred to,<sup>(1)</sup> was undertaken for several reasons.

(i) The fact that most existing theories of learning were rather



inadequate to explain the remarkable ability of the brain to sift through a vast amount of input information and yet to make snap decisions with a fair degree of survival value. The brain endows us with a remarkable ability to learn and use facts, and languages, and the complicated decision making techniques which have been evolved by mankind over many generations, and it gives us common sense.

(ii) A theory was required to form a bridge between the various empirical facts relating to the different forms of learning, on the one hand, and the anatomical and physiological evidence on the other. Different groups have stressed the importance of trial and error learning, intuitive behavior, contemplative thought processes, logic, etc., without any attempt to present these as different aspects of some whole.

(iii) In the brain there are cell masses (e.g. the cortex) in which there appears to be a very large random factor in the distribution and interconnection of neuron cells and fibers. There is consistency in parameters such as density of distribution of cells and fibers, but in many regions no evidence has been found to suggest that individual connections follow any precise or uniform pattern.<sup>(2)</sup>

(iv) Since the hereditary information which determines the structure of these cell masses is contained in the relatively small compass of the chromosomes one would expect to find a structure which can be described and specified reasonably economically. The types of structure that require the minimum of information to describe them are, on the one hand, those in which there is a high degree of order and, on the other hand, those in which the detailed structure is left to chance. As there is little evidence of any detailed ordering of interconnections in many regions, and no evidence of a high degree of order, one is inclined to think that the second alternative is more likely.

The obvious question arises, is it possible to explain at least a part of the behavior of living organisms in terms of the activity in a randomly connected network? The difficulty in answering this question, in relation to the cell masses in the cortex, lay in the lack of accurate and complete information about the properties of cortical neurons. To overcome this, the approach adopted was to consider the activity in a hypothetical random network of cells, for which precisely defined properties could be postulated. These



properties were deliberately chosen to be as close as possible to what were thought at the time to be properties of cortical neuron cells. Accordingly the following properties were assumed:

1. A random spatial distribution of cells.
2. Interconnecting fibers branching and terminating randomly within the type of probability distribution measured by Sholl.<sup>(2)</sup>
3. Connections are established by chance between neighboring cells, the probability of connection between any two cells decreasing with their distance apart as one might expect it to from the work of Sholl on the density of associated fibre systems.
4. Each cell, when it becomes active, excites those neighboring cells to which its axon system is connected.<sup>(3)</sup>
5. Cells integrate the received excitation and are capable of becoming active if this exceeds a threshold.<sup>(3)</sup>
6. Each cell, when it becomes active, undergoes a slight change which makes it more easy to activate on a future occasion.<sup>(3)</sup>
7. After a refractory period, following activity, a cell regains sensitivity and can again take part in the corporate activity of the mass.<sup>(3)</sup>

### 3. BULK PROPERTIES OF THE MASS

#### (a) *Stability*

It will be noted that no mention has been made of any inhibitory influence between cells. At the time, although there was evidence for inhibition in peripheral nerve cells, there was very little evidence as to how, if at all, inhibition might function in the cortex, and biologists were reluctant to accept the assumption of some inhibitory influence without evidence. For this reason great care was taken to avoid the assumption of any direct inhibitory influence. As a result, when the bulk properties of the cell mass were examined it was found that activity in it was of a very unstable nature. If activity was started in the mass, it would either die away rapidly to zero or, if the stimulus was stronger, it would increase more and more rapidly until all the cells in the mass were involved in what was referred to in the earlier paper as a saturated surge of activity. (See Fig. 1.) If one has a network of interacting elements in which the sole interaction is excitatory, then the activity of the network will necessarily tend to instability. Now, in actual cortical

material one does not appear to encounter this all-powerful, all or nothing, saturated surge of activity. Moreover, it is difficult to see how this sort of response could mediate any fine subtleties of behavior whereas, if there is some means of stabilizing or controlling the degree of activity at some intermediate level, to give an "unsaturated" surge of activity (Fig. 2), a wealth of interesting behavior results. It is thus essential to consider the question of stability.

In the earlier analysis, in view of the care that had been taken to avoid postulating direct inhibitory effects, it was only possible to introduce stabilization by a poly-synaptic chain of events. This was, in fact, the most unsatisfactory feature of that treatment. If we may assume that some direct inhibitory influence exists, then this can account for stabilization of the activity of the mass and it becomes unnecessary to introduce the rather clumsy concept of a poly-synaptic chain. Biologists now seem much more ready to credit the possibility of direct inhibition, mediated either as a synaptic effect or as a field effect. Either type of effect would be sufficient to account for the stabilization of activity but, as yet, there is insufficient evidence to show with any certainty how it comes about.

Nevertheless, it is only by asking questions that we discover facts, and one is prompted to point out that there is an interesting anisotropy introduced by the apical dendrites. It is known that when a large proportion of cells in one region of the cortex are active, there is a potential difference built up across the cortical layer in the direction of the apical dendrites. It is also known that, if a potential difference is applied artificially, the cortex may either be excited or inhibited depending on the direction of the applied potential difference. Is it possible, one may ask, for these two effects to combine to provide the stabilization we have been discussing?

#### (b) *Attenuation*

If, with some stabilizing influence present we test the response, we find that, as before, a very small stimulus, given once, merely produces a small local disturbance that dies away quickly. If, however, we stimulate more strongly or persevere with our weak stimulus, we may initiate a stronger, but stable response that may

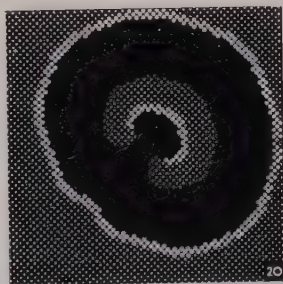
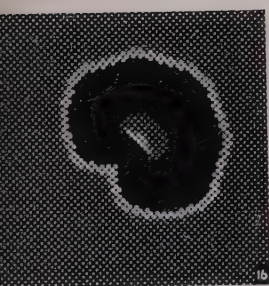
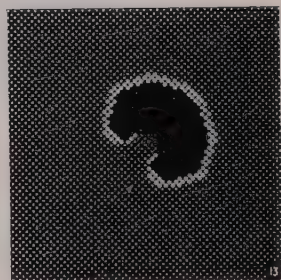
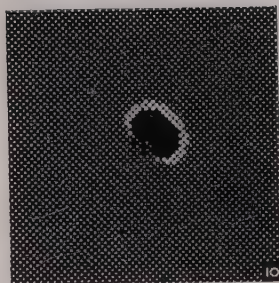
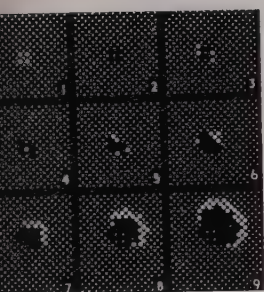


FIG. 1

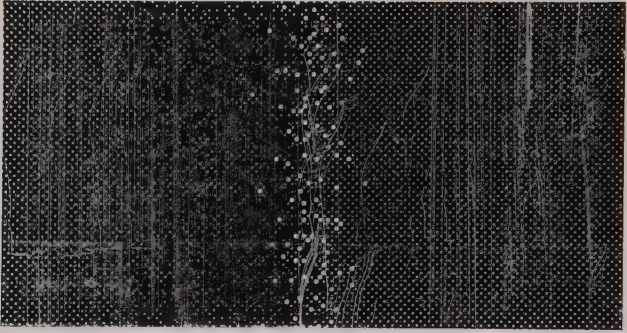


FIG. 2



FIG. 5



be maintained for some time and may spread. This behavior is strongly reminiscent of the reaction of the cortical surface under electrical stimulation. It is interesting to note also, that this same sort of excitability is shown by a colony of coral polyps, one of the most elementary "nervous systems".<sup>(4)</sup>

The nature of the stabilizing influence, and the particular way in which it intervenes, can have a marked effect on the form of the surge of activity. Under the action of the stabilizing influence, the surge waveform may be very different from the waveform derived in the earlier paper, for a mass of cells without any stabilization. One might even be left with a diffuse surge of activity in place of a clear-cut wave, but this does not invalidate the main principles which that paper aimed to point out.

The quantitative aspects of the excitability of the mass, and their dependence on past experience, are the most important features in relation to learning and memory. In particular, as a result of the sixth cell property, the attenuation of the cell mass for a particular surge of activity decreases slightly each time that surge travels through the medium. It is this that gives us the first basic adaptation process of trial-and-error learning.

#### 4. FUNCTIONAL PROPERTIES OF THE MASS IN AN ORGANISM

##### (a) *Learning and Memory—Trial and Error*

It is not hard to see that the dependence of attenuation on past experience provides a simple basis for trial-and-error learning. This was demonstrated in the original treatment with the aid of Fig. 3, which shows a hypothetical organism. This diagram was deliberately made as simple as possible in order to demonstrate that in principle the ability to learn by trial and error is a property of the mass of cells. Certain sensory inputs are defined by the discriminator as inherently satisfactory or unsatisfactory, and the discriminator accordingly aids or disturbs the passage of activity through the mass. It will be evident that, even though the network is randomly connected, the only forms of activity able to continue undisturbed will be those that result in a motor output which elicits a reaction from the outside world which is defined as satisfactory. Only these forms of activity will be able to build up paths

of low attenuation through the mass and, as long as the outside world remains consistent, these low attenuations will enable the organism to reproduce satisfactory responses to familiar environments in the future.

This ability to discover a suitable response to a new environment is the first stage of learning by trial and error. It is also the first

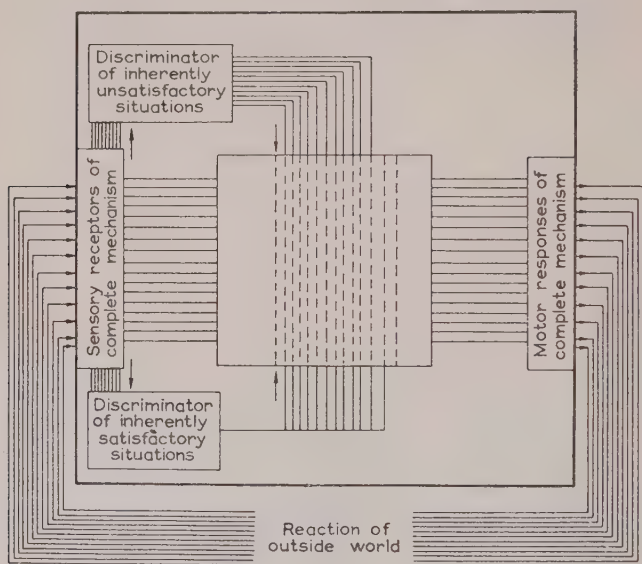


FIG. 3.

stage of recognition, for the ability to choose a suitable response to a familiar environment is tantamount to recognition of that environment. At the same time, since the low attenuation paths will persist, we have a simple form of long-term memory, namely memory of the behavior which has been found to produce satisfactory results in certain familiar environments. It is important to note that even a minute change in threshold, as long as it occurs in a sufficiently large number of cells, is sufficient to bring about the reduced attenuation. We do not need to look for a major modification of any cells.



*Abstraction of relevant features in the environment.* Now in real life things are generally not as simple as in elementary experiments on learning. Situations which have to be dealt with are seldom identical in every respect with situations which have been encountered in the past. In this sense every environment is a new one, and therefore every response, even though it is based on previous learning, must be regarded as a tentative one subject to further reinforcement or suppression according to its success or failure in the future. The ratio of success to failure depends on the extent to which it is the relevant features in the environment that are stimulating the response, and not the irrelevant features. The second stage of trial-and-error learning must therefore be the gradual growth of the ability to isolate just those particular relevant features of the environment for which the learned response is appropriate, and of the ability to ignore irrelevant and variable background detail. This necessarily involves further trial and error, for it is only by experiment that it is possible to find out which recurrent features are relevant to the choice of the response.

*What happens in the mass of cells during the second stage?* Suppose we present it with a series of different environments, all of which contain certain consistent features for which a particular response is appropriate. The mass will be receiving excitatory stimuli from three sources:

- (i) A constant stimulus from the consistent features.
- (ii) A variable stimulus from the irrelevant background.
- (iii) Any randomly scattered excitation which may be present at the same time.

The excitation from (ii) and (iii) may be lumped together as being variable, and irrelevant to the choice of the particular response we are considering. As a result of this scattered excitation, there will be a continual fluctuation in the path of activity through the mass of cells, so that it will only be gradually that, by trial and error, the tendency of the scattered excitation or "noise" to deflect the activity towards unsatisfactory responses will be eliminated. The process of adaptation must not be pushed too fast. If the noise content is too great the thread will be lost, and the organism will be back at the blind, groping, trial-and-error stage.

*The growth of selectivity.* As this process continues the presence of noise together with the stabilizing influence tends to help eliminate some of the redundant activity that will inevitably arise at first in a randomly connected network. The trial-and-error learning involves a cycle of events in which the generation of the satisfactory response is essential, and cells which do not contribute to the production of this response will fire less frequently and thus tend to drop out. There is thus a gradually increasing economy of effort, until the activity becomes restricted to a number of parallel interrelated paths between which the activity fluctuates according to the dictates of the noise content of the scattered excitation.

Taken slowly, the process can take care of an increasing excitatory "noise", the magnitude of which may ultimately be much greater than the particular relevant features which cause the response. With experience, involving trial and error, discrimination will grow until the cell mass behaves like a high selectivity filter responding perhaps to some minute but highly significant set of features, even though the bulk of the environment is variable. Whenever this familiar set of features is present, this particular response will be produced as the initial tentative response, even if the background against which it is presented is unfamiliar. This is the beginning of generalization, environments are classed together as meriting one particular response if they all contain the same set of familiar features.

*Memory and choice.* We now have an organism with some versatility, in that it can respond to a novel environment which is largely unfamiliar, provided there is some recognizable set of features for which a response has been learned. Obviously, it will sometimes happen that there are two sets of features present for which different responses have been found appropriate in the past. Faced with a choice of this sort it will generally be the product of the strength of each set of features and the strength of the associated memory trace which determines which response is most likely to be produced. What happens when there is no recognizable set of features for which a response has been learned? This is when the conditioned response comes in.

(b) *Memory, Intuition and Choice—The Conditioned Response*

In considering how the conditioned response could arise in a random mass of cells, the original analysis was handicapped by the adoption of a rather unsatisfactory concept for the stabilizing mechanism. Because of this, the possibility of the mass generating the conditioned response was illustrated in a very restricted form, that of two plane waves intersecting in a non-linear medium. If we have a direct inhibitory influence introducing intrinsic stability in the mass, then the interaction between two surges of activity and the medium takes a much less restricted form. Two components of a surge of activity, each arising as a result of some external stimulus, may now interact with each other and with the medium to produce a "memory trace" that records their simultaneous occurrence. For this to lead to meaningful behavior, as in the conditioned response, it is essential that the cell mass shall have had some prior "education" in the form of trial-and-error learning.

To visualize how the conditioned response can arise, we must bear in mind the response structure which has been imposed on the cell mass during trial-and-error learning. We shall already have a number of established paths along which surges of activity can travel. The surges of activity are essentially of a co-operative nature, and there are normally a number of parallel interrelated paths leading from stimulus to response, of which only a proportion will be active at any one time. Figure 4 is a simplified illustration of a few cells in which activity is the result of impulses of excitation from three stimuli:

- (i) A stimulus "A" to which there is an established response A'. The diagram illustrates two parallel paths leading from A to A'.
- (ii) Two stimuli B and C which, because they have not occurred frequently in the past, may not have a well-established response.
- (iii) A random excitatory stimulus S which sometimes does and sometimes does not excite a particular cell.

Let us now suppose, for simplicity, that the scattered excitation supplies a mean value of one impulse per three cells, and that the cells have a threshold level of three impulses. The probability of any particular cell receiving a scattered excitation of more than one impulse is then given by Poisson's series as

$$1 - e^{-1} = 0.28.$$

Similarly the probability of excitation by more than two pulses is

$$1 - e^{-t} - \frac{1}{3}e^{\frac{t}{3}} = 0.045$$

and by more than three pulses is

$$1 - e^{\frac{t}{3}} - \frac{1}{3}e^{\frac{t}{9}} - \frac{1}{18}e^{\frac{t}{18}} = 0.005.$$

To see how this small group of cells will react in relation to the conditioned response, let us suppose that for a while  $A$ ,  $B$ , and  $C$

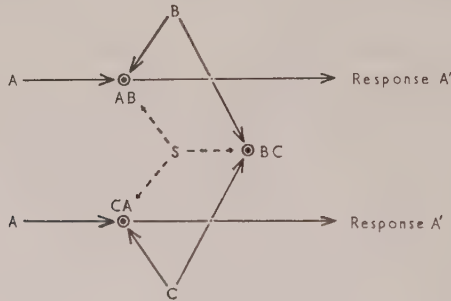


FIG. 4

all occur with the same frequency, but  $B$  always occurs simultaneously with  $A$  while, for comparison,  $C$  always fires independently. When the combination  $A$  and  $B$  occurs the cell  $AB$  requires one scattered impulse to fire, so it will fire with 0.28 times the frequency of occurrence of  $A$  and  $B$ . Cells  $BC$  and  $CA$  on the other hand require two scattered impulses and will only fire with a frequency of 0.045 times  $A$  and  $B$ . On the occasions when  $C$  fires alone, the cell  $AB$  will require three scattered impulses, and so will fire with 0.005 times the frequency of  $C$ . Cells  $BC$  and  $CA$  requiring one impulse, will fire with a frequency of 0.045 times that of  $C$ . Since  $A$  and  $B$  together, and  $C$  separately, all fire with the same frequency, the overall probability of cell  $AB$  firing is

$$P_{AB} = \frac{0.28 + 0.005}{2} = 0.14.$$

Similarly, the probability of  $BC$  and  $CA$  firing is

$$P_{BC} = P_{CA} = \frac{0.045 + 0.045}{2} = 0.045.$$

Thus the cell  $AB$ , which obviously forms an association element between  $B$  and  $A$ , fires three times as often as the cells  $BC$  and  $CA$  which do not. This difference is entirely due to the fact that  $A$  and  $B$  fire simultaneously. If  $B$  had fired independently the cell  $AB$  would only fire with the same probability as the others.

Now if, as we have assumed, the cell becomes more sensitive each time it fires (e.g. by lowering of the threshold) the cell  $AB$  will, after a number of simultaneous stimuli  $A$  and  $B$ , have a threshold appreciably lower than the others. We must also remember that we have guarded against the whole network becoming too sensitive, by assuming the existence of a stabilizing mechanism which will control the overall level of activity. This will maintain the effective average threshold at something like its original level of three, but the cell  $AB$  will still be lower than the rest.

Suppose the threshold of  $AB$  has reached a value of, say, one impulse unit lower than the rest, and stimulus  $B$  is applied alone, as in the traditional conditional response experiment. Then the cell  $AB$  will require one impulse of scattered excitation as against two for cell  $BC$ . Thus, when stimulus  $B$  is applied,  $AB$  will fire with a frequency  $0.28/0.045$  times that of  $BC$ , that is, approximately six times as frequently. Stimulus  $C$ , on the other hand, will show no tendency to fire  $AC$  more often than  $BC$ .

This example has deliberately been made very simple, and it should be noted that there will be a slight tendency for the association cell  $AB$  to fire preferentially even if the threshold has only altered a little. Even a slight tendency is still important, if it occurs in a sufficiently large number of " $AB$  cells", so that the integrated effect is sufficient to provide a statistically reliable association path from stimulus  $B$  to surge  $A$ . Not only will all these association cells act in parallel to launch surge  $A$  under stimulus  $B$  but, once a weak surge  $A$  has been launched, the combined effect of stimulus  $B$  and all the " $AB$  cells" with reduced threshold further on will reduce the attenuation of the medium for surge  $A$  and help to build it up. Thus, again, a minute change is sufficient to provide a memory trace, as long as the change occurs in a large number of cells scattered throughout the medium.

We can now see the relationship between trial-and-error learning and the conditioned response. As trial-and-error learning progresses, the irrelevant background detail becomes less and less



important, and it is just the familiar features that trigger the response. The conditioned response takes this a stage further. We now find that if features enter the background which are found to be relevant in the sense of being correlated over a significant period of time with the original (conditioning) stimulus, they may effectively become part of that stimulus. So much so, that not only does the response continue in the presence of these new features but, if on some occasion the original stimulus disappears, the original tentative response may still be produced. The nature of the world being what it is, this is often a satisfactory response and, if this is the case, it will be subject to reinforcement rather than suppression. On such occasions the ability to transfer a response in this way will obviously make for great economies in time and effort, while maintaining a high degree of plasticity. It will frequently happen that a novel situation can be correlated in this way with more than one response learned in the past. When such a conflict arises the first tentative response will in general correspond to the strongest correlation. If this fails to produce a satisfactory response then the others may be tried in order.

Why is it that the tentative responses chosen by the conditioned response show an economy? Followed to its conclusion, this question might lead to deep philosophical argument but it is worthy of note that there is a similarity between the criteria which determine the conditioned response and those which are indicated by Bayes' theorem. Following this theorem we may write an expression of the form:

Probability of action  $A$ , being appropriate in response to stimulus

$$S = P(A_i/S) = \frac{P(S/A_i) \cdot P(A_i)}{\text{Sum of all products } P(S/A_i) \cdot P(A_i)}$$

In other words, the probability that a particular response will be satisfactory, given certain features of the environment, is proportional to the frequency of simultaneous occurrence, as measured by past experience, of the stimulus simultaneously with the response. (It being assumed that the response is satisfactory or it would not have continued.) This must be divided by a normalizing factor which reduces the sum of probabilities to unity.



Now this is very similar to what happens in one mass of cells where the strength of the bond correlating the conditioned stimulus with the response may be expected to increase roughly in proportion to the frequency of simultaneous occurrence. Then again, the very fact that we have a stabilizing mechanism controlling the overall level of activity introduces an inhibitory effect, the magnitude of which must be dependent on the overall excitation provided by all the conflicting correlations. This inhibitory effect is somewhat analogous to the effect of the denominator in Bayes' theorem. Now, it is true that it is often difficult to apply Bayes' theorem rigidly to a practical problem, nevertheless it is interesting to see the evident relation between the conditioned response as presented here and the structure of the theorem.

Following Bayes' theorem, one might have argued, in justifying the introduction of a stabilizing influence, that the need for both excitatory and inhibitory influences is rather fundamental. The fact that in the equation quoted above both the numerator and the denominator contain quantities dependent on the stimulus means that in deciding the most appropriate course of action the stimulus must exert both an excitatory and an inhibitory influence.

### (c) *Thought Processes—Retrospection and Speculation*

So far we have only discussed elementary learning processes which give a trial response which follows the stimulus with little delay. On the other hand, it is common experience that prolonged thinking processes play a large part in the determination of human reactions and, to a lesser extent, those of the other higher primates. This is where the ability to regenerate the internal representation of a sequence of external events becomes important. The mechanism by which such sequences could be regenerated was pointed out in the paper already referred to. Any means, which allows activity to circulate more than once through the same region of a mass of cells, enables events which are separated in time to record their sequential relationship in the form of a memory trace analogous to that responsible for the conditioned response. Thus, if  $B$  has frequently followed  $A$  we shall have a memory trace  $A \rightarrow B$ . Moreover, again by analogy with the conditioned response, a surge of activity corresponding to, say, event  $A$  in the outside world may react with the memory trace  $A \rightarrow B$  to

regenerate activity which is the internal representation of the event *B* which has frequently followed *A*. Then, in turn, the internal representation of *C* may follow that of *B*, etc., etc.

The ability to regenerate a sequence of memories in the order in which they have happened on past occasions, corresponds to retrospection. Of more significance, is the fact that the same basic process should make it possible to regenerate a sequence of most probable succeeding events, which may not necessarily be an actual sequence that has occurred as a whole on any past occasion. In a most probable sequence of events, each event is followed by the event which has most frequently followed it in past experience. Such a sequence, in which the order is determined by the probability of succession, as measured in past experience, is by its very nature a train of thought about the future.

It is not difficult to see that this process could be elaborated, taking into account not only the probability of pairs of events, but also more widely separated transition probabilities. Just as the conditioned response combined with trial-and-error learning to give economy of time and effort, so the ability to follow a train of thought about the future, prior to taking action, can obviously effect further vast economies in effort. With the higher primates, we can visualize these three processes going hand in hand, the importance of the latter increasing as information is accumulated with experience.

## 5. SHORT TERM MEMORY

What evidence have we for the existence of a short term memory as distinct from, and in addition to, the long term memory which has been discussed so far? Some of the evidence comes from experiments on learning, and this is supported by the fact that concussion or shock can sometimes eliminate recent memories without apparently affecting memory of incidents further back. This would argue for the existence of some form of memory which can be lost in deep sleep or coma.

Now it has frequently been suggested that short term memory might be explained on the basis of circulating activity in chains of neuron cells. Moreover, it has been shown that such reverberatory chains of neurons may arise in randomly connected networks.

Figure 5 gives a simple illustration of this. It is not necessary to have specific predetermined chains built in for each perceptual element. It is true that local anisotropy or variations of connectivity, or other parameters within a random network, may assist the formation of such chains, but this is not the same thing as having regular sets of reverberatory circuits.

To explain short term memory in terms of random networks would be attractive, but at first sight this explanation appears to have one insuperable difficulty. If the network in which these reverberatory chains arise is randomly connected, how can one relate activity in any particular chain with the occurrence in the outside world which was responsible for its initiation? It is useless to record information if we have no means of playing it back in a meaningful form. Each recent occurrence in the outside world will have started activity in one particular reverberatory circuit, and it is not difficult to see that outgoing axons from neurons in each reverberating circuit will carry impulses that show that it is in an active state. If we monitor these axons we know when their circuits are active, but how do we know which circuit is connected to which axon, and which sensory input is connected to which reverberatory circuit if everything is randomly connected?

Fortunately there is a very simple answer to all these questions. It is that the long term memory mechanism discussed in the previous section, being itself based on a randomly connected mass of cells, is inherently capable of accepting information coming along randomly connected fibres. To begin with, this information cannot make sense but, in the course of time, the long term memory is capable of making sense of it by adaptation during the process of learning. Thus, if we think of the short term memory network acting as an intermediary between the sensory receptors and the long term memory, the latter solves the problem of random connectivity for the former. This relationship is illustrated diagrammatically in Fig. 6.

The concept of a short term memory acting as a temporary store of information between the sensory receptors and the long term memory also neatly solves a difficulty which has not yet been dealt with in relation to the establishment of long term memory. The difficulty is, that the whole exposition has been based on the fact that the attenuation of the mass for a particular

surge of activity is reduced significantly if that surge travels through the mass on many occasions. But how, one may ask, does this explain the fact that an important incident can occur once only, and yet it may be remembered for many years and perhaps for the rest of one's life? The interposition of a temporary storage of information explains this very simply, because information about an important incident may obviously be retained in the temporary store for some time, and during the whole of this time it may be

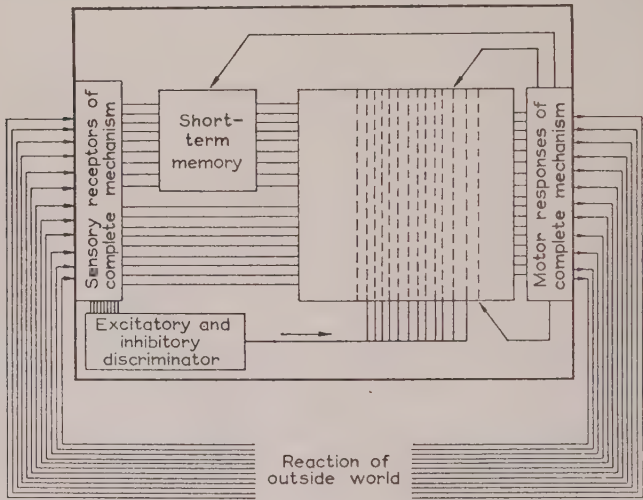


FIG. 6

fed continually into the long term memory. This also fits in very well with facts that have already been quoted about loss of memory of recent events. These would be the events which have not yet been imprinted firmly in the long term memory, and which are therefore lost if something destroys the information in the short term memory.

Another phenomenon which fits in very well with this is the "telescoping" of memories. When one recalls a sequence of past events, one does so in a space of time very short compared with the time taken for the events to occur. One picks out all the important features of the main incidents, and the mere passage of

time, and the occurrence of trivial incidents in between, are just omitted. This could readily be explained if only the "highlights" and important incidents were retained in the short term memory and passed on to the long term memory.

#### 6. DISCUSSION—THE RELATIONSHIP BETWEEN DIFFERENT FORMS OF MEMORY

One of the main objects of this paper has been to present various forms of behavior as different aspects of the behavior of one basic element—a mass of cells. As it has been the object to point out basic principles and relationships, the diagram of Fig. 3 has deliberately been kept as simple as possible. If the behavior which has been discussed here can be obtained in principle from such a simple basic relationship of component parts, then it will cause no surprise that a structure as complex as the human brain can show the remarkable corporate behavior that it does.

We have seen that the various functional properties which have been discussed, trial-and-error learning, association by the conditioned response, thought processes and short term memory are closely bound up with each other. In discussing the first two, trial-and-error learning was taken before the conditioned response because it facilitated demonstration of the dependence of the latter on the former. Chronologically this is the relevant order too. We start with a characterless mass of cells with no knowledge, ability, aptitudes or intuition, but only aspirations, represented by the associated discriminators. Initially, in these circumstances, the behavior of the organism towards an unfamiliar environment can only be the result of pure trial and error. Without the benefit of experience as a guide, the responses will be picked by chance from a wide range of possibilities. Later, as soon as even a small amount of experience has been assimilated, the conditioned response becomes important.

The place of the conditioned response in the pattern of behavior is that it allows the organism, when faced with a partially unfamiliar environment, to choose the response which has been found satisfactory on similar occasions in the past. The nature of the world being what it is, this more often than not speeds up the process of trial-and-error learning by picking more likely trial responses first. The concept of the conditioned response implicit in this description



differs somewhat from the traditional form, but it is felt that this conveys more realistic and, at the same time, more useful pictures.

The conditioned response can thus operate immediately the organism has begun to learn by trial and error. At first it may seem of little importance but as more and more basic responses are learned the conditioned response can become more and more effective in restricting the choice of alternatives which have to be explored during further trial-and-error learning. There is thus a very close relationship between the two. Trial-and-error learning makes the conditioned response possible and, at the same time, the conditioned response makes trial-and-error learning economically feasible by eliminating the wastage of time and effort which would be involved in trial and error operating with unrestricted choice of alternatives.

This close relationship between trial-and-error learning and the conditioned response is emphasized if we look at the sequence of events in traditional experiments with the conditioned response. This may be described as follows:

- (i) When the experiment starts it is found that there is already a satisfactory response for a particular environment containing features which we may label *A*.
- (ii) A new background feature *B* is introduced. This means in fact that we are dealing with a new environment which is similar to the first in that *A* is still present. A satisfactory response must now be learned by trial and error to this new environment and this process will be assisted by the continued presence of *A* if the response that was learned for *A* is also appropriate for *A* and *B* together. This will often, but not always, be so.
- (iii) The original feature *A* has disappeared but *B* remains. We are now dealing with a new environment to which a response must be learned by trial and error. Again the process will be speeded up if the response that was learned in stage (ii) for *A* with *B* is still suitable for *B* without *A*. This again will often, but not always, be so.

Now the second and third stages may be regarded as an example of trial-and-error learning of a response in an environment that is new but similar to the environment of the preceding stage. But the



first two stages may also be regarded as the initial stages in the development of the conditioned response which finally culminates in stage three.

One might say, are not trial-and-error learning and the conditioned response merely two aspects of one form of learning? This is indeed so. In the intelligent animals the two go hand in hand, with the conditioned response playing an increasingly important part until we reach what might be called the adult stage, when a large proportion of the tentative responses are satisfactory. One then tends to lose sight of the two learning components in what appears to be fully established "intelligent" behavior and it is only when we are surprised at the successful choice of a tentative response that we speak of "intuitive inspiration". But the behavior is only fully established as long as it is reasonably compatible with the environment and, the world being the changeable place that it is, minor revision and modification of behavior patterns is frequently called for. The reason that learning at this later stage appears so different, is that new behavior is built up on existing well-developed behavior patterns by adding to these and combining them, and learning is therefore a much more refined process than in the early stages. We now come to the part played by thought processes.

Thought processes are an even more powerful method of eliminating the waste of time and effort involved in trial-and-error learning. We have already seen that to work effectively, these depend both on the ability to regenerate most probable sequences, and on the intervention of a short term memory acting as an intermediary between the sensory receptors and the long term memory. The short term memory is essential because it makes possible the recording of isolated important events with a telescoping of the time scale which eliminates the mere unimportant passage of time between. The long term memory is essential because it provides the very basis for the relatively permanent recording of events in a form which allows the all important "most probable sequence" to be pieced together at a later date. The two forms of memory have been spoken of as though they were stored quite separately. This has been done deliberately for the sake of clarity. One cannot, of course, exclude the possibility that the two effects might in fact occur in close spatial relationship

within one mass of cells. The interconnecting bundle of fibers shown in Fig. 5 would then be superfluous.

The generation of the internal representation of a most probable sequence of events is related to the conditioned response, in that the same basic mechanism is involved. In the same sense in which Bayes' theorem was related to the conditioned response, it will also be related to the choice which is exercised subconsciously in compiling each step in a most probable sequence. The regeneration of a most probable sequence is also very closely related to trial-and-error learning, because the latter will have been the major factor determining the nature of, and also the structure of, the internal representation of events. That this is important, will be very obvious if we consider the question of tuition.

It is easy, in considering simple learning processes in man, to lose sight of the importance of tuition, yet it is a fact that most of the accumulated knowledge of civilization is handed down either by the instinctive learning of child from adult, or by some form of deliberate tuition. Tuition is a learning process superimposed from outside but working hand in hand with the intrinsic learning processes discussed above. In tuition it is the tutor who supplies the reward or punishment and not the world itself. He is assisted by the fact that communal mental activity appears to be something which is itself pleasurable (i.e. defined as satisfactory by the discriminator in Fig. 3). The advantage of tuition is that learning can be incomparably faster than learning by direct experience of the world, because the accumulated experience of generations is available once the pupil has learned (or been taught!) to learn. Man would be helpless in the present day world without this tuition.

In tuition much use is made, sometimes unconsciously, of the conditional response principle and of the ability to memorize and compose sequences. A very powerful structure is superimposed on the whole by the deliberate teaching of accepted mathematical and logical formulae to guide sequential thinking or check intuitive inspiration, and by the often less deliberate teaching of ethical codes. Moreover, the way we learn may have an overwhelming influence on our approach to further study.

It is a problem in teaching science to provide the pupil with an adequate foundation of existing knowledge without so constraining

his thinking processes that his ability for independent thought is impaired. It is difficult to estimate how forcefully the superimposed structure constrains our thinking along accepted lines which, while they have the advantage of having been tried and tested, may at the same time destroy our breadth of vision. Similar logical structures are superimposed by man on a computing machine during programming. To some extent this answers the much debated question of the similarity between a digital computer and a brain. The answer is that the same type of logical thinking structure is superimposed more or less successfully on both—by man.

Many of the participants in this Symposium are, I know, more interested in making self-organizing systems than analysing those that occur naturally. I do not pretend that the system I have discussed is by any means an ideal one. I have merely tried to suggest some principles that may be found relevant when we understand more about the spread of organization in the masses of neuron cells found in the brain. Perhaps by examining the natural system in operation we can derive some hints as to how an artificial one might work. It would not be the first time this has happened.

#### REFERENCES

1. R. L. BEURLE, Properties of a mass of cells capable of regenerating pulses, *Phil. Trans. Roy. Soc. B* **240**, p. 55 (1956).
2. D. A. SHOLL, *The Organization of the Cerebral Cortex*, Methuen (1956).
3. J. C. ECCLES, *The Neurophysiological Basis of Mind*, Oxford University Press (1956).  
J. C. ECCLES, *The Physiology of Nerve Cells*, Oxford University Press (1957).
4. G. A. HORRIDGE, The co-ordination of the protective retraction of coral polyps, *Phil. Trans. Roy. Soc. B* **240**, 495 (1956).

#### DISCUSSION

PLATT: Dr. Beurle, could you tell us if you have thought about how you could make an analog of this behavior or represent its input?

BEURLE: I have used a digital computer to simulate a relatively small array of interacting cells. Beyond this, I have no practical suggestions about making an analog. I think we should find out more about the system we are studying, and at the same time study the logic of the practical problems we wish to solve. These two approaches might ultimately converge to suggest a practical "thinking machine".

I imagine that one reason for wishing to make a thinking machine is that our own individual mental capacity is severely limited. I sometimes wonder whether someday we shall discover a really efficient method of communication which would allow us to pool the mental resources of several human brains much better than we seem to be able to do at present.

BOWMAN: It has been tried in a good many crude forms; they call them governments.

SHERWOOD: There are several points I want to ask about. First, how do you localize the surge in actual time and space? Second, do you identify these waves with EEG activity? This you must not do. We have not enough evidence. Alpha activity does not correlate well with input activity. It is not even stimulus locked.

Next, what is the nature of your stimulation? Is it the local electrical one of the brain or nervous system, or is it a so-called physiological one, an on-and-off light or sound, or something of that sort? This is a very serious difficulty, the identification of brain waves with actual signal processing.

BEURLE: As far as EEG waves are concerned, I don't correlate them with any certainty with what I have been talking about. In fact, I have three alternative explanations of the alpha rhythm.

ROSENBLATT: The models which Dr. Beurle has just described come very close to some which we have studied in our program on the Perceptron. I think I might just introduce a few items of correlated evidence here which might amplify what he has just described. We have studied several classes of systems which are extended in depth beyond the Perceptrons which we usually consider, which are generally just three-layer systems consisting of a mosaic of sensory elements, a layer of association cells, and a single layer of output transducers. If we introduce additional association layers, so that we extend this section in the form of a long column of cells, and if we introduce inhibition, which Dr. Beurle has found is necessary to control level of activity, in this column and keep it from exploding into total activity as you go through, then we do indeed get stable performance; that is, we find some level of activity at which this settles down. The most essential property that seems to emerge here is that the system improves in its discriminative activity. Patterns introduced at the input end which are ordinarily very difficult to distinguish become rather easy to distinguish.

On the other hand, patterns which were initially very easy to distinguish tend to become somewhat more difficult. We tested this, for example, with an environment of horizontal and vertical bars, and if we ask our system to put all the horizontal bars into one class and all the vertical bars into the other, the simple three-layer Perceptron finds this to be a very simple task. The task becomes somewhat more difficult as we start multiplying association layers of the system.

On the other hand, we can give it an alternative test in the same environment, and require that it place every odd-numbered bar, numbered according to its location, into one class, and every even-numbered bar into the other class. This ordinarily is a much more difficult task for the single association layer system, but it becomes just as easy as the first task as we increase the number of association layers. The tasks apparently become identical in difficulty with the aid of discrimination. Moreover, we are developing an invariant here, in terms of the level of performance available. We are developing an invariant with respect to the size of stimuli, the type of intersections which can occur among them on the retina. If we increase the number of



stimuli present in the environment, the task becomes more difficult, but the performance becomes independent of dichotomy and independent of the separation or relative discriminability of classes as described at the retinal end of the system, and we are paying for this. We are gaining discriminability at the cost of generalizability.

**RAPOPORT:** I would like to clear up a point that has bothered me for a number of years, and that is in connection with other very similar models of trial-and-error learning proposed by Shimbel in 1950. If you remember, his stimulus went by random connections to a series of responses, and, in principle, his response is random but actually becomes dependent on the fluctuation of threshold. Whatever threshold happened to be low, that is the response that occurred. Then there is a success center which sent out a message, which he called a "to whom it may concern" message, a message of "lower threshold" sent to everybody. But those connections where activity had just been present received the message most effectively. In other words, these were where the threshold was lowered the most, and that led, as can be shown very well, to the narrowing down of the response on to the most successful response.

The thing that bothered me about this model was the ad hoc assumption that the threshold was lowered essentially everywhere but it was lowered most at the place where activity had just taken place. What I want to ask you is this: does your model still make that explicit assumption, or is that assumption obviated somehow by the surge of activity attenuating itself?

**BEURLE:** No, I didn't make that assumption. I deliberately avoided it because I didn't think it was necessary. It is obviated. There is no assumption of a general message such as you have mentioned. Each neuron merely has the property of becoming slightly more sensitive each time it becomes active. The problem is taken care of by virtue of the fact that you can only store memory by a very large number of repetitions.

Shimbel, on his assumption, could effectively record a memory with a small number of repetitions. With this theory, it requires a large number. Each repetition reduces the attenuation of the material only very slightly, for just that particular surge of activity. In other words, it becomes a more and more selective filter for just that particular surge of activity.

**RAPOPORT:** I am not quite sure that Shimbel's model makes possible a single, one-shot fixation, because, you see, it is the average threshold that is lowered by a little bit. Then a low threshold becomes a more formal problem.

**SPERRY:** You mentioned that the more often A and B are paired, the more firmly the new combination becomes ingrained. There is a thing called "aperiodic reinforcement" where the food or the shock is not given every time but only in a fraction of the trials, randomly scheduled. This procedure gives stronger conditioning than where the two are repeated together every time. Does your model take this into account?

**BEURLE:** Not in its very simple form. As I said, I made this model extremely simple because what I really wanted to find out were the basic principles. I am sure the brain is vastly more complicated. Having arrived at these basic principles, you have effectively got bricks out of which you could build a very much more complicated network, if you wish.

**SPERRY:** Just one word in favor of non-random networks. There are lots of places in the nervous system that look like randomized interconnection of similar neural elements under the microscope, but which can be shown to consist of elements that are qualitatively specific and interconnected in a highly selective pattern—though you may be on relatively safe ground in some

local areas of the mammalian cortex. I often think there might be advantages for such model building if one were to use instead the midbrain tectum of the fish. The midbrain optic lobe is only about the size of a large pin head, yet it is adequate for the perception, learning, and memory of color, pattern, and many other discrimination habits. Its histology is pretty well worked out and there is considerably less in the way of apparently randomized networks.



**JOHN R. PLATT**

*Physics Department, University of Chicago*

## HOW A RANDOM ARRAY OF CELLS CAN LEARN TO TELL WHETHER A STRAIGHT LINE IS STRAIGHT\*

The retina, when we look at it under a microscope, looks random. It would be surprising if it were not. Most of our tissues seem to be. Certainly one cell can be hit by a cosmic ray and die, or one may double more rapidly than its neighbors. A necessary result is that after many cell divisions have taken place during embryonic development, a certain randomness must appear in every tissue.

This randomness puzzled me greatly in considering pattern perception, because it raised the question, how can we ever distinguish regular patterns or discover whether a line is really straight, for example? One could believe that a line might appear straight to one man because it crossed a certain set of retinal elements which his brain "knew" were in a straight line from the beginning. But suppose the man has a twin brother whose genetics is as nearly identical as possible, but who has a slightly different developmental history or has been hit by cosmic rays in slightly different cells. Then one might expect that "corresponding cells" in the retinae of the two brothers, if one can make any such identification between the retinae, would lie in a slightly different pattern; and the line that appears straight to one of the brothers should appear full of little wiggles to the other, and vice versa.

In short, the question boils down to: How does the brain "know", and how accurately does it know, where particular retinal cells are? Or can it improve its "knowledge" in the course of visual operations?

---

\* This is a summary of the two more complete discussions that have already been published<sup>(1,2)</sup>.

Helmholtz said that the brain "knew" by "local signs". We have all seen the paper by Lettvin, McCulloch and co-workers last November, the dramatic and fascinating paper on "What the frog's eye tells the frog's brain",<sup>(3)</sup> which appears to show that particular points on the retina are connected to particular points on the cortex, at least within their experimental accuracy of about 1 degree of arc. Nevertheless, one must still suppose that there is some uncertainty in the location of a "particular" cell or in the "straightness" of a particular subset of cells which were intended genetically to be in a straight line, just because of the developmental irregularities. And the amplitude of these uncertainties would determine the amplitude of the wiggles each of us would see in looking at a straight line.

But the fact of the matter is, there are no such wiggles (except in rare instances of post-childhood retinal damage). The wiggles that ought to be there seem to have been ironed out by some more precise post-developmental mechanism. Our ability to perceive that a line is indeed straight is fantastic. We cannot tell a straight line from a gently curved one, but we can tell a straight line from one with breaks and wiggles in it, and from "S" curves and the like.

I got a clue to at least a possible mechanism that could do this job when I heard Ditchburn talk about the scanning motions that the eye is making all the time, which are too fine and fast for us to be aware of.<sup>(4,5)</sup> He has classified these motions into three types. First, a "tremor" motion, of about one-half minute of arc, with frequency from 50 to 200 cycles per second. Second, a "drift" motion, such that a point of light one is fixating on actually drifts across the fovea. Third, a motion he calls "flick", which is a sudden resetting of the fovea such that the fixation point moves back to the center.

The discussion of these motions and of how vision ceases whenever the motions are canceled out to give a "stabilized retinal image", as shown by Ditchburn and by Riggs and co-workers in this country, made me wonder whether such scanning motions would help a random mosaic of cells to tell when a line in the external field is indeed straight. It turns out that this is possible in principle, and that one can determine the straightness of lines and the regularities of several other kinds of geometrical patterns

by scanning motions, by moving the retina across the patterns, without knowing where any of the cells are located in advance. This is a dynamic method; and I have been unable to think of any static method that would make similar discriminations with similar accuracy and that would overcome the objections I have raised about the uncertainty of individual cell location. This dynamic method would make it necessary for the young child or for the visually naïve adult (such as one who has just had cataracts removed) to spend some time on scanning operations before the locations of the cells could be determined accurately or before the straightness of a line could be accurately judged; but I know of no data that would exclude the necessity for such a learning period. I should emphasize that I do not claim that the human eye actually learns to detect straightness of lines or any other pattern property by the scanning methods I am discussing, although these methods certainly lead to many suggestive and quasi-human consequences. But I do claim that a mechanical array or mosaic of receptor cells at the focus of an optical system could learn to discriminate straightness and other pattern properties by such scanning operations, even if the cells were arranged at random and wired at random and the location of individual cells were initially unknown.

If we have a random array of cells across which a straight line passes, we can displace the mosaic of cells along the line or the line along the cells, and the cells that the line crosses after displacement are the same ones as before displacement. However, an "S" curve or a general wiggly line will necessarily lie on a different set of cells after displacement from the set it lay on earlier. A straight line is therefore distinguished by the property that it is "self-congruent", that is, that it is congruent to itself after an arbitrary displacement along itself.

This property of "self-congruence under displacement" is obviously an "invariant" of a straight line pattern. The problem of detecting such an invariance is the problem of detecting the invariance of a set of sensory signals, which is obviously about the weakest requirement we could make of a mosaic receptor system, since it does not involve any examination of what the details of the signals are. In a straight line that has a break in it, the break can be detected because the signals it produces are not

invariant under displacement, as Weymouth and co-workers emphasized long ago.<sup>(6)</sup>

Besides straight lines, uniformly curved arcs are also self-congruent under displacement, provided the displacement contains a certain amount of rotation. But the eyeball is capable of rotations about its optic axis by large angles, so there is nothing inherently impossible in our use of a scanning method to determine whether an arc has uniform curvature or not.

Another self-congruent pattern is a set of parallel lines. If the lines are not parallel, they are not self-congruent. (I am speaking here exclusively of small patterns only two or three degrees in diameter, on a plane surface perpendicular to the optic axis, which is the simplest case to discuss first.) With a displacement-self-congruence method of determining pattern relationships, parallel lines will have a unique perceptive quality that non-parallel lines will not have. Other primitive perceptions of such a mosaic receptor obviously will include the perception of concentricity of circular arcs or circles; the perception of equidistance, using discrete displacements; and so on. In detecting equidistance, we would get one pattern of stimulation and then would move the eye by a discrete amount and see whether the original pattern of stimulated cells is repeated again, using that as a test of whether the second set of pattern elements has the same spacing as the first set. In short, we can use the retina to establish a metric.

It will be obvious to most physicists and mathematicians what the other possibilities are for patterns self-congruent under linear or angular displacement. We can therefore go on to list a few of the advantages of the displacement method\* in pattern perception, advantages that would deserve consideration in an artificial random mosaic and that certainly would seem to be helpful in a natural system, like the eye.

We have already discussed the primary advantage, that pattern-perception using displacement does not require knowledge of the position of individual cells. A second, and related, advantage is that it does not require knowledge of or regularity in the

---

\* The "displacement-self-congruence" method is better called the method of "functional geometry" for the reasons given in Reference 1, but I have chosen not to emphasize this relatively unfamiliar, name with its larger implications, in the present short talk.



sensitivities of the individual cells. If a cell has been knocked out by a cosmic ray, a pattern that crosses it can still be self-congruent after displacement. If cells vary in sensitivity, as commercial photocells certainly do, you could still use them perfectly well for this method of detecting pattern. And one would certainly expect biological cells to vary in sensitivity, as they do in size and shape.

A third property which is important is that the output signal that tells the brain that a pattern is present, can be independent of the mosaic structure. All the output signal needs to say is "Same array!", and even though there may be errors associated with the grain size, possibly errors comparable to the grain dimensions, these errors or this mosaic structure never appears as an element of perception. The mosaic structure goes out of the picture, because the only signal you finally get through is a signal of congruence or non-congruence. One can imagine methods of tremoring back and forth across a line to get a time-dependent array of primary signals making a very much finer discrimination of boundary positions than would be given by the size of the mosaic structure itself, as Marshall and Talbot suggested.<sup>(7)</sup> In this case, the time-dependent array at one time can be compared with the time-dependent array after a gross displacement, for a super-fine discrimination. The accuracy of discrimination may be limited rather by signal-to-noise considerations for the time involved in studying the pattern or making the comparison, rather than on the graininess of the mosaic structure. But in any case, the mosaic structure need not appear explicitly in the perception.

A fourth, and most suggestive property, is that the perception of straightness or any other pattern relationship by scanning methods is independent of the actual shape of the image on the retina. It makes no difference whether the image of a straight line in the external field lies on a straight line on the retina or on a curve or in a distorted wiggle; and it does not even have to lie on a connected section of the retina (for example, when it crosses the blind spot). Because, as the eye scans from one part of the straight line to another, all this method asks is whether the signals afterwards are the same as they were before. If the straight line was imaged onto an "S" curve on the retina originally, then as we scan along it, it continues to be imaged on the same "S" curve. It



is not the actual straightness of the image, but its "operational straightness", so to speak, that counts.

The straightness (or other pattern relationship) that we detect is therefore a straightness in the external field, and not in any internal field; and straightness in the internal field is unnecessary. I think this can blow away one of the old biological riddles, which was: How is it that the mapping of a straight line on the cortex is far from a straight line, and the mapping of a set of equidistant points on the cortex is far from equidistant? Such questions become irrelevant with this method of pattern perception, and it is obviously important that they be irrelevant in a plastic biological system.

Moreover, the fact that the perception of a straight line means a straight line in the *external* field has the consequence that we can discuss such patterns publicly without the intrusion of personalism. The straight line you see is the straight line I see. Both are in our external fields. There is nothing private about straightness or parallelism or concentricity, or equidistance. There may be a good deal that is private and unique about our particular network connections which signal straight lines, but the straightness itself is an invariant property in the external field. This makes it a subject for public discussion and education and also makes it a subject for public language. Such a method of perceiving pattern therefore has a number of implications which fit the philosophical and linguistic aspects of the human situation in a very satisfactory way. If alternative ways can be devised by which a random mosaic receptor could discriminate patterns, it will be important to compare them with this method of displacement-self-congruence in respect to these larger implications.

Another larger implication is that such a method of pattern perception requires learning. The fact that a certain set of cells is stimulated in a self-congruent way means that (given certain other obviously necessary limitations on the character of the pattern) they lie on a straight line. Their "addresses" in the network and their positions in space relative to each other have been learned by means of their association in a straight line perception; and the brain has determined their relative addresses in the course of using them. It is a considerable advantage in a million-element array of cells to determine addresses in this way rather than

having to put them each initially in exactly the right spot and wire them up individually in exactly the right way. This is what makes our present artificial computing and perception systems so expensive at the present time. All this pre-assembly information has to be put in by an army of high-school girls working at long assembly lines for months or years. By comparison, a baby is very cheap and simple to produce. The only price we pay is that it comes out with all these random arrays of cells, and then takes twenty years before it locates them all and learns to make all the discriminated perceptions we need as adults. But would it not be nice if we could make also a quick-assembly mechanical system that would do some of the same things?

#### REFERENCES

1. J. R. PLATT, Functional geometry and the determination of pattern in mosaic receptors, in *Information Theory in Biology*, pp. 371-98, Eds., YOCKEY, QUASTLER and PLATZMAN, Pergamon Press, New York (1958).
2. J. R. PLATT, How we see straight lines, *Scientific American* **202**, No. 6, pp. 121-9 (June 1960).
3. J. Y. LETTVIN, H. R. MATURANA, W. S. MCCULLOCH and W. H. PITTS, What the frog's eye tells the frog's brain, *Proc. I.R.E.* **47**, pp. 1940-51 (1959).
4. R. W. DITCHBURN, Eye-movement in relation to retinal action, *Optica Acta* **1**, pp. 171-6 (1955).  
R. W. DITCHBURN and D. H. FENDER, The stabilized retinal image, *Optica Acta* **2**, pp. 128-33 (1955).
5. L. A. RIGGS and co-workers, The disappearance of steadily fixated visual test object, *J. Opt. Soc. Am.* **43**, pp. 495-501 (1953).  
L. A. RIGGS and co-workers, Motions of the retinal image during fixation, *J. Opt. Soc. Am.* **44**, pp. 315-21 (1954).
6. E. E. ANDERSEN and F. W. WEYMOUTH, *Amer. J. Physiol.* **64**, pp. 561-94 (1923).  
H. L. AVERILL and F. W. WEYMOUTH, *J. Comp. Psychol.* **5**, pp. 147-76 (1925).
7. W. H. MARSHALL and S. A. TALBOT, Visual mechanisms, *Biological Symposia*, Vol. 7, pp. 117-64, Ed. H. KLÜVER, Jacques Cattell, Lancaster, Pa. (1942).

#### DISCUSSION

ROSENBLATT: It seems to me when a theory reaches this degree of specificity it becomes both possible and imperative to undertake a quantitative analysis in order to find out whether it actually meets the known quantitative facts of the situation. It seems to me that it is plausible that optical tremor might account for vernier acuity. On the other hand, it is not immediately clear that this is the case. Vernier acuity permits us to detect slight relative displacements

of two segments of a line which is considerably less than the diameter of a retinal element, and on first consideration it would appear that if you now translate such a line in such a manner as to try to obtain self-congruence, that both sections, the slightly out-of-line section and the prior section, would both be translated across the same array of cells.

Now clearly there will be a slight difference in the sets of cells picked up. As one is translated parallel to its length, the slightly displaced section is going to strike the edges of a few cells that have been displaced, which are slightly out of line with the others, so that the theory which is proposed does have a certain plausibility. On the other hand, the question is, do we actually have a high enough probability of picking up additional cells or a sufficient number of additional cells with a 30-second tremor, or whatever the amplitude is, to account for the degree of fineness of discrimination that is possible here? This is one consideration.

The second is that there are a number of demonstrations, now, of pattern perception of a very sophisticated sort which is possible in four microseconds or less, that is, with three or four microseconds exposure, and then it is possible to tell that this is a picture of a man or to describe a scene which is being exposed. Now clearly you do not have optical tremor approaching a megacycle, which would be necessary if this is to be the actual mechanism of pattern recognition under such circumstances. Now this criticism is not applicable to vernier acuity, which clearly does not occur under these conditions. You cannot expect to get the sort of vernier acuity in four microseconds that you can if you are permitted to examine a card with a straight line on it for several seconds.

On the other hand, if I understand the position taken in the *Scientific American* article in which this has appeared, I think there is a suggestion that an extension of the same techniques of figural analysis might account for more complicated pattern recognition, the perception of angles and this sort of thing, which is clearly coming into play in recognition of scenes, figures, and so forth. Do you have any comments on quantitative studies of this sort, and also what happens in the case of very brief exposures where pattern vision is still possible?

PLATT: Yes. While it is true that an experienced eye, an adult eye, can detect a man or can read print over about one degree of angle on a tachistoscopic exposure of a few microseconds, nevertheless the naïve eye of an adult—say one who has been operated on for cataracts and has been fitted with a lens which works as far as the ophthalmologist can tell us—this adult eye takes months before it can tell a triangle from a square or a tree from a book on the table or even a red triangle from a blue triangle. I think therefore that the eye needs to *learn* its organization; and that when we have this tachistoscopic recognition without scanning we are perhaps comparing the pattern with a pattern learned years earlier. We are perhaps comparing the image of the man with another set of images of a man seen long ago. This is one possible answer.

It is true that this general theory of perception is fairly specific and suggests a host of experiments, none of which have been done. For example, it is possible to measure the motions of the eye in scanning over particular patterns or in scanning over visual illusions such as Zollner's illusion. The scanning theory suggests that certain motions of the eye should then recur with very high probability. The people who measure motions of eyes have never looked at tremor and drift motions on line patterns. I think the measurement of

motion during scanning of patterns, particularly the measurement of rotation around the optic axis, which has been quite neglected, would be most important. I do not want to claim that my functional geometry method *is* the way the eye sees. But this *might* be the way the eye sees; and it would be a way to get an artificial mosaic system to perceive pattern. It certainly raises questions about whether the eye does, in fact, see this way.

SPERRY: Another comment with respect to "learning the addresses" between eye and brain. I think it is clear enough that the whole topographic, map-like projection of the retina, first on the three alternate layers of the geniculate nucleus and then on to the visual cortex, as well as that on the midbrain optic lobe and superior colliculus, is built in by orderly growth forces with learning unnecessary. Also, finer aspects of visual organization that are not anatomically demonstrable and are much more complicated than the discrimination of simple straightness of lines can be shown to be built-in in various species. Recall, for example, the "visual cliff" experiments of Dr. Gibson and others, the starlight navigation of migratory birds, and so on.

I am not sure that evidence drawn from cataract cases, or the use of Lucite cups and other diffuse light conditions, is something we can count on, because there are elements in the visual system that seem to depend for their firing upon edges, contrast effects, and rapid on-off light changes for their stimulation and possibly for their normal development and maturation. These, and the related units involved in perception may well therefore undergo something like disuse atrophy under prolonged exposure to diffuse light, just as does most of the retina in prolonged darkness.

PLATT: I would certainly agree that much of the visual system, a tremendous amount of perception, and possibly perceptions, like color perceptions, are genetic. Genetic determination of patterns exposed to would seem to be a desirable scheme throughout the lower organisms. Naturally we would have preserved much of any such desirable genetic scheme. But what I would suggest is that perceptions like straight lines which we make very precisely, which are not part of the normal environment of the fish and the monkey, may be *learned* elements, which can be superimposed on a pre-addressed perception system which is genetically determined.

Likewise, I think that any pattern-perceiving artificial mosaic system will also have a tremendous pre-addressed section which would have to be wired up properly in the beginning in order to work at all. Only, on top of that could it begin to have some sophisticated learning elements.

MCCULLOCH: About two weeks ago I went down to visit Riggs; he has given me a reference to it, but I have lost it. He has a girl working with him who has just done a very careful job on the immobilized image. When you immobilize an image, you have a short time before it fades. By working in that short time she was able to prove that our vernier vision was at least as good as, perhaps slightly better than, with the image moving with respect to the eye, so that in the case of the human eye we know for the first time that this is probably not necessary particularly. It is necessary for the preservation, but not for the resolution or the detection of shapes. I am sorry; I do not have her name.\*

\*Ed. Note: Dr. Platt has called our attention to the exchange of letters between himself and Dr. Riggs in the November 1960 issue of *Scientific American*, touching on the matter of Dr. McCulloch's comments here.





**GEORGE W. ZOPF, Jr.**

*Electrical Engineering Research Laboratory, University of Illinois*

## ATTITUDE AND CONTEXT

### HOPES AND PREJUDICES

Although I have been concerned with self-organization for ten years, and active in its study for five years, I must admit to considerable bewilderment. To be sure, most of my bewilderment comes from the fact that this is a bewildering subject; I suspect, however, that part of my bewilderment may be laid, not to the complexities and confusions of the new field of cybernetics itself, but to the activities of cyberneticists. In fact, when I told one of our symposiasts here that I intended to talk on constraint in cybernetics, he snorted, and suggested that, instead, I talk on the restraint of cyberneticists. I cannot so easily relinquish my original intent, but he most certainly has a point. Perhaps I can do his subject implicit justice.

B. Delisle Burns,<sup>(1)</sup> the neurophysiologist, once wrote, in a tone I thought more wistful than expectant: "It would be helpful . . . if the authors of papers on the upper end of the central nervous system were to state, before the section on methods, their prejudices and hopes. After all, it is distressingly easy to find what one is looking for and remarkably difficult to discern the unsuspected or the unwanted."

A request so plaintively put should be answered. I should like to state, so far as they are available to me, my hopes and prejudices, and thereby constrain the dimensions of the space within which I judge and am willing to be judged.

My prime hope is a simple one. I hope only that cybernetics, and that subsection concerned with self-organizing systems, can illuminate just one cybernetic problem: that of control and communication within and between men. You may thus expect

me to squirm in my seat during descriptions of the amusing, the tricky, or the technologically useful; I am willing to consider symbols and black boxes, slimes and sludges, factories, computers, and missile systems, only in the expectation that inferences from these may somehow apply where my concern really lies.

My prime prejudice is cognate to my hope. It is that there is no problem logically or practically prior to our understanding of ourselves. Under both of these terms, I can answer any question of the form: "How or what can we know about  $x$ ?" with, "First know the knower."

I believe it is logically prior, because many of our problems are not objective, are not to be solved by considering that everything is "out there". In fact, many of our problems *are* problems only because we demand wholly objective reference.

I believe it is practically prior, because I see embedded in every question of how to close missile gaps, correct space lags, encourage consumption to go from trot to gallop, that which begs the question. Why do we ask these questions? Why do we proceed as if they were both necessary and important? Just what is their connection to the main business, that of the extension of our knowledge of ourselves and our relation to the world?

This is indeed a violent prejudice, so much so that I am intolerant of those who regard the whole of biological data, of the phenomena of biological organization and intelligence, as not more than a grab bag from which to abstract technological goodies. My intolerance is tempered only by the belief that such casual abstraction cannot succeed. Do not mistake me. I have no doubt of our capabilities to build, someday, both automata and intelligent automata. I do doubt, however, that we can gain much from casual sampling of biological particulars, neither in explanation nor exploitation. Those who are interested in useful automata should be reminded that the biological systems can only be suggestive; they are rarely systems that actually perform the tasks we want machines to do, and, to date, they are not systems that perform those tasks *in vacuo*, as isolates. One may phrase this as an Irish bull: Machines now do many things better than man, but they are things man does not do.

I do hope that we shall not fall into the trap for a time occupied by a friend of mine who, speaking of a device he had built, said,

in all seriousness, "If the conditional reflex were only like this, then what I have built is a model of the conditional reflex".

### ON OBJECTIVITY

It is quite true that we have had a remarkable historical success in the objectification, the externalization of entities. One might even characterize science as the activity of doing just this. We have, to give a pertinent example, been able to construe order as belonging to, present in, the systems of the external world alone. Order, in its avatar of entropy, has been rendered a function of state. But notice that entropy is a defective concept. It maps into, not onto order. It becomes a measurable, objective property only on the assumption of an arbitrary reference, a base ordering. We cannot speak of order, but only relative order: we must choose some maximum or minimum ordering quite arbitrarily. Popper<sup>(2)</sup> has wisely commented that we cannot assess the order of anything; we can only state that this or that specified order is or is not present, or in some cases, state its degree.

So much for the idea of the adaptive or self-organizing system as something that, in a general way, seeks, finds, and devours the order of its environment. To counter the metaphorical appeal of this characterization, one need only note that we can detect only those orders that we already possess; at best, we can only detect the difference between our reference orders and those "in" the environment. We, and our fellow self-organizing systems, are propositions that such and such an order might exist in our environment. In this we are like a filter circuit; we differ in that we have access to several ways of detecting the same order, and to several orders which we can detect. But we are not free to find just any order, to order just any finding, we are constrained creatures. I shall touch on more of this when I discuss the game of  $2^{2^N}$ .

Similarly, we have attempted to place memory, learning, intelligence, as intrinsic properties of objects or systems, and have thereby set bars to the understanding of these phenomena. They are surely properties not of this or that system, but of our relation to those systems. In fact, the attribution of these properties to any system always amounts either to a confession of ignorance on

our part, to the statement that our knowledge of the present complexion of input and internal states of the system is deficient, or to the redefinition of the terms.

This latter is the case when we speak of the "memory" of the digital computer. However much we may regret that we do not employ the more non-committal term of "store", it remains that we can talk quite lucidly and accurately about computer memory. It is only when we fail to recognize that the operation of the computer depends only on the present states of its components, including those of the memory block, it is only when we attempt to compare machine memory with its human counterpart, that we have trouble. To say that a system "has" memory is to say only that we do not, or cannot (or choose not to) know its present complexion, but that we can render its behavior explicable by reference to previous input states. Ashby has made this point abundantly clear; I need not belabor the point.

To say that a system is intelligent is to say only that we cannot produce sufficient evidence for the determination of its behavior in certain "problem solving" situations; it is to say, by a deliciously circular but inevitable argument, that it is intelligent because we must assume its intelligence in order to make its behavior intelligible. Note how many computers drop in I.Q. as soon as their lid is lifted! Note how the provenance of intelligent machines often requires us not to let our right hand know what our left is doing. We must know in full deterministic detail what we are doing to build a complex machine; to call such a machine intelligent, requires that we forget or ignore our knowledge of just how it does what it does. Perhaps we can more comfortably describe devices as intelligent when we start providing the sorts recommended by Pask and Beer, where it will be impossible (or at least inconvenient) to specify just how the device solves a particular problem.

Note also the hidden arrogance. Save in jest among safe company, we rarely deny our own intelligence, and we tend to ascribe intelligence to those machines with which we can identify ourselves, in terms of the problem presented, the solution reached, and *the degree of ignorance of the intervening processes*.

We say that a system learns when, while indubitably remaining the "same" system, it shows now a functionality it did not show



then. This amounts to the assumption that the input *now* is identifiable with the input *then*, and that the system has preserved its formal identity. Again this is an assumption based either on ignorance of, or the impossibility of discovering, the internal states of the system in all completeness.

It is the fashion for some of us to claim a wider ambit for learning than the rest of us will generally accept. Such people will include the merely plastic or frangible under the learned, and claim that the breaking-in of an engine, or the apt conformability of an old glove to the hand are examples of learning. Not so. I can demonstrate, without reference to experience, that the engine and the glove *are* altered. If you will bring me two indistinguishable gloves, indistinguishable, that is, in every respect save that one snuggles aptly on my hand, and then tell me that this one has had a lot of experience with hands, then will I admit to learned gloves.

Or bring me two men, and tell me that one of them has learned Sanscrit. Be I the most gifted of neurophysiologists, and a scholar in Sanscrit to boot, there is presently no way that I can identify the linguist without reference to performance. We do not know the states of the machine which are indices of "knowing Sanscrit". I dislike such extrapolations, but I doubt if we shall ever have such state knowledge. I doubt that we shall ever have need to, or want to. To believe that we must, is the objective, the analytical, heresy.

I think this is what Gordon Pask intends when he gives us the option of regarding a system as composed of elements that behave "as if they couldn't help it" or "as if they decided". I am sure this option will inflame all those of analytical temper, and those who conceive one main task of science to be the killing of the dragons of anthropomorphism and pathetic fallacy. But this is not a fallacy, it is a conscious fiction: it is not a *mystique*, but, if you will, an *heuristique*. Pask is not maintaining, I trust, that elements must and do will and want and choose; to do so is as absurd as maintaining that they do obey a strict and discoverable determinacy. He is saying, I hope, that there are certain bewilderingly complex systems of which we can get a *working* knowledge, a successful interaction, on the assumption of wilful internal direction. He is suggesting that a relational approach, while not giving us the wealth of data, and the certainty of state that the



analytical approach might give us, is justified if it gives us what we need, not just what we think we have to have.

Such an approach is of utility only if our concern is to get on with the matter, only if our concern is with winning, not just with playing the game. If we consider that explanation or understanding can consist only in the analytical and atomic dissection of the matter, then his technique is as useless to us as is a tennis net to a chess player. But if our concern is a working relation, a successful interaction with systems called self-organizing, then his technique is promising.

Yesterday Stafford Beer cited one of Gordon Pask's sludges as performing the inversion of a matrix of just sub-astronomical order. I say that it did, and it did not. It did, in that if we, with or without computer aid, were required to give analytical reproduction of the operation performed by the sludge, we should have to have recourse at some point in our computation, to the mathematical operation known as matrix inversion. It did not, because at no time, in no place, was there a discernable sub-region of the sludge that could uniquely be identified as a matrix inverter.

If we are to regard knowledge as a game played in certain fixed ways with certain fixed counters, then it will remain uncomfortably constrained.

#### SEPARABILITY, DESCRIBABILITY, AND FUTILITY

I have some minor prejudices that I neglected to state at the outset. They are minor in that I shall not mind terribly being wrong, but like most minorities, they color the lump more than their proportion would indicate. I am prejudiced, for example, against the assumption of high orderliness in the universe, and the assumption that whatever order there is, it is knowable within techniques now available to us. I believe that we stand in relation to the world more as inventors than as discoverers. As Suzanne Langer<sup>(9)</sup> observed, "Our world 'divides into facts' because we do divide it". We have a choice of divisions, a variety of ontic decisions, but our choice is not unlimited. We can posit no kind or degree of order beyond that which we possess. We are finite sub-regions of a universe that may be either finite or infinite, hence Ashby "variety" or Shannon "information" can only apply to us,

not to the world. Surely the central core of both of these concepts is the idea of a message set—a finite set. The variety we may express is always less than that of the world. What we express cannot hope to be more than a homomorph of the world; only because we can choose (or are constrained to choose) among several homomorphic mappings, several descriptions or interpretations, can we be considered as self-organizing or order-detecting.

It will follow from this that I am prejudiced against much of the analytical temper. Sir Herbert Read made the distinction that the artist is concerned to *represent* the world, the scientist, to *explain* it. I see scientific explanation as an excessively ruly mode of representation. Henry Eyring commented that the scientist is not concerned with being right, but only with proof. Perhaps our concern for reliable knowledge has blinded us to the utility of unreliable knowledge, to the embarrassing fact, now troubling those in the artificial intelligence game, that you have to start with some knowledge in order to assess its reliability. And our assumption that we know the criteria for reliability, that they are thus and so, has led us to the fallacy I shall call the fallacy of a single frame of reference.

I am sure we should all laugh at the idea of a small, finite set of rules that we could write down on a card small enough for vest-pocket or purse, which should be applicable in all generality to every specific case. Yet among those of us who laugh, are those who not only are devoted to the idea of some universal frame of reference, but also are convinced, in anticipation, of what its form shall be, and of the procedures by which we shall find it.

The method and form they refer to is that of analyticity. This faith is equivalent to that in the universal language; the belief that there can be an extensional language in which what cannot be the case cannot be said; a language in which any licit combination of its symbols yields an expression representing an actual or possible state of affairs in the referent world; a language in which any expression may be uniquely decomposed into ultimate elements which represent the necessary and sufficient atomies of the referent world.

Some of us have believed that mathematics is this language, but mathematics is overdetermined; it states many, many cases to which no reality can be made to fit, nor does it furnish, within its

formal structure, the means by which we may discriminate between the actual and the possible. To make it at all useful, we must frequently go outside it, back to the empiric empire. Those who resent such excursions are mad, or mere gamesmen. *Vide* Wittgenstein,<sup>(5)</sup> no mean analyst himself: "The idea is to express in an appropriate symbolism what in ordinary language leads to endless misunderstandings . . . . Now we can only substitute a clear symbolism for the imprecise one by inspecting the phenomena which we want to describe . . . . An atomic form cannot be foreseen. And it would be surprising if the actual phenomena had nothing more to teach us about their structure".

The notion of analyticity implies a belief in the existence of a single frame of reference, a universal context. One of the essentials of this notion is that of independence and separability. One way to look at the history of the physical sciences, particularly physics, is to view it as a search for separability, as a hunt for that reference frame that permits the independent variation of the elements. We have had a remarkable success at this sort of game; perhaps that is why we treat it as panacea. To abort a quotation I shall deliver full term later, "we are made to accept as truth that which is only advantage".

I am not a philosopher, and I rarely aspire to that irreversible state. I do not really care whether the world is necessarily analytical, or, what amounts to the same thing, whether what is reliably knowable is limited to the analytical. I do become fretful, however, when it is proposed that as a matter of course, we should be limited to analytical interaction with the world.

Now it is a matter of experience that separability is a matter of context. We find that we can vary quantities independently only if we choose the appropriate context. What are we to do if we cannot find, or have not yet found, the appropriate context? The answer, of course, is that we use various makeshifts. We pretend that certain "weak" parameters are not there, we employ approximations. We find that if we ignore certain factors or connections, the remainder becomes separable. That is, we repartition the world, we look for a new system in which separability is possible.

It is true that each of these defective models is true within a limited context. What I suggest is that the process of thinking involves coming to terms with a world that is analytical *only* in

restricted contexts. Our task is not to find the context under which an organism can render the world intelligible, but rather to find how it renders the world intelligible under a flux of contexts. It may be useful to consider the organism not so much as a device which tries to reach that state wherein the world, as disturbance, is fully regulated, but rather to find out how a device, restlessly wandering among contexts, can make sense of a world that is true under none of these, but somehow understandable under several, serially or in parallel.

### SEEING AND BELIEVING

Wittgenstein, speaking of psychology and of set theory, said that the former has experimental method and conceptual confusion; the latter has conceptual confusion and methods of proof. I would say the same about our treatment of self-organization, or, more particularly, our approach to the modeling of the "higher" mental functions.

Few things inflame me more than the all-too-general practice of constructing "seeing" machines, "thinking" machines, "chess-playing" machines, "translating" machines, and, when it is pointed out that this is not seeing or chess-playing or thinking or translating as we perform them, of claiming that there was no serious or obligatory intention of modeling our "seeing", etc., but only of showing a possible or alternate mode of "seeing". It is difficult for me to understand just what this means. I am tempted to take the dogmatic view that there *is* no alternate "seeing"; that all we know of seeing, all its criteria, are applicable only to *our* seeing, or to that which we extend by courtesy to those creatures with which we can identify ourselves.

Of course I am willing to grant that what such a seeing machine does is *like* seeing, even as I am ready to grant that there are machines that do something *like* playing chess, or thinking. But I am also ready to grant that hallucination is *like* seeing; that dreaming is *like* the perception of a public world; or even—and this is the clinching case—that thinking one knows the answer is *like* knowing the answer. We must constantly remind ourselves that we know very little about our own seeing or thinking or chess playing.



Now I believe that we shall be able eventually to produce thinking machines and the rest, distinguishable in their behavior only by being built, not born. But I do ask that the test of their success in doing so will be marked by our inability to treat them other than we treat their fleshy counterparts; if we lower our voice in the presence of the *thinking* thinking machine, attempt to cheat, or bluff, or rattle the chess-playing machine, and worry about the problem of censoring the input and output of the reading machine.

The test will surely be operational: the machines will show the claimed function only if we extend to them the interaction appropriate to that function. I reiterate: if I can deal most efficiently with a machine only on the assumption that it thinks; then it thinks.

And the test will surely be extensional: we will not allow that a machine thinks unless it is freely exchangeable with an indubitably thinking creature in that wide range of situations that would, for that creature, require thinking.

I think part of our confusion on this score comes from equating all thought with logic. In our eagerness to believe that there is an analytical expression for everything, and to find that expression in a particular case, we mistake the nature of these "higher" human activities. Even the demonstration that what *we* should have to think to do, can be done by a strictly deterministic mode, by a combinational machine, for example, says only that any finite behavior, if it is fully specified, can be reproduced by a machine. It can give little or no insight, however, into *our* processes.

Let us say that over a finite period of time, I have categorized sheep, lambs, goats and kids as members of the same set. Now few of us would doubt that it is in principle possible to build a machine that effects the same categorization. But now it turns out that I also include hydraulic rams in this categorization. The question is, how similar to the original machine would a machine be which also includes hydraulic rams? Put in another way, how should we go about building a machine that, with only the initial knowledge about sheep, lambs, goats, etc., will also categorize hydraulic rams accordingly? We have tended to shy away from problems like these, and to concentrate our attention only on those tasks where objectification can be taken for granted; where,



wrongly or rightly, the criteria for decision can be considered to be given in the "properties" of the input universe.

### THE GAME OF $2^{2^N}$

To begin one consideration of constraint, I shall use the Petrine method, and deny it at least thrice. The framework within which I shall do so is the game of  $2^{2^N}$ . I am well aware that this game has been considered a legitimate problem; I hope to show that if it is a problem, it has no more standing in the world than a chess problem, and even less verisimilitude.

Even if we take the utilitarian view of games as rehearsals for reality, it must be recognized that even a dress rehearsal differs in important respects from a performance. Usually the difference lies in the degree and kind of constraint. The game of  $2^{2^N}$  is tainted as a representation of reality because it is so little constrained, only the game of  $L^{M^N}$  appears freer.

We begin with a player, a black box; and an opponent: Nature. The box has  $N$  inputs and one output. At each moment, Nature may render the inputs independently active or inactive. There are then  $2^N$  distinct input patterns possible. For each of these inputs the black box may have its output active or not. There are thus  $2^{2^N}$  distinct ways in which the box may associate output states with the input patterns.

The game is not wholly free of constraint. We must assume, for example, that Nature knows clearly what she wants in the way of a response on the output of the box; she must be consistent about the function she has in mind. We must also assume that she is harsh and relentless; one mistake on the part of the box, and she wins.

Now there are several ways of playing the game from this point, differing in the kind of black box we use. In the Economy Edition, we are provided with a box of fixed or single function, considering the output of the box as a Boolean function of the  $N$  inputs as two-valued variables. In this version, we put down our box, Nature fiddles with the inputs, and we pray that the red light indicating error will not flash this side of eternity.

In the Delux Edition, we have a box with a  $2^{2^N}$ -position switch on the side, each position selecting a different function. We thus

are allowed a preliminary choice of function, but the game is played the same thereafter.

The Masters' Edition adds a means of preprogramming the switch positions as any function of the current input and output states. This is equivalent to saying that the box may be programmed as any sequential machine with  $2^{2^N}$  input states and the same number of internal states. My mathematical informants tell me, with a low level of confidence, that there will be  $M^{M^2}$  such sequential machines or programs, where  $M = 2^{2^N}$ . To those who thrill at the sheer magnitude of  $2^{2^N}$ , this new number should give them sensations unendurably exquisite:

$$2^{2^N} (2^{2^N})^8$$

At the risk of diluting their pleasure, I must point out that many of these sequential machines, while distinct as matrix representations, do not differ ultimately in their action.

Whichever edition of the game we buy, the game in essence consists in estimating the odds of survival of the black box. It is curious but true that any box has a finite chance of surviving for a finite time. While Nature is constrained to have only one function in mind, she is not constrained to any ergodicity in her presentation of patterns. For the case where  $N = 4$  and where she expects only one of the sixteen possible patterns to be affirmed, and all others denied, the box set at a total denial can survive forever if she never presents the pattern to be affirmed.

Even more curiously, there is no advantage whatsoever in buying the more elaborate versions of the game; they are merely come-ons, appeals to snobbery. So long as nature is relentless, and punishes capitally for single error, no amount of preprogramming or feedback reprogramming can avail. In order to go right, one must occasionally be permitted to go wrong without immediate disaster.

Now I have tried to render this game absurd by describing its play. Judging from his remarks following Beer's paper yesterday, Dave Willis intends to attack it structurally, showing that even if Nature were less adamant, or ground more slowly, we should still be unable to build such a box when  $N$  is, say, five or more. Thus there are two ways of reducing the game to an absurdity.

I have said that this game denies constraint. Constraint was denied when we permitted Nature to choose just any function, and to present the patterns in just any order. It was denied again when we required that one error meant death.

Treble, and most strongly, it was denied whenever there was an implication that the game has anything to do with the real problems of successful adaptation or survival of a real organism or device in a real world. Take just the matter of there being  $N$  predicates, all independent. I should be surprised indeed if any of you could demonstrate even two such predicates, such that one can be altered and all else remain the same. We most certainly can talk of, and perhaps conceive of, pink elephants and mauve crocodiles, but it is only in talk and conception that such differences make no other difference.

In a world suddenly without elephants, all else may still appear to be the case—for a while. But a world without elephants is a world without elephant diet and elephant dung. The very shapes of some African trees would alter, and we should at least be put to the trouble of renaming such items as elephant flies and elephant grass. (This brings up the parenthetical question of whether a re-labeling is a difference. For man, it certainly is; it may be *all* the difference.)

I think we may safely say that the ultimate independence of a finite number of properties is just not the matter of concern. Each of us knows that there is a sense, a context, in which two properties are unconnected and a sense or context in which they are connected. Yet who can maintain that there is that ultimate, most general context in which some finite number of properties lie all unconstrained?

Similar considerations restrict the freedom of Nature to present patterns in just any sequence, without temporal constraints. Only where we can impute something lawful about their succession, can we hope to make sense of them and survive. It is appropriate to point out here that since the permutation group is not cyclic, and thus cannot be generated by a repetitive serial operation, a world with even the weakest of succession constraints cannot generate all patterns with equiprobability, and thus would have improved habitability, since it improves the black box's chances of not encountering those patterns for which its computation is wrong.

Another matter that renders the game inapposite to reality is that it provides no means for the black box to react on the world. None of its efforts can have the slightest effect on the future sequence of patterns it receives. There is no possibility of interaction, no reciprocity. The game thus requires that the box not be *of* Nature, but somehow aside from it. We can beg the whole question by pointing out that one cannot confront an arbitrary black box with an arbitrary world. Any real black box comes out of its world, and shares at least some of its particularities. Its very texture will reproduce some of the constraints on the world, and it will have that much of a head start in adapting to a world to which it is already partly adapted.

I can think of nothing crueller, or stupider, than a Nature which would produce children armed only with logic. I can think of nothing more salutatory for us to realize than that great hunks of competence, in terms of a particular world, are inbuilt in every organism. That which is most invariably so about the world, its rigidest constraints, are not left for the organism to discover, but are given as *a priori*s, embodied in the machine. Such things as the willingness to make induction, preferences for certain measures of similarity, predilections for certain partitionings of the world, are in the structure of the machine. The machine does not search through  $2^{2^N}$  possible worlds—most of these it cannot conceive of as possible, and need not conceive of as possible.

Each organism lives in the world it *had* conceived of, or the world its ancestors conceived of.

In sum, I am at a loss to know why the game was ever seriously proposed as a problem, if indeed it ever was. Can it be that there are those who believe in a world of  $N$  properties to which a real device can ultimately adapt? Are there people who think that this game is something that living organisms have played and solved?

#### EPPUR SI MUOVE

One of the general observations on living systems that bears emphasis and re-emphasis is that all or almost all of the constancies seen in such systems are there only through dynamic maintenance. It is easily forgotten that not only the functional constancies



(behavior) but also the formal constancies (structure) are there only through the active interplay of several variables. Thus the level at which homeostasis is present is often lower than we would assume.

That you and I stand erect is not because we are rigid bodies that can be balanced on our bases. We stand erect because of the continued tonic interplay of a numerous system of muscles. That a living bone has a constancy of form does not indicate that it has been cast once and for all in that form, but must instead be referred to steady states of the constant war of osteoblast and osteoclast.

If one were constrained to model these low and inobvious levels of homeostasis in electronic hardware, it would be required that not only the so-called active elements be dynamically maintained in certain states, but also the passive elements—including the wires, knobs, and chassis. Such a model could not be turned off with impunity; to do so could destroy its validity as a particular structure.

I think that this is a consideration that, ergodically, has crossed each of our minds at least once. I find it curious, however, that we take it so little to heart. When we approach a biological system with intent to model—either in explanation or exploitation—we doggedly proceed to render most of the dynamic constancies into static equivalents, or where this is not easily possible, we remove the maintenance of the dynamic constancies from the purview of the model.

When we model a neuron, we note that the interior of the real neuron shows a largely constant resting polarization relative to the extracellular medium. We then supply such a resting potential by a battery or other power supply of equivalently constant output, but we make quite sure that the maintenance of that polarization is in no way a function of the states of the neurone model. Such a procedure is permissible only if we can show that the important or significant activities of the neuron—those we wish to model—are indeed independent of both the degree and kind of regulation in its subcellular mechanisms, or independent of extracellular influences other than neural input.

In other words, in order to separate these low-level constancies, we must show that they are not under neural control, or that



they are under the control of neural subsystems independent of the one of our concern. We have the mathematics, the techniques, of separability so readily at hand that I sometimes suspect us of studying under Procrustes, of being more willing to force the elephant into the teapot than to find a container appropriate to elephants. While it is all very well to search doggedly for that frame in which a selection of variables are separable, it is much less praiseworthy to pick the frame in advance, or to demand that there be such a frame.

Sometimes I feel that we have missed an important point in our pants-wetting eagerness to swipe goodies from the biological grab-bag. The point is, it is not a grab bag. The items therein show connection, and to attempt to draw out just one glittering generalization entails a host of contradictory loose ends. Apparently many of us believe that we can model the "Higher Mental Functions" without regard for the lower mental functions, or indeed, for the non-mental functions of whatever rank. The brain, or its "interesting" parts (if the word "part" may be admitted at all) is an integral part of a larger system, not only integral with, but most likely specific to, the remainder of that system. To withdraw it blindly will give us something quite other than what we hoped for.

Even when we admit that the nervous system cannot be plucked whole, like a hairy turnip, and replanted successfully in a different soil, we do not go on to note that the great bulk of the nervous system is not dedicated to those "useful" activities we hope to replicate. It is not devoted to the luxuries of pure thought on arbitrary problems, but rather to the restricted dirty work of daily life—maintenance and coordination of some dull and inobvious little constancies. In fact a strong case can be made that the "higher" mental functions are strictly in the service of some pretty grubby lower activities.

A quotation from Albert Szent-Gyorgi states the case: "The brain is not an organ of thinking, but an organ of survival, like claws and fangs. It is made in such a way as to make us accept as truth that which is only advantage. It is an exceptional, almost pathological constitution one has, if one follows thoughts logically through, regardless of consequences. Such people make martyrs, apostles, or scientists, and mostly end on the stake, or in a chair, electric or academic".

I think it is about time for us to admit that much of what we want machines to do, much of what our present techniques and understanding will inevitably lead us to, are functions that at best, map only into pathological man. One can even state the pathologies we are aiming for: they are the pathological states of being a scientist, and of being a slave. Insofar as we want both, we shall have to look elsewhere than in man for most of our design clues. I am saying no more than what Ross Ashby said last night, that we must have regard for the essential selfishness of adaptive systems, meat or metal, born or built.

I am of the same belief still, that we *can* build "self-organizing" systems of utility to ourselves. There are only two restrictions: one, that we must insure that the utility is one that is shared between us and the system, and, two, that to "build" such a system does not imply that it can be done by the exhaustively specifiable designs that presently accompany our ideas of practical synthesis. These will have to be systems that are constrained to a functionality, not synthesized to one.

#### PLUS CA CHANGE . . .

The idea of adaptation has a teleological taint. As a progression toward aptness, it suggests betterment, so that it is difficult to take a pejorative view of adaptation. I suggest that we replace it with the idea of equilibration, save where we are quite sure we intend betterment, and can state its context. It is to be understood, of course, that equilibration covers steady states, cycles, dynamic equilibria of all sorts.

To speak of the equilibration of a system is to say that there is a change that does not impair its ability to be recognized as the same system. Hence equilibration is a relational property of the system, hence not a property of the system alone, but also of our regard of the system.

There is an old philosophical rubric: "Titus is taller than Caius". The truth of this statement is impaired if Titus wanes, Caius waxes, or both. To say that a system persists, yet adapts, says one of two things: either our definition of the system is not exhaustive and finite, so that the system may have at least two states and still be the "same" system, where one of the states *is* better in a context:

or it is a statement that our standards of what constitutes the system, the contexts in which we can regard the system, change in the course of observation. "If we want things to stay the same, we'll have to make some changes."<sup>(4)</sup>

I submit to you that this latter condition is the one operative when we talk about "self-organizing" systems. The dynamism, the alteration of order, is not purely a property of the system, but may just as well reflect changes in the contexts with which we regard the system. The cybernetics of self-organizing systems is the study of changing contexts.

I suggest that this interpretation may explicate several problems of order and organization, particularly some of the peculiarities of perception, of the evolution of a homomorphic representation of a world incompletely specified and constantly changing.

It may be applied at many levels. Take the linguistic curiosity cited by Steve Sherwood yesterday: "The cat washes itself". To construe this in blatant extensionality gives it this form

$$a R b$$

then, noting that surely  $a = b$ , we permit the extensional substitutions

$$a R a$$

$$b R b$$

neither of which conveys the sense of the original. Identity is not a transparent concept; its validity also depends on context. Frege's characterization of it:

$$[a = b] \equiv [(AF)(Fa = Fb)]$$

gives us a view of identity both Olympian and intolerant; it is a matter of experience that we posit identity safely in cases that fall well short of his strict requirements. We do so safely by a tacit change of regard; a situation is now recognized as syncategorematic, now as atomically extensional. It is only when, as Wittgenstein says, "language is idling", that we construe "The cat washes itself" as  $(a R b)$ ; when language is working we construe it also as  $(R_b(a))$ .

In a frame perhaps more pertinent to cybernetics, we clamber bewilderingly up and down the lattice that extends from the system

with all its states distinguished to the system with all its states merged. Many of us feel that there must be one position on the lattice that is the "right" one, that there is a partitioning of the system, and a functional relation among the elements of the partition, that is correct. I suggest that finding such a position is not important, that there are too many situations of concern where we cannot characterize a system by a single partitioning or a single function. We cannot characterize all situations by stating what is invariant over their states: we must often examine how the definition of their states may be altered.

Take the statement, "I recognize Caesar". This is usually taken to mean that in a wide but limited range of circumstances, all different, I can pick out, identify that pattern we call Caesar. Or, in the jargon of the profession, I can pick out Caesar as something invariant over alterations of size, position, distortion, etc. Now this suggests that there is something invariant in the patterns of which we say "Caesar is there". This in turn suggests that we may, by rational examination of the presented patterns alone, come to discover those configurations categorized as Caesar. Then all we have to do is back up an artificial sensory apparatus with a logical net rendering that which is Caesar's.

Now it is certainly true that for a wide variety of observations on perception and recognition, such an objective construction only over the pattern elements themselves is possible. I question whether this is the universal, or even the usual situation.

You may remember the alternate construction that Oliver Selfridge<sup>(6)</sup> gave to pattern recognition: "A pattern is equivalent to a set of rules for recognizing it". As I interpret this, the fact of recognition involves what the organism must do in order that a pattern fit a class. Whether or not the device has a *choice* of activities, of operations it may perform on a sensory image, is not important. What is important is that the device need not be a combinational machine during pattern recognition. It is not restricted to a single context, a single mapping function, over all the patterns which are called Caesar. Its context of similarity need not remain constant throughout the processes we call recognition. To find that which is Caesar does not imply that there is a single internal state specification such that all those Caesarian patterns are mapped into the set of responses called "seeing Caesar". That



we can detect that a line has a break in it in no way implies the presence in us of a uniform and constant break detector, nor even that the family of internal states that permit break detection have an obvious similarity. I think it is obvious, for example, that tachistoscopic performance, and the normal, more leisurely performance, are linked, identified, only nominally. Seeing in flashes and seeing at leisure, are linked only descriptively. That we describe both as "seeing Caesar" describes no functional unity over the two processes.

Here again we return to the idea that the perceptive and adaptive organism regards the world under contexts, applied simultaneously and in series. Identification, recognition, are not phenomena within one context, but over several. Our job should be to discover the nature of the contexts under which we view the world, to find if they are finite in number, to discover the rules for their sequence. I am tempted to identify changes of context with the step functions of Ashby,<sup>(8b)</sup> but this should wait for closer examination. I suggest that perceptual processes involve just such sets of competitive or alternative mappings, and that the organism searches for that coupling, serial or parallel, that names the pattern.

I mean "name" quite literally. If the recognition of a pattern is equivalent to a set of rules for recognizing it, then the application of the rules results in the labeling of the patterns. We, and our animal friends, do not categorize just for the sheer joy of pigeon-holing, but in order to create the simpler situation where we need only deal with the *labels* on the pigeon holes, without fret for the variety within. We are, in Medawar's terms<sup>(8)</sup> "elective" devices; the variety we show does not indicate the amount of selection we exerted, and may be greater.]

The surprisingly low channel capacity of neural channels again suggests that the message set from which neural messages may be drawn, cannot be discovered just by examination of the messages nor our response to them. A given string of pulses and not pulses, simultaneously interpreted under several contexts, perhaps obtained by variation of threshold, can convey less than appears to be the case, for any one context, and more for all. The code of the neurones may have sufficient constraint to be a Baconian cipher; a message is not *a* message, but several,



dependent upon context. And if neural noise is grey rather than white, so that under one context or another the message is relatively noise-free, the organism may range among contexts until it finds that one relatively free of ambiguity. But these are items for fuller development elsewhere. I hope here only to suggest that there is not *a* neural code with a single decoding specified, but rather that there is a schema which may be interpreted in overlapping but different ways.

#### SUMMA SINE LAUDE

In going over what I have said, it appears that I am fonder of denial than affirmation. I am tempted to defend the greater validity of nay-saying than yea-saying by noting the sufficiency of Sheffer strokes. But perhaps I should ask the physicians present: Just what is the mortality rate for Sheffer strokes?

Having paraded my prejudices and outlined my bigotry, I should like, as farewell, to demonstrate my tolerance. After all, some of my best friends are cyberneticists.

Sommerhoff<sup>(6)</sup> quotes Quine as saying that the younger a science, the more its terminology rests on the uncritical assumption of mutual understanding. While this is something I must deplore in principle, I am not sure but that this sin of youth is necessary, even as youth is. I must admit to a certain tolerance for uncritical interaction. Rigor applied too soon may be rigor mortis.

I should like to end as I have progressed, anecdotally. For the first item, I shall have to commit the sin of anachronism. This presentation of mine was given by title only during this symposium, but as editor of the proceedings, I have had access to material after the fact, namely, the errors of transcription made by our stenographers. Some of these I regard as truly creative and instructive. I was much impressed when I found "Lebesgue integral" rendered as "a vague interval", but I reserve my admiration for the alteration of "Caratheodory" to "paratheoretic".

I heartily recommend this neologism to your use. For its justification by parable, I give you the observation that the natives of Mont St. Michel have developed a technique for crossing quicksands. This consists in stepping quickly and lightly, never

letting one's full weight bear for long on any one small area. Gentlemen, I give you paratheory, that will not bear critical weight at any point, but which yet suffices to carry one from here to there.

And for a bedtime story tonight, I suggest Hans Christian Anderson's "The Emperor's New Clothes". Fellow emperors—and weavers—may we become as little children!

#### ACKNOWLEDGMENT

The work on which this presentation was based was done under Air Force Contract 33(616)-6428 and Office of Naval Research Contract Nonr 1834(21).

#### REFERENCES

1. B. DELISLE BURNS, *The Mammalian Cerebral Cortex*, Arnold (1958).
2. KARL R. POPPER, *The Logic of Scientific Discovery*, Basic Books (1959).
- 3a. W. ROSS ASHBY, *Introduction to Cybernetics*, Wiley (1958).
- 3b. W. ROSS ASHBY, *Design for a Brain*, Wiley (1950).
4. G. DI LAMPEDUSA, *The Leopard*. Pantheon (1960).
5. LUDWIG WITTGENSTEIN, Some remarks on logical form, *Aristotelian Society Supplementary Vol. IX*, 163-4 (1929).
6. OLIVER SELFRIDGE, Pattern recognition and learning, *Third London Symposium on Information Theory*, Academic Press (1956).
7. SOMMERHOFF, *Analytical Biology*, Oxford University Press (1950).
8. P. B. MEDAWAR, B.B.C., Reith Lectures (1959).
9. S. LANGER, *Philosophy in a New Key*, Harvard Univ. Press (1942).

**ALBERT B. J. NOVIKOFF**

*Stanford Research Institute, Menlo Park, California*

## INTEGRAL GEOMETRY AS A TOOL IN PATTERN PERCEPTION

This talk will be largely didactic, explaining what is in a sense a well-known part of the mathematical literature, but a part of the mathematical literature that I hope is not so well known that I am telling people only things they already know. I hope, therefore, you forgive me if I take what may seem to be a patronizing tack in the didactic portion, but I would rather be clear about a few simple things than address myself to profundities with only a lick and a promise. Then there is a portion that is not purely didactic, namely my proposal that this chapter of the mathematical literature may be of use in designing pattern recognition or pattern discrimination devices.

This work all began because of an attempt to clarify a detail concerning the behavior of Rosenblatt's Perceptron. I was told that the way in which the Perceptron would distinguish between, say, a circle and a square of the same area was, roughly speaking, to have in its own internal *modus operandi* a replica of, say, the circle with which it would go hunting around the retina, checking for overlap, and that in some sense circles prefer circles to squares of the same area, with regards to the amount of overlap. Someone had given me this idea—I do not say Rosenblatt because I myself have never been able to read the Rosenblatt reports and rely on “filters” to acquire their content, and in fact one of these “filters” was in error.

Now I smelled a rat concerning that statement and looked into integral geometry as the subject in which to find out precisely what is true about the overlap between circles and circles as opposed to the overlap between circles and squares. I did indeed find the rat,

but the body of the rat is not anywhere in the Perceptron. Perhaps in the remarks after the talk we can clarify what it is the Perceptron actually does do, because there is a connection. That at any rate is how this inquiry all began. The device I propose has a very specific and modest end in view. We suppose that there is a prescribed alphabet of patterns, "characters", which are to be recognized, and that the designer of the device is very well aware of what this alphabet is, he has at his disposal as much geometrical description as he requires. However, he doesn't know the location and orientation of the pattern on the retina of the device. He wants to be able to discriminate, to be able to tell which of the alphabet has arisen, and my remarks concern how to design such a device.

Now I will give an instance of the kind of technique of pattern recognition which I will employ repeatedly. This example is chosen chiefly because it is intuitive; it rests on a theorem of integral geometry, as does everything else that I will say, but this particular example rests on a theorem which is conceivably popular and well known, although I do not think that most of the other theorems I exploit are popular. This example is in fact not especially simple to fit in the framework of integral geometry as a whole, but taken by itself it is a very simple one to understand.

Suppose that the two patterns to be recognized are both an infinite grid of infinitely extended parallel lines and the grid width of one pattern  $d$  and the grid width of the other pattern is  $d'$ . Assume  $d < d'$  (Fig. 1). How might we go about recognizing which of these two patterns is presented when we do not know the orientation of the pattern? We recall now a famous theorem which goes back to 1760, in common with a lot of other bright ideas, and which is usually associated with the name of Buffon, under the name of the "Buffon needle problem". Stated in the customary language of geometric probability, this states that if a "needle", that is an oriented line segment, of length  $l$ , smaller than the grid width, or at least no larger than the grid width, is tossed at random on a grid, the probability that the needle intersects a grid line (and does not just fall between), is  $(2/\pi)(l/d)$ . Then the designer of the device picks a needle whose length is the smaller grid width  $d'$ , tosses it repeatedly and averages the number of times the grid has crossed the line. If he gets a number which closely resembles  $2/\pi$ , it was the smaller grid width he was dealing

with. If he gets a number which approximates  $(2/\pi)(d/d')$  or which is suspiciously less than  $2/\pi$ , he will report, "I was looking at the coarser grid".

It is clear that this method, roughly described, is independent of orientation and location of the grid. If the method of looking is itself independent of orientation and translation, the object being looked at can suffer orientations and translations.

We could actually outline a precise statistical test of hypotheses that the designer would go through in order to make his decision.

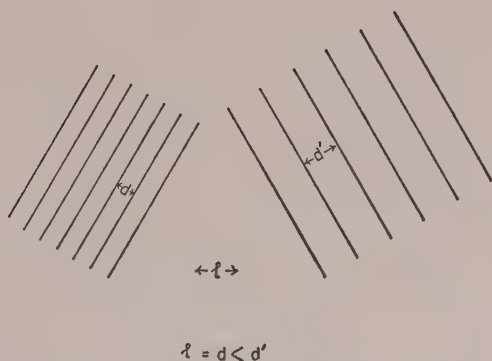


FIG. 1

I will omit these hypotheses testing and decision-theoretic questions. The above summarizes the idea of how a designer could in principle make such a decision.

Now observe the following facts. One, I have already observed. The method is independent of orientation and translation. Second, the designer should be able to make some estimate of how many tosses of the needle are required in order to get an answer of a given reliability. This, however, is one of the statistical questions that I will not address myself to further, but answers—sensible statements—can be made on this question. (Unfortunately, some of these sensible statements require the use of higher moments of a random variable, not their first moments, and for many of the theorems I am going to describe only first moments are available; however, this simply indicates there are some outstanding useful



problems to be solved; I do not think it is a major defeat in my proposal.) The most important thing to notice is that this method requires the knowledge of a theorem. We know the theorem of Buffon, which tells us what number to expect when we know what the pattern is and when we toss a needle upon it, i.e.  $2/\pi$  and  $(2/\pi)(d/d')$ . The more such theorems are available, the more recognition devices are possible, sensitive to more and more interesting patterns.

For example, it can be shown that one need not toss a needle, that is, an oriented segment, of length  $d$ . One can toss an arbitrary oriented, rectifiable curve of length  $d$ , for example a circle of radius  $d/2$ , and again, the probability of intersection is just what it would have been for a needle, strange to say, so the flexibility of device mechanisms varies with the amount of theorems we possess. I am making a case, in other words, for storing a large number of theorems of this character.

Another remark. In this case there was no "retina" upon which the character was displayed (or rather the retina was, unrealistically, the whole plane). In general, recognition devices have sort of edge effects due to the fact that the character or pattern will be presented in a certain finite location. How do we take that into account? I will have more to say about this later. In order to construct a physical device which performs pattern discrimination according to this idea, you want a device which can, first of all, toss a needle at random. It should be able to tally intersections. Now questions as to speed of repetition and the ability to tally I regard as questions of an engineering nature and I will say no more about them, but what it means to toss a needle at random is not an intrinsically clear notion without further remarks.

I will conclude this introductory portion by giving an instance of the perils of loose reasoning in the realm of geometric probability and this instance is, again, offered solely to orient the listener. It too, is very well known, but I think it will serve to warn you why we must be clear about the notion of tossing objects at random before we can proceed to design devices based on that notion. This example is known as the Bertrand paradox, and is of mid-nineteenth century origin. We wish to toss a chord at random on a circle, and we ask, what is the probability that the chord should

have a length which exceeds the length of the inscribed equilateral triangle. Of course, there are many inscribed equilateral triangles, but the side of an equilateral triangle inscribed in the circle is independent of the choice of the equilateral triangle. The problem has the following solution, according to some. By reasons of symmetry you may regard one end of the chord as fixed so that it is only the location of the second end point which determines the

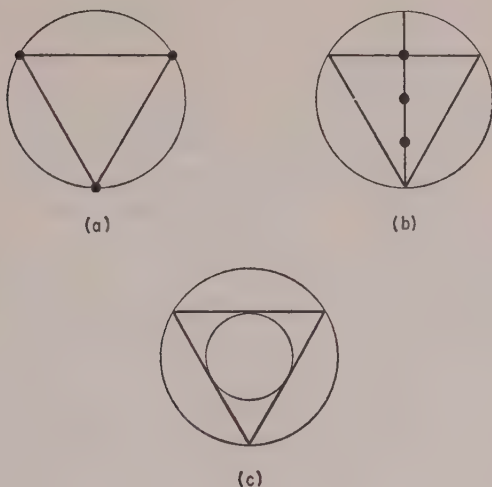


FIG. 2

chord. The circle can be trisected into three arcs, each of which is equally likely when we toss a chord at random (Fig. 2a). If the second end point lies in either of the two adjacent arcs the chord is not a "successful" one, but in the opposite arc it is. The three arcs being equally likely, the probability is one-third.

The second solution is this. By reasons of symmetry we may regard the angle of the chord as fixed and proceed to examine only one family of parallel chords, taking as fixed the common diameter that they are perpendicular to. We ask what fraction of this family is successful (Fig. 2b). In this case an elementary study of what an equilateral triangle inscribed in a circle looks like, shows that if you quadrisect that diameter, the chords lying

in the extreme two-fourths, whereas those crossing the central two-fourths are "successful". Thus, of the four equally likely divisions of the diameter, two of them are successful and two are not. So the probability of success is  $1/2$ .

Finally, the length of a chord is completely determined by the location of its center. Now we examined which locations for the center are successful cases and which are not. We see that in a concentric circle of a radius  $r/2$  lie the centers for a success (Fig. 2c). The probability for a point chosen at random in a large circle to lie in a small circle is given by the ratio of the areas, which in this case is  $1/4$ . Summarizing, we get the answers  $1/2$ ,  $1/3$ , and  $1/4$ , and others can be provided.

Each of these answers corresponds to an appropriate physical device for generating random chords. If you build the device by imitating the method of choosing one end-point fixed and the other one at random, the answer  $1/3$  is correct, and you can build the device in that fashion. In fact, I will not bore you with this, it is clear you can build physical devices motivated by the different notions of "at random" that are implicit in the three different solutions. There is not a single notion of tossing a chord "at random" without further remarks. There are many probability measures that can be put on the family of all chords drawn in a circle. Now since I am going to refer to theorems which depend on the notion of throwing a line at random, throwing a curve at random, throwing an arc at random, I must in every instance be precise in what I mean by the phrase "at random", so that there is no ambiguity in the construction of a device based on the theorem. We will not have an ambiguity in the theorems themselves that I will cite, we will ask the notion of randomness that we use in the theorems to enjoy a property which selects out of all possible probability measures that we can consider, precisely one. Speaking now roughly, before I get down to cases, we will require as a condition on our measures that they be invariant with respect to translation and rotation. In the case of the chords of a circle, which I have just mentioned, that condition would mean the following. You would want a notion of random such that the probability that a chord intersects some figure in the circle would be the same as the probability that it intersects any translation or rotation of that figure anywhere else within the

circle. (Of the three "solutions" mentioned above it is only the middle one which conforms to that subsidiary condition.)

Now we turn in general to the notion of a random line, since my first collection of theorems will deal with random lines. Well let us first consider the more basic notion of random point. What is meant by the notion of choosing a point at random in the plane? Here is a case where restricting ourselves to a retina would be a big help. I know what it is to choose a point at random in the retina. I would assume the retina has area 1 and then use area ("Lebesgue measure") on the retina for my probability measure; but what does it first mean to choose a point at random in the entire plane? I must ask this question because many of the theorems—I am not restricting myself solely to retinal theorems—deal with lines drawn on the plane as a whole and not constricted to lie in a sub-region which we call the retina.

Well, let us address ourselves to the more simple question, what is meant by a point at random in the plane? We observe the following: Whatever meaning we want for the notion of random point in the plane, we want the probability that the point lies in a given set  $A$  should be the same as that it lie in any translate of that set or any rotate. I insist on that as a condition that we impose on the measures we look for.

Well, the plane is obviously large enough to contain infinitely many disjoint translates of a given set, so if this one set had a positive probability, the entire plane would have an infinite probability. We have either to give probability zero to every finite set or abandon the search for a rotation and translation invariant probability measure for random points on the plane. That is a fact of life. We do abandon the search, and I will not use the word "probability measure" for the plane as a whole but I will illegitimately speak of "random" when I am not legally entitled to. We will exhibit a measure on the plane, but the measure of the plane as a whole will be infinite, and not 1; it will not be a *probability* measure. That is not cause for alarm. Later, probability measures will reappear, namely, when we have retinas, and in addition, when they do not appear, it is my belief, not yet sustained by experiment, but I believe sustainable, that I can construct devices without appealing to finite measures. I have to perform "experimental" numerical integrations that do not correspond to

finite measures, but they will be of finite integrals. We therefore do not look for probability measures, we look for measures in general. What we want, therefore, is a measure on the plane which is invariant under rotations and translations. Now a measure is, for my purposes, a way of associating non-negative numbers to point sets in the plane, with the following properties. If I have a denumerable collection of disjoint sets,  $A_1, A_2, \dots, A_n$ , the measure of their union—let me use some special notation to indicate they are disjoint,  $\Sigma' A_n$ —should equal the sum of the measure  $m(A_n)$ , that is, we want the additivity property  $m(\Sigma' A_n) = \Sigma m(A_n)$  to hold, non-negative (although it might possibly be infinite for some sets). Also the measure of the empty set (no points) should be zero. These are three conditions we automatically ask for a set function before we call it a measure. And now the crucial fourth condition is this: if  $A$  is congruent to  $B$ , I will require that the measure of  $A$  be equal to the measure of  $B$ .

Given  $A$  and  $B$ , they will be called congruent if one element of the group of rigid motions can transform  $A$  into  $B$ . Now (recall I am discussing measures on points as a preliminary to discussing measure on lines), I can now describe briefly how to find such a measure. Points, you see, have a simple coordinate system. We can introduce the  $(x, y)$  coordinate system in the plane, and now I will show you how we can find a measure that satisfies all of these properties. In fact, you all know how we can find such a measure. Here is the formula for a measure that will preserve these conditions,

$$m(A) = k \iint_A dx dy.$$

(I am skipping over the discussion describing the class of sets  $A$  which are to be regarded as measurable. I am even delighted to do so. I know about measure theory. I assume you either do know about it or do not care about it. In either case, I am justified. Besides, the class of figures that I will apply this to will be elementary geometrical figures, by and large. I have not seen anything in any slide today that would indicate that people are worried about discrimination of non-measurable patterns.)

Now this formula does have the desired property. It does not



describe a unique measure. There is a free constant  $k$  which is adjustable. There arises the question, are there other formulas for measures of points in the plans which will also be invariant under rigid motions? The answer, which is the key to the whole business, is that there are no other measures that satisfy, in addition to the basic requirements of being a measure, the condition of invariance. If you restrict yourself to the first three conditions, there are plenty of measures. For example, something of this form

$$m(A) = \iint_A \phi(x, y) dx dy.$$

will do where  $\phi$  is an integrable function, non-negative, and that even does not exhaust them all. The condition of invariance singles out the functions  $\phi$  which are essentially non-negative constants.

In general, when a transformation group acts transitively on a manifold (that is, can take any point into any other point), there can be at most one measure which is invariant apart from a scale constant. It can happen that there is no measure at all which is invariant, because the group may act over-transitively. That is, it may be possible for several transformations of the group to take a point  $A$  into a point  $B$ . For example, the group composed of translations, rotations and *dilations* acts over-transitively and it is easy to see there is no measure on the plane invariant with respect to that. I have little to say about the recognition of patterns which are allowed to swell, to be dilated.

The preceding is all I have to say about "random points" on a plane. To me the notion of random points on a plane is the construction of a measure function, which happens not to be a probability measure, and the knowledge that there is only this one family of possible measure functions. The invariant measure is essentially unique; the phrase here, "essentially unique", means there is only a free multiplicative constant. Since ratios of measures will appear all the time that multiplicative constant will not disturb us. Now what do we mean by "random line"? I simply want to imitate this procedure. I want to define a notion of measure on the space of all lines. Guided by analogy, it would be a good idea to coordinatize the space of all lines so that I could have something like  $dx dy$  with which to write the resulting

measure. It is easy to coordinatize the space of all lines. (In fact, it is too easy, like the Bertrand problem, which shows that "it is easy" to find chords at random . . . there is more than one way to coordinatize all straight lines.)

For example, you may use the two intercepts  $(a, b)$ , to describe the line  $x/a + y/b = 1$ , or their reciprocals  $(u, v)$  to describe the line  $ux + vy = 1$ . The preferred system of coordinates, however, is  $(p, \theta)$  that occur in the normal form for the line

$$x \cos \theta + y \sin \theta = p.$$

Here  $\theta$  and  $p$  are the angle of elevation and distance to the origin of the line respectively. Among the various line measures we might attempt are those which assign to a set  $A$  of lines the measure

$$\iint_A dadb, \quad \iint_A dudv, \quad \iint_A dpd\theta$$

respectively, or more generally

$$\iint_A \phi_1(a, b) dadb, \quad \iint_A \phi_2(u, v) dudv, \quad \iint_A \phi_3(p, \theta) dpd\theta.$$

There are other possibilities as well.

If

$$\phi_2(u, v) = \phi_3(p, \theta) \frac{\partial(u, v)}{\partial(p, \theta)}$$

the last two are equal, etc. Generally we abbreviate the notation and write  $dadb$ ,  $dudv$ ,  $dpd\theta$ ,  $\phi(a, b) dadb$ , etc., to denote the measure being referred to. How do we find a measure which satisfies the fourth condition, invariance under translations and rotations?

Without telling you the proof, which is a charming fact, but which I am obliged to omit for reasons of time, the formula  $k dpd\theta$  with the free non-negative constant  $k$  describes precisely the only rotation and translation invariant measures that can be put on the family of straight lines. It is easy to see—I do not think I will labor the point—that  $dudv$ , for instance, is not; if you translate a set of lines so that it gets very far away from the origin you will diminish its measure. As a matter of fact,

$dudvJ = (1/p^3)(dpd\theta)$ , so that if  $dpd\theta$  really is invariant with respect to translation, since  $p$  is not,  $dudv$  cannot be. Now this way of defining a measure, is, in fact, independent of the choice of the origin in the plane and the choice of the axis through it that permitted us to define  $p$  and  $\theta$ , which it must be if it is going to have geometrical meaning (Fig. 3). One can base the proof of the invariance property of  $dpd\theta$  on this fact alone.

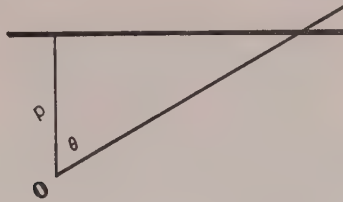


FIG. 3

Now, let us retreat to a retina; say, a rectangle. Suppose we consider only those straight lines which are going to penetrate a rectangle (Fig. 4). That family of straight lines is a submanifold of the manifold of all straight lines in the plane, and it is only that submanifold which we use for pattern recognition when we know

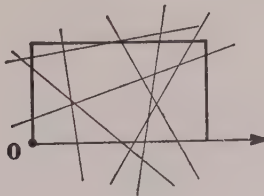


FIG. 4

the character will be displayed within a retina. Figure 5 shows the points  $(p, \theta)$  corresponding to this submanifold. Since we are integrating  $dpd\theta$  over this region it is the area of this region which corresponds to the measure of all the straight lines penetrating the rectangle. How do we compute this area? In the case of a retina as simple as a rectangle, it can be done by brute force. Integral geometry, which I did not invent and am not claiming as my own,

has another way of computing this area, by answering, in fact, a more general question. We will not only compute the measure of all the straight lines which cross the rectangle, we will compute the measure of all the straight lines which cross any convex curve, and the rectangle is only one of that family of curves. We will, in fact, answer that question by answering a slightly modified



FIG. 5

question. We will compute a *weighted* measure of all straight lines that cross any rectifiable curve whatsoever, closed or open, convex or not.

I will just skip to the result. The result is this: Let  $n(p, \theta)$  denote the number of times (possibly infinite) the line  $(p, \theta)$  crosses the curve  $C$ . I will compute

$$\iint n_C(p, \theta) dp d\theta.$$

Now for convex curves, one can readily show that a line either fails to cross a convex curve at all, crosses it twice, or touches it in some tangential way. The "tangential" lines can be shown to be of measure zero and can be neglected in this integration. Apart from these tangential exceptions, when a line crosses a convex curve at all, it crosses it twice, so that if I apply this formula for  $C$ , a convex curve

$$\iint n_C(p, \theta) dp d\theta = 2 \iint A_C dp d\theta$$

where  $A_C$  is the set of all lines intersecting  $C$ . I will have computed the formula that I need for my retina.

I will now state the result, which is of enormous generality. The result is this, that

$$\iint n_C(p, \theta) dp d\theta$$

is equal to twice the length of  $C$ . I will just make the following remark. If  $C$  is a crinkled curve, or a stretched-out version, you get the same result, which might seem at first alarming, because more straight lines cross an "unwound" curve than cross a "tightly wound" version. But when the tightly wound version is crossed, it is crossed more often. That is why the formula looks like this. The method of proof is certainly a key to the whole thing and is a beautiful thing, but I have my choice as to what I will lecture on and I will not give the method of proof, unless someone by chance should ask me afterwards, in which case I will be delighted.

Now we know the following. For closed convex curves the measure of straight lines which intersect the closed convex curves, that is

$$\iint A_C dp d\theta$$

is exactly the length of  $C$ . Now I can construct a bona fide probability measure on the manifold of all straight lines which cross a convex retina  $R$ . It is the following: My bona fide probability measure is  $dp d\theta / l(R)$ . It is of the form  $k dp d\theta$  with  $k$  chosen to be  $1/l(R)$ . Now if I take  $A$  to be all lines in the whole retina with respect to this measure  $m(A) = 1$ .

Now I have got a bona fide probability situation. It is easy to design, at least conceptually, a device which can compute the length of a figure presented in any location or orientation, on the retina.

Suppose, for example, I have an alphabet of patterns which differ in their length. By the way, they do not have to be curves in the ordinary sense. For example, the letter  $y$  is not a curve in the ordinary sense (it is not the continuous image of a piece of straight line), or a triple of circles is not a "curve" in the ordinary sense. If my alphabet of patterns (which can allow this degree of generality) has the property that the members differ substantially in



their length alone, then I can use that single geometric parameter and design a device, a sort of sophisticated Monte Carlo method, for finding which pattern I am looking at. I do not believe I need to elaborate how you would design a device in terms of this. You can in practice throw only a finite number of lines. Their average number of crossings is, with high probability, very near the average of all retinal lines, which by the above result is twice the length of the pattern being examined. The key formula in the working of such a device is that the expected value of the random variable  $n_C(p, \theta)$  defined as the number of crossings of the random line  $(p, \theta)$  with  $C$  is twice the length of  $C$  over the length of the retina.

Now by tossing fancier devices you get fancier formulas. For example, suppose I do not consider the random variable which counts the number of crossings. Suppose I am dealing only with patterns that are smooth so that they have a notion of curvature, and suppose I count every time a line crosses, not the number of points of intersection, but the sums of the curvatures at these points. (Incidentally, when I once mentioned this idea, for more or less pedagogical reasons only, and pooh-pooed it as being probably not physically realizable, I was told that measuring curvatures with electrostatic means is a rather common thing, so that possibly a device can even be made doing this.) If we consider the sum of the curvatures as our random variable, it can be proved that its expected value is twice the total curvature of  $C$ . In symbols,

$$\iint K_C(p, \theta) dp d\theta = 2(C)$$

where  $K_C(p, \theta)$  is the sum of the curvatures at the intersection points of the random line  $(p, \theta)$  and the curve  $C$  and  $K(C)$  is the total curvature of  $C$ , that is the total change in angle of the tangent line to  $C$  as  $C$  is completely traversed. These two numbers are very different, for two curves  $C_1$  and  $C_2$ , although their lengths may be similar, so that it might be actually more feasible to perform the discrimination with the curvature trick than with the length trick.

There are many other remarks crying to be made that lack of time prevents. For example, what is the expected value of the

random variable,  $m_C(p, \theta)$ , which is 1 when the random straight line  $(p, \theta)$  crosses the figure  $C$  and zero when it does not? That is just

$$\iint m_C(p, \theta) dpd\theta = \iint A_C dpd\theta$$

you see, rather than

$$\iint n_C(p, \theta) dpd\theta.$$

I have already remarked that when you compute this a closed convex curve, you report back the length of the curve.

What happens when you apply it to a curve  $C$  not necessarily either closed or convex, to a squiggle? What you get is the length of the convex hull  $D$  of  $C$  ( $D$  is the smallest convex figure which can be drawn containing  $C$ ). Therefore, if two figures have sufficiently different convex hulls  $D_1$  and  $D_2$ , I may use the random variables  $m_{C_1}(p, \theta)$  and  $m_{C_2}(p, \theta)$  as the ones on which to base a device. The letter "A" has a convex hull which is independent of the precise location and angle of the cross bar. In a sense I have a machine which knows what a letter "A" looks like without it having to be printed terribly carefully.

Now I should remark about one obvious problem that you would raise if you were confronted with the problem of designing a device and would like to know *how many times* you must toss your random line to get with high probability a very accurate average. This requires knowing the second moments of the random variables and not just the expected value. The second moments have never been calculated. Integral geometry does not generally seem to shed much light on them. I can calculate the second moments for  $n_{C_1}(p, \theta)$  when  $C_1$  is a segment, and I can calculate it for  $n_{C_2}(p, \theta)$  when  $C_2$  is a semi-circle. Both of these are elementary. You can assign them as undergraduate calculus problems. But I cannot so easily compute the second moment for the letter "D", composed of a semi-circle and a segment, because the random variables  $n_{C_1}$  and  $n_{C_2}$  are not uncorrelated, so that their second moments do not add. There are a lot of interesting and, I think, amenable questions attached to building devices on this principle.

I would also like to point out that simulation provides answers

that may be perfectly useful to the design of devices. There is no reason at all why you are obliged to get an analytical estimate of the number of tosses required, if you know the patterns you are going to be looking at beforehand. You may resort to simulation.

Let me now speak briefly about various extensions of these ideas. You can put a measure on the manifold, not of lines, but of *pairs* of points. That is a perfectly good manifold. It is a nice four-dimensional thing. Obviously an invariant measure is  $dx_1dy_2dx_1dy_2$ ; however, you can also introduce other coordinates—you do not have to use the four cartesian coordinates. In fact, in proving some theorems you would never use the four cartesian coordinates; you would indeed use some other coordinates. Incidentally, the cleverest proof of the fundamental theorem above involves not using  $p$  and  $\theta$  to compute the expected value of  $nc(p, \theta)$ , but making a change of variable to a coordinate system adroit for purposes of the proof. In an alternative coordinate system for the family of pairs of points you use the straight line connecting the two points as an item in the coordinate system, so that you can use the results that I have already described concerning the measure of straight lines.

In terms of the manifold consisting of pairs of points, a natural variable to consider is  $r(P_1, P_2)$ , the distance between  $P_1$  and  $P_2$ . If we consider the submanifold of pairs of points  $P_1$  and  $P_2$  lying in a given convex domain  $D$ , we can, as above, introduce a unique invariant probability measure on the submanifold, and we are justified in calling  $r(P_1, P_2)$  a random variable. It is closely related to the random variable  $\sigma(p, \theta)$  where  $\sigma$  is the length of the chord on  $D$  formed by the line  $(p, \theta)$  passing through  $P_1$  and  $P_2$ . These two are related by the identity

$$\iint \sigma^n(p, \theta) dpd\theta = n(n-1)\left(\frac{1}{2}\right) \iint r^{n-3}(P_1, P_2) dP_1dP_2$$

$n = 2, 3, \dots$ . It is easy to verify that

$$\iint \sigma(p, \theta) dpd\theta = \pi \cdot (\text{Area of } D)$$

and we saw earlier that

$$\iint dpd\theta = 2 \cdot (\text{perimeter of } D).$$

These results cover the cases  $n = 1, 0$  respectively.

For  $n = 2$  we obtain

$$\iint \sigma^2(p, \theta) dp d\theta = \iint r^{-1} dP_1 dP_2$$

equals self-potential of a uniform lamina on  $P$ , and for  $n = 3$

$$\iint \sigma^3(p, \theta) dp d\theta = 3 \iint dP_1 dP_2 = 3 \cdot (\text{Area of } D)^2.$$

From these we can deduce the first three moments of the random variable  $\sigma(p, \theta)$  in terms of geometrical (and physical) parameters of  $D$ .

The next major class of formulas are those formulas which deal with manifolds somewhat more general than manifolds of straight lines or manifolds of pairs of points, including them both as degenerate cases. The manifold consists of all possible locations of a specified curve. Take a specified curve  $C$  and consider all of its various motions in the plane, its various positions that can be obtained by translation and rotation. Every such location is regarded as a point in this new manifold. I want to coordinatize that manifold. This can be done, for instance, by imagining a half-line attached rigidly to one point  $P$  of the curve, and using as coordinates the cartesian coordinates  $(x, y)$  of  $P$  and the angle of elevation  $\theta$  of the half-line from the  $x$ -axis. I want next to deduce a measure to put on that manifold conceptually no different from the measure put on the manifold of straight lines. I want this measure to have the property that it is invariant under rotation and translation.

It is not difficult to see that  $dx dy d\theta$  is such a measure, and that the most general such measure is a non-negative multiple of this. Of course other triples of coordinates can be introduced in terms of which to write the measure. When we specify the multiplicative constant to be 1 we refer to the resulting invariant measure on the manifold of locations of a given curve as the *kinematic measure* of the mobile curve. This measure can be related to the previously introduced invariant measure on lines (more exactly, on half-lines, or oriented lines) by introducing a random oriented line as one of the items in a different coordinatization of our manifold. This permits the use of earlier

results and that, in fact, is the order of events in the derivation of the theorems I will state next.

First let me call your attention to the special case when the mobile figure is an oriented line segment (the "needle" of Buffon, in its full mathematical dress) of length  $\lambda$ . It can be shown that the kinematic measure of the set of all such needles (that is, all locations and orientations of a mobile needle) intersecting a convex curve of area  $A$  and perimeter  $L$  is  $2\pi A + L\lambda$ . The degenerate case in which the convex curve is a doubly traversed segment of length  $l$  ( $A = 0$  and  $L = 2l$ ) shows that the kinematic measure of  $\lambda$ -needles meeting a segment of length  $l$  is  $4Ll$ .

This can also be interpreted as being the integral of  $n_C(x, y, \phi)$  the number of intersections of the  $\lambda$ -needle with the segment of length  $l$  (the integration extending over all  $(x, y, \phi)$ , i.e. all locations of the needle); for, apart from a set of measure zero,  $n_C$  is either 1 or 0, and so its integral coincides with the measure of the set where it is not zero, that is, locations in which the needle meets the segment.

Reinterpreted this way, the result generalizes as follows: if  $C_1$  is a mobile rectifiable curve of length  $L_1$  and  $C_0$  a fixed rectifiable curve of length  $L_1$ , and  $n_{C_0}(x, y, \phi)$  is the intersection number when  $C_1$  is in the position described by  $(x, y, \phi)$ , then

$$\iint n_{C_0}(x, y, \phi) dx dy d\phi.$$

This result is due to Poincaré, and so precedes the emergence of a general theory of Integral Geometry by about thirty years. It also furnishes a generalization, in a very natural way, of the notion of length when the fixed set is not a rectifiable curve in the ordinary sense, but the integral still exists, namely  $1/4L_1$  times the integral.

In a sense, a retinal version of the  $\lambda$ -needle case of this theorem can be provided which, as before, lets us renormalize our invariant measure, turning it into a true probability measure. Namely, if we consider all the needles which enter a retina of area  $A_R$  and perimeter  $L_R$ . They are not restricted however to lie entirely within the retina, merely to enter it. Then their total kinematic measure is  $2\pi A_R + \lambda L_R$ . Now, by choosing the reciprocal of this as the multiplicative constant in the formula  $k dx dy d\phi$ , instead of one, we get a measure which is just as invariant as the kinematic



measure, but which attaches the number one to the set of all needles which are retinal in this somewhat artificial sense. (It would certainly be preferable to regard only those needles lying *entirely* within the retina as the set which is to have the measure one, but the kinematic measure of this set so far escapes me, although it can be bruted out for special choices of the retina.) When this probability measure on needles is introduced, we can introduce random variables associated with the number and length of intersections of a random needle on a pattern exhibited within the retina. I will not carry out this program here, nor will I go into corresponding detail when other figures than segments are used as the mobile figures, but rely on your goodwill and sense of analogy to state the concluding results.

First among these, there is an explicit formula for the kinematic measure of all  $\lambda$ -needles which intersect both limbs of an angle (the limbs are regarded as infinitely long or equivalently, long enough so that further prolongation will not enlarge the set of needles meeting both). If the angle is  $a$ , the kinematic measure is

$$\frac{\lambda^2}{2}(1 + (\pi - a) \cot a).$$

Thus you have an angle-measuring device which is independent of the location and orientation of the angle being measured. In all honesty, it must be confessed that this is not a retinal theorem, and there seem to be substantial problems associated with the cases where the vertex of the angle is near the edge of the retina. But, even if the language and some of the tools of probability theory cannot be introduced, I believe a designer can take advantage of the ability to make finite samples from the family of all possible locations and safely estimate the angle being displayed.

Second, suppose we take for the mobile figure a bounded region of the plane bounded by a finite number of simple closed curves. Consider a fixed region of the same degree of generality; we study the arcs of overlap of the fixed figure with the mobile one. This was the case which struck me as relevant in my initial concern over the alleged description of the Perceptron. The integral of this over all positions of the mobile figure (I would call it the expected value of this random variable if I had a probability measure to work) turns out to be  $2\pi$  times the product of the

areas of the fixed and mobile figures. Corollary: nobody can successfully tell a circle from a square of the same area by examining the "average" amount of overlap of either with a mobile area of any shape whatsoever—the result only depends on areas, not shape.

Third, if we compute the perimeter of overlap rather than area, and again "average" this, that is integrate this with respect to the kinematic measure for all locations, the result is  $2\pi(A_0L_1 + A_1L_0)$  where the subscript 0 refers to the fixed figure and the subscript 1 refers to the mobile one (the result is symmetric anyway).

If we compute the total *curvature* of overlap for each position of the mobile figure, and again integrate this over all positions using kinematic measure, we get  $2\pi(A_0K_1 + A_1K_0 + L_0L_1)$ . In particular, if both fixed and mobile figures are convex domains, their total curvatures  $K_0$  and  $K_1$  will be  $2\pi$ , as will the total curvature of the overlap in every position in which there is overlap. Since the integrand is then constant, we can factor  $2\pi$  out of the entire identity (as we did with the constant 2 when considering intersections of lines with a convex curve) and we obtain the kinematic measure of all locations of a convex mobile figure which overlap with a convex fixed figure to be  $2\pi(A_0 + A_1) + L_0L_1$ . This again permits a sort of retinal version of locations of an arbitrary convex figure "within" a convex retina where unfortunately the word "within" misleadingly means "intersecting" or equivalently, "overlapping". As before, knowing the kinematic measure of all the retinal locations, we may divide the kinematic measure by this constant, producing a true probability measure on the submanifold of retinal locations.

As the last of this little bouquet of theorems, I want to mention a slightly different direction of inquiry. Let us consider a doubly infinite array of congruent figures or cells filling the plane without overlap, as with parallelograms or hexagons; in each consider a congruently placed replica of a curve, of length  $L_C$ . Now consider a mobile curve, of length  $L_1$ , not necessarily small enough to be contained in any one cell. The kinematic measure of this mobile figure may be taken to be  $dx dy d\phi$  where  $(x, y)$  are the coordinates of a point fixed in the mobile figure. Let us call this point the *base point* of the mobile figure. If we integrate the number of intersections on the infinitely repeated fixed figure with the

mobile figure, over the submanifold of all positions of the mobile figure having the base point within any one cell, the result is  $4L_0L_1$  where  $L_0$  is the length of the fixed figure and  $L_1$  is the length of the mobile one. This is again a theorem with a retinal or probability version, since the kinematic measure of all locations and orientations of the mobile figure within any one cell of area  $A$  is

$$\int_0^{2\pi} \iint_A dx dy d\phi = 2\pi A.$$

If we divide the integral (or equivalently, the kinematic measure with respect to which it was performed) of the number of intersections by this normalizing constant we get the average number of intersections, for a random position of the mobile figure with base point in the cell  $A$ . This then is ratio of the last two, namely  $2L_0L_1/\pi A$ . The special case in which the cells are rectangles of sides  $a$  and  $b$ , and the fixed curve is a pair of adjacent sides, while the mobile curve is of length  $l$ , above that the average number of intersections is  $2(a+b)l/\pi ab$ . If we let  $a$  tend to infinity we get  $(2/\pi)(l/b)$  for the average number of crossings of the mobile curve with an infinite grid of parallel lines of distance  $b$  apart. This is the general form of the Buffon needle problem referred to at the beginning. As you see, it is not the simplest theorem in the subject, viewed in this framework.

My final remark is this. There has been some progress made as to the question of the number of tosses that are required in this kind of random process to get specified accuracy. Back in 1841 Cauchy proved a theorem in what would now be called Integral Geometry: if you project a closed convex curve on the direction making an angle  $\theta$  with the  $x$ -axis, and average the resulting projections, regarding  $\theta$  as chosen at random between 0 and  $\pi$ , the result is  $2/\pi$  times the length of the curve.

Now this remark has been used by the mathematician Steinhaus in 1930 to design a machine very similar in idea to the ideas that I have now. (I did not know of Steinhaus' work when I first suggested that Integral Geometry would be a good trick for pattern recognition.) Steinhaus was interested in designing a so-called longimeter, a means for measuring lengths in the field

of a microscope by using this trick of projections. He knew that Cauchy's result called for the use of the average over all angles between 0 and  $\pi$ . He asked what was the error caused by using a numerical integration, replacing the integral by a sum involving six equally spaced terms. Now that is an extremely small sample, only six, and it is not a sample of six independent observations from among all angles. The six are dependent: having chosen the first, all the others are found by advancing thirty degrees at a time. He was able to show (and it is really a trivial fact having to do with the accuracy of numerical integration for the cosine function, which the general case reduces to immediately) that by using such a sample you are always within 2.26 per cent of underestimate and 1.15 per cent of overestimate of length. So there is at least some evidence that some finite samples are very good approximations to these averages.

**DAVID G. WILLIS**

*Lockheed Aircraft Corporation, Missiles and Space Division,  
Sunnyvale, California*

## THE FUNCTIONAL DOMAIN OF COMPLEX SYSTEMS

### INTRODUCTION

In this paper three things are presented. The first is a brief review of some of the fundamental results which have previously been obtained dealing with systems which might be said to be self-organizing. Secondly, there are presented a number of arguments from the viewpoint of switching theory, demonstrating that the domain in which such systems operate cannot be all functions of  $n$  variables when  $n$  is large. Finally, a possible domain of operation of such systems is described.

There exist complex, highly organized systems, the detailed functioning of which we understand only very poorly, if at all, and which have been termed self-organizing. Two interesting examples are the human brain and the system which produced it, the evolutionary system. Theoretical results over the past thirty years have led to the speculation that it might be possible to construct machines which display some of the flexibility and power of these complex systems. Turing,<sup>(1)</sup> for example, investigated and developed the general notion of automata. He was able to demonstrate the existence of universal machines which could do anything with information that any other machine could do. Further, McCulloch and Pitts<sup>(2)</sup> have demonstrated that it is possible to build machines which will do anything that can be precisely and completely described using only simple elements which closely resemble neurons, the building blocks of biological information-handling systems.

Von Neumann<sup>(3)</sup> has demonstrated that the results of Turing in



principle allow us to construct machines which are able to reproduce not only themselves but machines that are more complex than themselves, thereby proving the theoretical possibility of evolution.

Finally, von Neumann,<sup>(4)</sup> and later, Moore and Shannon,<sup>(5)</sup> have demonstrated that it is possible to build such machines having arbitrarily high reliability from unreliable components.

None of these results, however, bears directly on the problem of how efficiently highly complex machines can be constructed. That is, how rapidly they will operate and what their requirements are in terms of physical components. In the following discussion one aspect of this problem will be considered. Specifically: what are the physical requirements for a machine which operates in the domain of all functions of  $n$  binary variables?

#### PHYSICAL REQUIREMENTS

In Fig. 1 is shown a schematic diagram of a simple model which is perhaps sufficiently general to include any of the systems which might be described as self-organizing. The model has three parts: an environment in which the system lives; a mechanism which

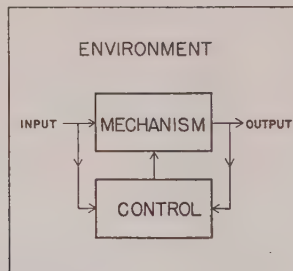


FIG. 1. Schematic diagram of a complex system.

receives inputs from the environment and also affects the environment by means of outputs; and a control which measures some or all of the inputs and outputs of the mechanism and has means of altering the mechanism.

The mechanism may produce any of a large number of functions. The control conducts a searching operation in which the search

for a desired function is guided by information derived from functions already tried.

If we view this as a model from the evolutionary system, the mechanism represents some species which is characterized by its genetic pattern. The control is simply a part of the species' environment which induces genetic changes and eliminates unsuccessful genetic patterns by natural selection. Alternatively, if we consider this as a model of the brain, both the control and the mechanism are part of the brain. New functions are learned by means of information obtained from the trial of other functions.

The question which we shall examine is that of the domain of functions which can be produced by the mechanism. That is, what is the domain of functions which are available to the control portion of the system in its alteration of the mechanism? We shall examine this question from the standpoint of switching theory in which all variables are two-valued. Later we shall indicate how our results may be generalized to cover situations in which inputs and outputs are not binary variables.

We shall examine the simple situation in which the inputs are represented by  $n$  binary variables and the outputs by a single binary variable. Let us assume first that the domain of functions which the mechanism can produce is that of all possible functions of  $n$  binary variables. It is well known that there are  $2^{(2^n)}$  of these functions. If this is the domain of operation of the mechanism, we may then examine a number of questions relating to the amount of hardware required in the mechanism and the time required for the system to organize or find the proper function.

### *Memory Requirements*

In general, if we require any mechanism whatsoever to be able to assume any one out of  $r$  states, we require a storage capacity of at least  $\log_2 r$  binary digits. That is to say we require at least this many binary storage elements, flip-flops, magnetic cores, or their equivalent. Any less than this would be insufficient to define uniquely the state which the mechanism assumes. Therefore this number of binary digits represents the lower limit for the memory requirement for the mechanism.

Clearly if the domain of operation of the mechanism in Fig. 1

is all  $2^{(2^n)}$  functions of  $n$  binary variables, its storage requirement will be  $2^n$  binary digits.

In Fig. 2 is shown a chart of this storage requirement as a function of  $n$ . Also shown, for purposes of comparison on the same scale, are a number of physical quantities of very large size. It seems fairly conclusive that we will never be able to build any

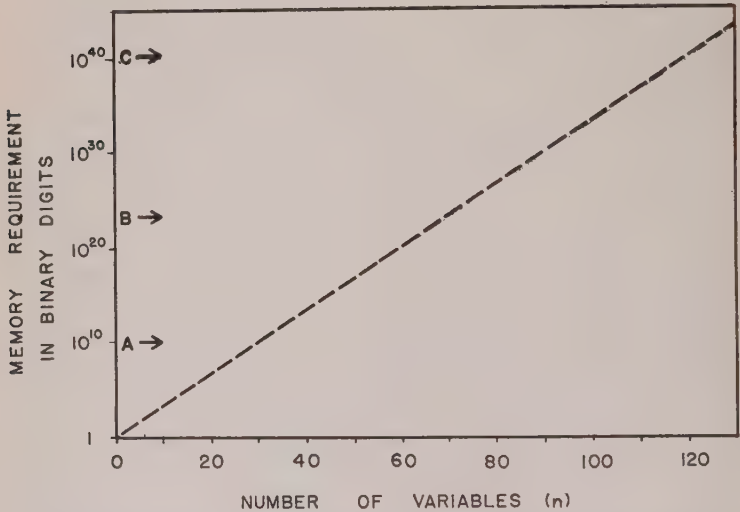


FIG. 2. Memory requirements as a function of  $n$ . A represents the approximate number of neurons in the human brain. B is Avogadro's number (the number of molecules in a gram molecular weight). C is the approximate number of electrons within the earth.

machine of this kind which operates on more than a few tens of variables, and that no natural system could operate on as many as a hundred variables.

### *Organizing Time*

We may now examine the time required for the control portion of the system to place the mechanism in some particular organized state represented by one of the  $2^{(2^n)}$  functions of  $n$  binary variables. We can, of course, postulate a serial searching procedure which, if we were fortunate, might provide an answer in a very short

period of time, or conversely, might provide one only after a very long time. However, the average times involved in a serial search are unduly long and we may therefore postulate a much more efficient searching technique.

The most efficient method of searching arises when one is able to utilize all the possible information in the most efficient way. In this case, each bit of information—the answer to a “yes” or “no” question—suffices to cut the remaining choices in half. No more efficient utilization of the information can be made. If there are  $2^{(2^n)}$  initial choices it will require  $2^n$  bits of information to select the proper function.

If we now pick a rate of information handling for the control portion of our system which is sufficiently high to encompass any known natural system and also to encompass any system we might build, we may obtain another estimate of the limitations on  $n$  for this kind of a system.

In Fig. 3 are shown the number of years required to find one function of  $n$  variables as a function of  $n$  for information rates of  $10^{10}$ ,  $10^{20}$ , and  $10^{30}$  bits per second. The largest of these rates is more than 20 orders of magnitude faster than the fastest known digital circuitry and is certainly sufficient to encompass the processing rate of any biological system.

Two time intervals of large magnitude are shown for purposes of comparison. Here we are again severely limited in the number of variables such a system could handle.

#### *Physical Component Requirements*

Under normal conditions of pressure and temperature there is room for less than  $10^{25}$  atoms in a cubic centimeter of solid material. Each atom could be one of approximately 100 elements, each of which could certainly have no more than 100 possible valence states. Thus the atoms could be selected in far less than

$$\binom{10^{25}}{10^4}$$

possible ways.

Let us make the extreme assumption that each atom might occupy any one of less than  $10^{100}$  significantly different physical

positions within the cubic centimeter, and that it might be in any one of  $10^{10}$  different energy states, orientations, etc.

Combining these data, we find that there must be less than

$$\left( \frac{10^{25}}{10^4} \right) \cdot [10^{100}]^{10^{25}} \cdot [10^{10}]^{10^{25}} \ll [10^{114}]^{10^{25}} < 2^{(2^{95})}$$

different possible physical configurations of matter within a cubic centimeter under normal conditions of temperature and pressure.

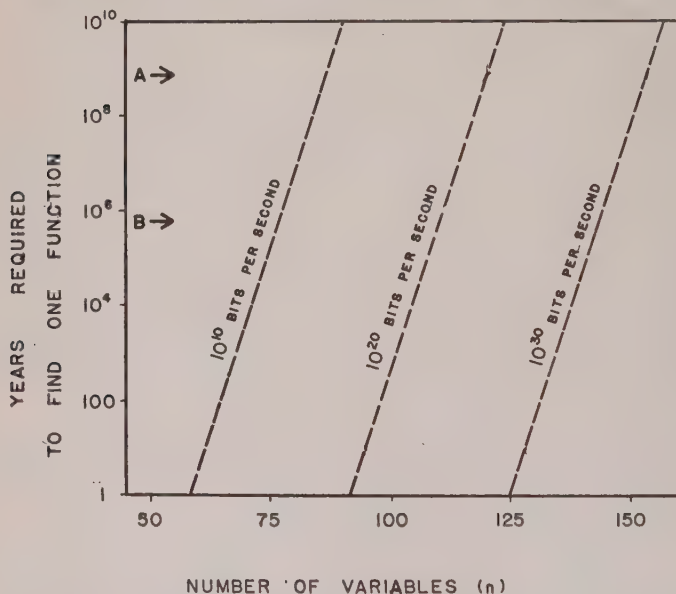


FIG. 3. Time required for finding one of the functions of  $n$  variables with information rates of  $10^{10}$ ,  $10^{20}$ , and  $10^{30}$  bits per second. A is the approximate age of the oldest fossil records of life on earth. B is the approximate time since man first appeared on earth.

Obviously for  $n > 95$  all but a vanishingly small proportion of functions of  $n$  variables cannot be mechanized within a cubic centimeter of space. There are simply not enough possible physical configurations of the matter to go around.

The analysis may easily be extended to show that even by allowing mechanization within the volume of the earth we only increase the limit on  $n$  to something less than 155.



Since liberal allowances have been made throughout the foregoing discussion, the actual limits on  $n$  are certainly much smaller.

We now examine the possibility of relieving the above difficulties by extending the domain of the analysis beyond that of switching theory to that of functions of multi-valued variables. If we leave the number of variables the same, we can only increase the number of functions by allowing the variables to assume more than two values, and thereby increase the memory, time and component requirements even more. On the other hand, if we reduce the number of variables at the same time as we allow the variables to assume more than two values in such a way that the total number of configurations of all the variables is approximately the same, we still have not reduced the number of functions and therefore we have not relieved the situation.

Since a continuous variable is simply the limiting case of a discrete variable which can assume an increasingly large number of values, we see that we can gain no advantage by introducing continuous variables. We also note that the assumed information rate for the control portion of the model was sufficiently high that it includes any possible increase that might occur by the use of multi-valued or continuous variables.

One other possibility is whether or not random or noisy processes in our system can help it organize any faster or more efficiently. The answer must be negative since we have postulated that we are using the information in the most efficient possible manner. Any errors in it will only lengthen our searching process. We are, therefore, forced to the conclusion that no matter how we wish to view the problem, the domain of operation of these systems, or any system for that matter, cannot be all functions of  $n$  variables where  $n$  is large.

We should note that  $n$ , in fact, is large, not only for the complex systems which we find in Nature, such as the evolutionary system and the human brain, where the number of variables being handled at any time may be on the order of millions; but  $n$  is also very large for many of the man-made systems, for example, large scale computers or telephone switching systems. For instance, the multiplier circuit for a digital computer characteristically handles between 60 and 100 binary digits as variables.

The difficulty we encounter arises not with the number of

variables that are involved but simply the fact that we have been trying to admit the possibility that any function of these variables might be required. One must therefore ask the question: is there any way that we may characterize the domain of those functions of  $n$  variables in which these systems operate, or possibly could operate? In the next section we shall attempt to outline a possible answer to this question.

### DECOMPOSABLE FUNCTIONS

We now inquire into the nature of the subset of all functions of  $n$  variables which are the domain of operation of complex systems. It is clear that this subset must represent a fraction of all  $2^{(2^n)}$  functions which is vanishingly small as  $n$  increases. Otherwise we would be faced with the same situation that we have when we admit all functions.

It is possible to think of a number of classes of possible switching functions of  $n$  variables which are vanishingly small as  $n$  increases. For example, we can introduce the symmetric functions, the threshold functions (which are modeled after neurons) and functions where  $m$ , the minimum number of terms in the equivalent Boolean expression, is small. Each of these classes of functions can be mechanized relatively inexpensively and is sufficiently small that it can be searched efficiently and specified with a small amount of information. However, such classes of functions still do not appear to be versatile enough to constitute in themselves the subset of all functions which are of interest to us. We could reasonably expect however that the subset we are seeking should include all of the above classes.

Aiken<sup>(6)</sup> has suggested the possibility that the only switching functions of  $n$  variables which are of interest to us when  $n$  is large, are those functions which are decomposable. The notion of decomposability of a switching function was first introduced by Shannon<sup>(7)</sup> who termed this property "separability".

Consider a switching function of  $n$  variables,  $f(x_1, x_2, \dots, x_n)$ . We may say that the function is disjunctively decomposable if there exists another function  $g(x_1, x_2, \dots, x_k)$  such that

$$f(x_1, x_2, \dots, x_k) = h(g, x_{k+1}, x_{k+2}, \dots, x_n)$$

and

$$1 < k < n.$$

Decompositions of this kind have been studied by Ashenhurst.<sup>(8)</sup> They are characterized by the fact that the variables,  $x_1$ , may be partitioned into two disjunctive or non-overlapping sets such that the members of one set are variables of the function  $g$ . Shannon<sup>(7)</sup> has shown that the proportion of all functions which are disjunctively decomposable is vanishingly small for large  $n$ .

We note that instead of a function of  $n$  variables, we now have a function of  $k$  variables and another function of  $n-k+1$  variables. The physical requirements for this class of functions are accordingly enormously less than those for all functions of  $n$  variables for large  $n$ . However,  $k$ , or  $n-k+1$  will in general still be so large that functions containing only a single simple disjunctive decomposition still represent a domain which is far too great for the operation of the complex systems which we are considering.

We may conceive of a class of functions which contain multiple decompositions of the disjunctive type. For example, we might be able to express the function  $f(x_1, x_2, \dots, x_n)$  as a function of  $k$  variables,  $\phi_i$  ( $i = 1, 2, 3, \dots, k$ ) each of which is a function of a disjunctive set of  $k$  or less of the  $n$  variables. Again, as  $n$  increases, this class of functions must be a vanishingly small subset of all possible functions of  $n$  variables.

There also exist other decompositions which are not disjunctive. For example we might be able to express  $f$  as a function of  $k$  variables each of which is a disjunctively decomposable function of the original  $n$  variables.

It appears reasonable to generalize the above notions in the following way. Let us consider the class of functions of  $n$  or less variables which may be expressed by no more than  $p$  functions of no more than  $k$  variables each. We can thus imagine situations such as those shown in Fig. 4. Obviously if we allow either  $p$  or  $k$  to be large enough, any function of  $n$  variables may be expressed by a decomposition of this kind. We must therefore inquire as to the maximum values of  $p$  and  $k$  which will allow reasonable physical requirements. We cannot do this in any very precise way since so much depends on the magnitude of  $n$  and on the complexity of the system being considered.

However, it is possible to determine an upper bound on the number of such decomposable functions in terms of  $n$ ,  $p$ , and  $k$ ,

and also to set an upper limit on the number of pairs of relay contacts which would be required to mechanize them.

First we consider an upper bound on the number of functions of  $n$  or less variables which may be expressed in terms of  $p$  functions of  $k$  variables. The  $k$  variables of any of the  $p$  functions may be selected from among no more than the original  $n$  variables and the  $p$  functions. Thus, there are less than

$$(p+n)^{(pk+1)}(2^{2^k})^p = (p+n)^{(pk+1)}(2)^{p(2^k)}$$

functions of  $n$  variables which may be expressed by  $p$  functions of  $k$  variables. This actually represents an extreme upper bound since

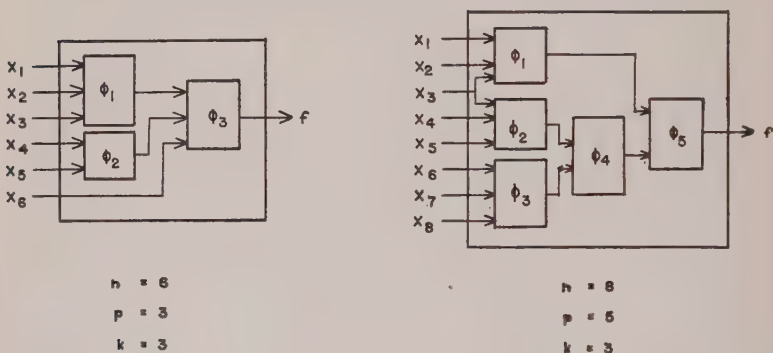


FIG. 4. Typical decompositions of switching functions into functions with fewer variables.

we have counted many duplications and have admitted circular functional relationships which are meaningless in the present context.

We recall that both memory and time requirements for our hypothetical system were proportional to the logarithm to the base 2 of the number of functions. It follows that for the class of decomposable functions under consideration an upper bound for these physical requirements is proportional to

$$(pk+1) \log_2(p+n) + p(2^k).$$

It is evident that this limit varies approximately as  $p$  and as  $2^k$ .

Shannon has shown that any function of  $k$  variables may be mechanized in a two-terminal network with less than

$$\frac{2^{(k+3)}}{k}$$

pairs of relay contacts. If we allow multilevel circuits, we require no more than  $p$  such networks in the mechanization. Thus the cost of this class of functions is always less than  $(p/k)(2^{(k+3)})$  pairs of relay contacts. Hence, the upper limit for relay contact cost varies as  $p$  and  $2^k/k$ .

It follows that for decomposable functions of a large number of variables, we may be assured of reasonable physical requirements only when  $k$  is very much smaller than  $n$  and  $p$  is very much smaller than  $2^n$ . We may therefore define the class of functions of  $n$  variables which may be expressed in terms of  $p$  functions of  $k$  variables where

$$\begin{aligned} k &\ll n, \\ p &\ll 2^n. \end{aligned}$$

For large  $n$ , the functions in this class represent a vanishingly small proportion of all functions and will have physical requirements which are considerably more realistic. They therefore represent a possible domain of operation of complex systems. Actually all that we are saying is that the only functions of a large number of variables which are of interest to us are those which may be built up from a reasonable number of functions of a small number of variables.

A number of further remarks should be made concerning this class of decomposable functions. It is easily demonstrated that the symmetric functions of  $n$  variables may be expressed by  $p$  functions of  $k$  variables where

$$\begin{aligned} p &\leq n^2, \\ k &\leq 3. \end{aligned}$$

A symmetric function of  $n$  variables may be completely characterized by a set of  $r$  different integers,  $a_1, a_2, \dots, a_r$ , where  $r$  is less than  $n$ , and such that the function has the value 1 whenever  $a_i$  ( $i = 1, 2, \dots, r$ ) of the variables have the value 1. Such a



function may always be expressed as shown in Fig. 5. Each variable feeds into one box. The outputs of the boxes carry information as to the number of variables to the left having values of 1. For example, the second output of the  $i$ th box would have the value 1 if exactly two of the first  $i$  variables have the value 1.

Each output need be a function of at most 3 variables. The outputs of the  $n$ th box may be combined as shown by two-variable functions to produce any symmetric function. It follows directly that  $p$ , the number of elementary functions, is less than  $n^2$  and none of them has more than three variables.

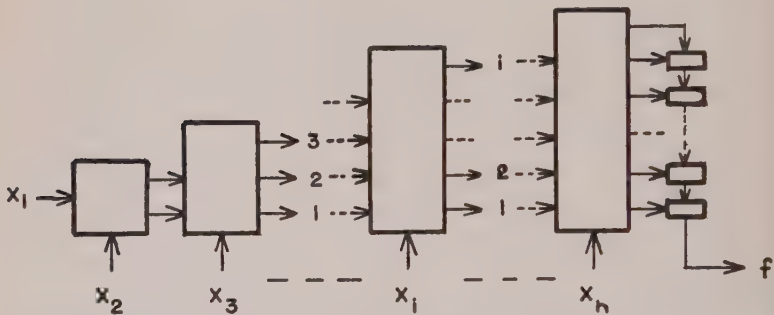


FIG. 5. Generalized decomposition of symmetric switching functions.

It is also easily demonstrated that any function of  $n$  variables which has  $m$  polynomial terms in an equivalent Boolean expression may be expressed by  $p$  functions of  $k$  variables where

$$p \leq mn - 1,$$

$$k \leq 2,$$

Consider first  $m = 1$ . Such a function may be expressed as shown in Fig. 6. Here, each box represents a two-variable function. The output of each box simply carries the information as to whether or not all variables to the left have the value given by the single polynomial term. No more than  $n - 1$  functions are needed. For  $m > 1$  there will be  $n - 1$  functions for each of the  $m$  terms. As with the symmetric functions it will always be possible to combine

the outputs of the  $m$  sets of functions by  $m-1$  two-variable functions. Hence,

$$p \leq mn-1,$$

$$k \leq 2.$$

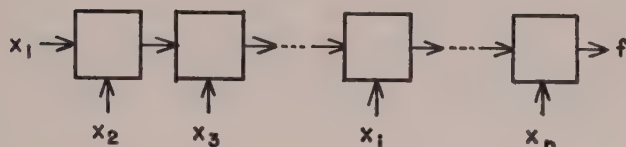


FIG. 6. Generalized decomposition of functions with  $m = 1$ .

### CONCLUSION

It has been demonstrated that no system can operate on arbitrary functions of  $n$  variables when  $n$  is large. The subset of all functions of  $n$  variables has been defined which may be realized by combinations of  $p$  functions of  $k$  variables where

$$p \ll 2^n,$$

$$k \ll n.$$

It has been shown that this class of functions is sufficiently cheap and sufficiently small in number that it could be a domain of complex systems. It has also been shown that two well-known classes of simple switching functions are members of this class.

We have really said no more than the following. The only functions of  $n$  variables which are of interest when  $n$  is large are those which may be built up from a reasonable number of functions of a very small number of variables.

### REFERENCES

1. A. M. TURING, On computable numbers, with an application to the Entscheidungs problem, *Proc. London Math. Soc.*, Ser. 2 **42** (1936),
2. W. S. McCULLOCH and W. PITTS, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5** (1943).

3. J. VON NEUMANN, The general and logical theory of automata, in *Cerebral Mechanisms in Behavior*, The Hixon Symposium, Wiley, New York (1951).
4. J. VON NEUMANN, Probabilistic logics and the synthesis of reliable organisms from unreliable components, in *Automata Studies*, Princeton University Press, pp. 43-98 (1956).
5. E. F. MOORE and C. E. SHANNON, Reliable circuits using less reliable relays, *J. Franklin Inst.* **262**, 191-208, 281-97 (1956).
6. H. AIKEN, private communication.
7. C. E. SHANNON, The synthesis of two-terminal switching circuits, *Bell Syst. Tech. J.* **28**, 59-98 (1949).
8. ROBERT L. ASHENHURST, The decomposition of switching functions, *Proc. of an International Symposium on the Theory of Switching, Annals of the Computation Laboratory of Harvard University*, Vol. XXIX, Harvard University Press, pp. 74-116 (1959).

### DISCUSSION

WILLIS: Let me make one comment. I do not want these things so elementary or universal that all you can use to describe them is the Boolean algebra, because this is what we have got.

NOVIKOFF: One thinks of a well-posed problem of just this sort. Someone once tried to compute the cheapest way that you can evaluate a polynomial of the  $n$ th degree, where you pay a price per multiplication and per addition. The conjecture is that the cheapest way—that the trick is roughly that you use the factor formula. When you first compute something of the form  $ax + b$  and then you multiply that by  $a_1$  and to it add  $b_1$ ; by  $a_{1x}$  and  $b_{1x}$  and then multiply that by  $a_{2x}$  and add  $b_2$ , and so on. He was able to prove that for cubics, which was as cheap as possible, but his proof is not particularly satisfying to me for quartics and the whole thing bogs down hopelessly about that. Now there is a perfectly simple problem which has resisted serious attempt at solution and is explicitly given.

HAYEK: One more speaker, I think.

AMAREL: I have a comment on the approach to the realization of switching functions. Now, we have in general  $2^n$  terms in a truth table; if we have say ten variables, we will have  $2^{10}$  terms. It is quite a large number. If we want to realize such function without any concern for efficiency we can easily do it, by using a huge number of components. Usually, we might be interested in functions of ten or twenty variables that are specified by a very partial truth table; perhaps ten or twenty terms per table. For these cases there are well-defined methods today, especially those developed by Quine and Mott, by which we can design quite efficient normal form realizations for any number of Boolean variables; there even exist computer programs for these methods. It seems to me that the essential problem now is to find out what functions can be realized by newly invented devices; it is possible that we can realize a desired class of functions by a small number of the new devices. I think that the remarks of Rosen are close to this point of view. I believe that the approach to the problem goes usually this way: you start with a given device and you try to find out whether the device can realize a certain function. The problem does not take the form: given a Boolean function or a class of Boolean functions, what are the ways in which the function can be decomposed in general?

HAYEK: Dr. Willis, do you wish to comment on this?

WILLIS: I agree in general with everything that has been said, but let me just make one remark. I think if we take functions like the threshold functions, which are, of course, in themselves general enough to build any circuit, and develop a switching theory around them and showing how we can put these things together to do other things, we are going to find much more structure, we are going to find search procedures in which we can find the most efficient way, I think we are going to be way, way ahead toward answering that first question, how to describe them.





## FRANK ROSENBLATT

*Cognitive Systems Research Program,  
Cornell University*

### STRATEGIC APPROACHES TO THE STUDY OF BRAIN MODELS\*

This is going to be a very informal discussion. I will try to follow some cues which I have collected from people here as to what they would like to hear about, since it is out of the question to present a complete résumé of the perceptron program in so short a time. I would like at least to touch on three main topics. The first is the general motivation of what we are trying to do, and I would like to deal with some common misapprehensions as to what our actual objectives and directions are. Second, I would like to indicate something of where we stand at the present time, and undertake some updating for those of you who have been following the perceptron program, as to the current problems that we are trying to tackle, and how far we have come. Third, I would like to say a few things about the relationship of all of this to biological memory.

First of all, as to motivation of the program: we are interested in brain models. By this, I mean that we are interested in the psychological properties of brainlike systems, not necessarily biological brains. I think the closer we come in the resemblance of our models to the biological brain, the more important it is to distinguish clearly between those occasions when we are talking about biological neurons and actual tissue and those when we are talking about hypothetical neurons and hypothetical nerve nets. The systems that we are interested in are fully axiomatized. They consist of abstract neurons in abstract environments.

---

\* In an attempt to preserve the character of the original extemporaneous presentation, the original wording and organization of these remarks has been preserved as far as possible. Those who are interested in the mathematical development of perceptron theory can refer to previous reports and papers (Refs. 1, 2, 5, 6).

At this point I would like to refer back to Dave Willis's remarks a few moments ago when he quoted von Neumann from the Hixon Symposium. Von Neumann's remark, as I remember it, was that if we can really define precisely and unambiguously what it is that we want a system to do, we can without fail construct a system to carry out this required behavior. I have no argument with this whatever. If we can really fully specify the behavior that we require from a system, then certainly we can construct a network to compute the required functions or control the required output of our device. The main issue which I have taken during the last few years with some of the theorists in this field concerns the adoption of this theorem (which dates back to Turing) as defining the basic strategy with which we will approach the brain model problem. The implicit assumption is that it is relatively easy to specify the behavior that we want the system to perform, and that the challenge is then to design a device or mechanism which will effectively carry out this behavior and which will (incidentally) resemble the nervous system. It is further assumed that such devices, once we have designed them, are intrinsically of interest as objects of study.

The position which I would like to represent is that it is both easier and more profitable to axiomatize the *physical system* and then investigate this system analytically to determine its behavior, than to axiomatize the *behavior* and then design a physical system by techniques of logical synthesis, which will in fact illuminate the functioning of the brain. Admittedly once the behavior has been axiomatized, there is no serious problem in synthesizing a hypothetical system to carry out the designated form of behavior. Having done that, I think there is a serious question as to the further interest of the particular model which is contrived.

First of all, it appears pragmatically that when brain models are synthesized in the manner of a logical computer, for a given *a priori* function, the systems which are produced tend to lack uniqueness. They are overdetermined and they represent "genotypes" rather than "phenotypes" (resorting to biological terminology). I think the laws of organization which we are interested in are the laws which characterize species of organisms rather than individuals within the species; and clearly from what we know about the organization of neural ganglia, it is most unlikely that

any two representatives of the same species really have the same logical structure or "wiring diagram" for their central nervous systems. Thus, while it is often useful in our perceptron studies to represent a particular nerve net in terms of McCulloch-Pitts neurons and find out what its logical properties are, the general laws which we would like to come up with are laws which can be stated best, I think, in terms of very general constraints on the organization of the system. These constraints permit us to define and to analyse broad classes of systems, without ever knowing precisely what logical functions are implicitly being made use of by the individual members of a given class. To require the precise logical structure of a nerve net in order to predict its behavior seems to me to be comparable to requiring the precise position and velocity components for every molecule in a tank of gas in order to predict its temperature.

There is a second angle which also helps discriminate among the various models which have been proposed. Many of the models which we have heard discussed are concerned with the question of what logical structure a system must have if it is to exhibit some property,  $X$ . This is essentially a question about a static system. Given the nerve net, which has certain fixed organizational properties, will it or will it not have property  $X$ , or how should a system be constructed in order to have this property?

An alternative way of looking at the question is: what kind of a system can *evolve* property  $X$ ? I think we can show in a number of interesting cases that the second question can be solved without having an answer to the first; that is to say, we can state the types of systems which will evolve certain properties of interest without being able to say precisely what are the necessary organizational constraints in our finished system once it has completed its evolutionary process. I think this comes close to some of Stafford Beer's remarks, and also to some of what Gordon Pask is trying to do in undertaking to evolve systems, the end results of which are interesting or fulfil certain conditions, without necessarily knowing what the final detailed structure is at the end of this evolutionary process.

Now I have said that we are interested in investigating closed, fully axiomatized systems. This is because these systems permit us to analyse their performance in detail and to perform experiments.

The object of analysis is a system (which we call the *experimental system*) which generally consists of three parts. First of all, we have a world  $W$ , which consists of a finite number of stimuli, patterns, or sequences. The axiomatization starts here. We want to know precisely what our system is going to be exposed to, and we may simplify this to the limit, provided we know exactly what it is going to see, or what probability is associated with its seeing each

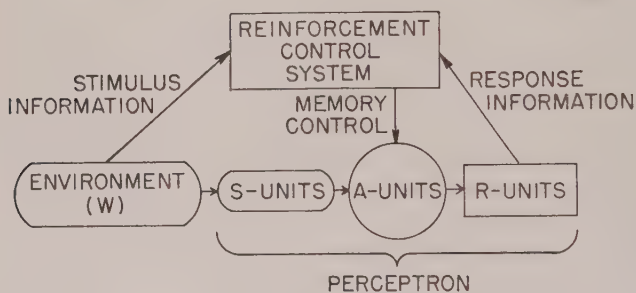


FIG. 1. Experimental system.

possible stimulus; if there are constraints on the sequences with which stimuli may occur, we want to know these constraints as well.

The second part of the experimental system is a network of signal transmission elements which are neuron-like in their basic properties; that is to say, they accept input signals and they emit output signals along connections. In particular, we are interested in a class of such networks which we call *perceptrons*, which consist of three types of neuron-like elements. First, there is a set of *sensory transducers*, which send signals into the central ganglion of *association units* which are really essentially McCulloch-Pitts neurons with a memory; the outputs of these go to one or more *response units*.

The last part of our axiomatic closed system is a *reinforcement control system*. This has in the general case two inputs. For one thing, it is able to observe the output of the perceptron, and take cognizance of the responses coming out of the system. It may also have an input directly from the environment or, alternatively,

from the sensory system. In the general case, the reinforcement control system is a correlating device which, on the basis of the output of the perceptron and the inputs from the external universe, decides whether the performance of the system is right or wrong at the present time. On this basis it may decide to "reinforce" some response. This decision is implemented in the perceptron by incrementing or decrementing the connection weights, which represent the memory of this system. The reinforcement control system may be a human experimenter (as long as he is following the axiomatic rules that we set up for this system) or may be part of a digital computer program, as in our simulation programs. For certain purposes we may cut one or another of these information channels. We have considered, for example, cases in which information from the environment to the reinforcement control system is totally lacking, and the r.c.s. makes its decisions entirely on the basis of the output of the perceptron. It is found that while such systems may occasionally achieve interesting types of organization, such phenomena are rather special and unusual. In general, if we want the system to learn things which are of interest to us as human experimenters, it appears that both information channels are necessary for the reinforcement control system. It must know something about what is going on in the "outside world" and it must have access to the output states of the perceptron itself.

For specificity, let us consider a *simple perceptron*, which consists of a mosaic of *sensory receptors* (the retina), a single level of *association units*, each of which has a number of input connections originating from points in the retina, and each of which has an output connection to a single *response unit* which emits a signal  $+1$  if it receives a strictly positive signal and  $-1$  if it receives a strictly negative signal. The input to the R-unit is the algebraic sum of all of the signals transmitted from the association units, at a given time,  $t$ . The  $S$  to  $A$  connections all have constant weights, but the weights of the  $A$  to  $R$  connections can be varied by a reinforcement procedure.

The types of experiments that we are concerned with are the types of experiments that you might perform on any animal whose psychology you are trying to investigate. The approach is essentially identical with that of comparative psychologists in trying to find out what it is that a rat or dog is able to perceive.



We undertake to run discrimination or detection experiments. Much of the problem in trying to study the potentialities of perceptrons, is in devising suitable experiments which will bring out particular psychological properties. We are trying to dodge the issue of defining ahead of time what we mean by perception, what we mean by cognition, or other terms which are being used here in a very loose sense.

We wish to be quite precise about the types of experiments which we wish to perform and the types of behavior which we are going to measure as indicating particular psychological functions. Memory, for example, is represented physically by changes in the weights of the connections; but if we ask what it is that the system remembers, we must determine this experimentally. We cannot go into the system, examine its state, and say that this is currently storing the image of a square. There may be no image of a square to be found. On the other hand, if we ask whether the perceptron will consistently discriminate between a square and a triangle, we may find that it does so, and we may similarly find that it discriminates between a square and a circle. However, it may fail to discriminate between a square on the left and a square on the right. For such a system we come to the conclusion that it has some general category for squares; it responds to these as being equivalent regardless of size and regardless of where they appear on the retina, but distinguishes between squares and non-squares. If prior to a particular training procedure it has *not* shown this consistency or this type of discrimination, and after a training procedure it does, then I think we are justified (operationally) in saying that the system now has a memory for squares.

We can show that the simple perceptron (with a three-layer *S-A-R* topology) does not recognize the analytic geometrical properties of the square; that is to say, if we show it four points situated at the corners of a field, the system does not abstract from its experience with solid filled-in squares to this new configuration, and recognize that there is a similarity there. This would involve recognition of relations, and it can be shown that recognition of relations is something which is considerably beyond the systems that we are dealing with right now. The properties on which a simple perceptron bases its identification are properties which involve intersections of figures on the retina. It must have

seen something very much like the square which it is asked to identify in order to recognize it, and as a matter of fact, it really is not fair to say that the system is "generalizing" at all, although this claim has been made from time to time. What this system is doing basically is mistaking cases of partial superposition of retinal images for true identity. In any case, it is the amount of overlap, rather than the geometric similarity, which determines the response.

Now to complete the description of the simple perceptron, let us assume the memory modification for this system to be of a form which we call the *alpha system*. This is a non-conservative system. Weights can keep growing, without bound. In this system we add an increment (a positive or negative quantity) to the weights of connections which are currently carrying a signal from active association units, under the control of a reinforcement control system. This system may employ one of several strategies. On the one hand, it may reinforce the perceptron universally every time a stimulus occurs, regardless of whether or not the response is correct. On the other hand, it may reinforce the perceptron only if the current response is *wrong*, and leave it strictly alone if the response is *correct* (according to some predetermined classification of stimuli). This second method is called the "error correction procedure", for which we have now proven the following theorem.<sup>(5)</sup> If we reinforce the system only when it makes a mistake, with the simple alpha system procedure of just incrementing active connections, and we leave it strictly alone as long as it is right, then always, 100 per cent of the time, if a solution to a given classification exists in the form of some assignment of weights to connections, the system will achieve that solution in finite time.

Now this means that the reinforcement control system only needs to know whether the current response of the system is right or wrong. It does not need to know anything more about the actual stimulus which is occurring, as long as it knows whether the current response is correct or not; on this basis alone we can guarantee that the system will converge to a correct solution if a solution exists. We cannot say at the present time what the probability is that a system constructed in this way *will* have a solution to a particular problem. We can say that if the system is large, the probability is very close to unity. We can also say that if the number of stimuli is equal to  $n$  and the number of association units

is less than  $n$ , then there will be some classification of the  $n$  stimuli which does not have a solution. On the other hand, in environments and classifications of the type that we are actually interested in, the probability that a system of reasonable size (where the number of association units may be much less than the number of stimuli) does not have a solution becomes very small.

Now we can approach the problem of system capacity by another means. We can examine the learning curves of this system analytically, and find that as we increase the number of  $A$ -units in the system, the probability of learning to identify a particular stimulus approaches unity, with an arbitrarily small error.

Let me now turn to some of the more interesting problems which we are working on at the present time.

In previous papers I have presented some statements about systems in which the reinforcement control system receives no information from the environment, and the perceptron is always reinforced as if its response is correct.<sup>(3,4)</sup> We now ask whether the perceptron forms an interesting dichotomy spontaneously under these circumstances, where "interesting" is defined in some arbitrary way by the observer, who no longer has any control over what the system is going to do.

We find that for simple (three-layer) perceptrons, "interesting" behavior is most exceptional under these conditions, although it may occur in certain highly constrained types of environments. For example, in the case of an environment of horizontal and vertical bars, certain types of perceptrons will tend to form a dichotomy with the horizontal bars in one class and the vertical bars in the opposite class. This will not occur, however, with arbitrarily chosen patterns, such as squares and triangles.

Now this opens up the question of what types of systems will, in fact, spontaneously recognize similarity between such things as translated patterns, regardless of their form and their position on the retina. We would like a system which will learn spontaneously that a triangle on the left and a triangle shifted across the field to the right are really similar to one another. We would like a system which will recognize this by itself. This cannot be done by "three-layer" perceptrons, having a simple  $S$  to  $A$  to  $R$  unit topology, regardless of what constraints we put on the retinal connections. If we introduce cross-connections between the elements of the

association system, give these cross-connections variable weights and give them a short time delay for signal transmission, then we have a system which shows some rather remarkable new properties. First of all, such a system will no longer respond simply to momentary stimuli but will begin to respond to sequences of stimuli. In such a system, the signal arriving at an association unit at a given time depends on the current inputs coming from the retina, combined algebraically with any inputs coming from other neurons within the association network. Consequently, the activity state of the association system is a function of the current retinal state and the preceding association state, which, in turn, is a function of the preceding association state a step before that, and so on back. Consequently, the state of association system at the present time (the set of cells which is currently active) is a function of the preceding sequence rather than the single stimulus which has just occurred.

Clearly, if we can get an evolutionary process to occur within this association system such that  $Q_{ij}$  (the measure of the intersections of the  $A$  sets responding to  $S_i$  and  $S_j$ ) becomes large for any two stimuli which are "similar" to one another, then we are beginning to evolve a system which will recognize similarity spontaneously. In other words, if we now try to teach this system, just as we did before, to distinguish between circles and triangles, it will no longer be necessary to show it a large sample of circles and triangles in order to get generalization. What we would like, of course, is to have a large intersection between the  $A$ -unit set responding to any two circles, and a small intersection between the set responding to a circle and a triangle, even though all of these stimuli may be disjoint on the retina. In the simple perceptron there would be no preferential bias in such a case. If we teach it to recognize a circle, the response will tend to generalize equally to the other circle and to the triangle.

There are two basic conditions which will permit such organization to occur in a cross-coupled perceptron. The first condition is a reinforcement rule for the cross-connections between  $A$ -units. This reinforcement is now entirely spontaneous, and is no longer controlled by the reinforcement control system. Let us state the rule for the alpha system for simplicity.

If two association units are both active,  $a_1$  at time  $t$  and  $a_2$  at



time  $t+1$  then a connection from  $a_1$  to  $a_2$  will gain in value by some small increment. If we want the system to be well-behaved, there should also be a decay effect, otherwise the system will ultimately become unstable.

The other condition is one on the environment. If the system is to learn to recognize similarity, it must see similarity, in some sense, exemplified in the outside world. We cannot show it an arbitrary collection of stimuli occurring in an arbitrary temporal fashion and expect it to acquire a concept of similarity equivalent to our own. We require an environment which is temporarily organized so that there is some sort of continuity present, where a stimulus is more likely to be followed by a transformation of itself under some admissible set of transformations, than by a completely unrelated pattern. A triangle is likely to be followed either by itself or by a displaced triangle rather than suddenly being transformed into an elephant or a clam.

These conditions (with a suitable choice of parameters) are sufficient for the model that I have just described to evolve an association system in which the  $Q_{ij}$ 's between similar forms begin to increase in magnitude. Having established a discrimination of squares and triangles in one part of the sensory field, this discrimination will now generalize to the other parts of the field, provided the perceptron has previously been exposed to a universe in which whatever it has seen (e.g. random dot configurations) has undergone the same sorts of transformations which it is now experiencing with squares and triangles. It has learned the *transformations* from its previous experience. We have demonstrated this effect successfully with a simulation program for a perceptron with 102  $A$ -units, fully cross-coupled (which means over 10,000 connections) all of them variable in weight.

The mathematical properties of this system have never really been properly stated at the present time. An attempt at a preliminary analysis was presented at the Chicago Symposium over a year ago.<sup>(6)</sup> This analysis really demonstrates only that the initial bias of the system (in the early stage of adaptation) is in the proper direction. It assumes that this initial bias will be maintained. Actually, the point at which the bias which we are looking for becomes strong enough to be felt, is also the point at which instability begins to occur. It happens that the effect we are after



comes in just a little ahead, so if we can catch the system just at the point before it becomes unstable, it is possible to keep the effect that we are after, and avoid the instability effect which would ultimately ruin the whole thing.

If we allow the weights to keep building up, the internal network essentially becomes "detached" from the environment, and new stimuli coming in can exert no influence at all on what is happening internally. So we want to catch the system at a point where the weights are just strong enough to affect the performance, but not strong enough to overpower the system. This can be done by a proportional decay mechanism, which leads to a dynamic equilibrium point being reached at a suitable stage. Currently, we are continuing the investigation of such systems, and we have hopes that we will soon be able to analyse the performance of these systems more rigorously.

Let me now conclude with some observations on the implications of these results for memory mechanisms.

First of all, there seem to be two basic types of memory dynamics which are useful in perceptrons. In the connections from *A*-units to *R*-units, where we would hope to use an error correction procedure, the most successful system investigated to date is an alpha system (reinforcing only active connections) with no decay, but where either positive or negative reinforcement is possible at the discretion of an "external" reinforcement control system. In the internal connections between *A*-units, however, the preference seems to be for a conservative "gamma system" (in which the active connections can gain only at the expense of the inactive connections, which must lose by a corresponding amount), where there is a steady decay of all weights at a rate proportional to their current magnitude, and where the reinforcement process goes on continuously, based only on local activity of the network, and without regard to the decisions of any external control system. The first type of mechanism permits the system to learn from an external "teacher", or by reward and punishment; the second type permits it to acquire an internal model of the "similarity structure" of its environment, as defined by temporal relationships of moving stimuli. While the biological significance of these conclusions is open to question, the relevant phenomena have now been clearly demonstrated in perceptrons.

## REFERENCES

1. R. D. JOSEPH, *Contributions to Perceptron Theory*, Cornell Aeronautical Laboratory Report No. VG-1196-G-7, Buffalo (1960).
2. F. ROSENBLATT, *Perceptrons and the Theory of Brain Mechanisms*, Cornell Aeronautical Laboratory Report No. VG-1196-G-8 (in preparation).
3. F. ROSENBLATT, Two theorems of statistical separability in the perceptron, in *Mechanization of Thought Processes* (Vol. I), H.M.S.O., London (1959).
4. F. ROSENBLATT Perceptron simulation experiments, *Proc. I.R.E.* **48**, 301-9 (1960).
5. F. ROSENBLATT, *On the Convergence of Reinforcement Procedures in Simple Perceptrons*, Cornell Aeronautical Laboratory Report No. VG-1196-G-4, Buffalo (February 1960).
6. F. ROSENBLATT, Perceptual generalization over transformation groups, in *Self-Organizing Systems*, (Eds. M. Yovits, S. Cameron) Pergamon Press, New York (1960).

## DISCUSSION

COWAN: May I ask what kind of projection have you got from your sensory units to your association units?

ROSENBLATT: It is usually a non-geometric projection. Let us assume that every association unit receives some fixed number of input connections. Some definite fraction of these are excitatory and some definite fraction are inhibitory, and each input connection has a weight of plus or minus one.

COWAN: What rule do you follow?

ROSENBLATT: I am getting to that. The only constraint will be a parametric one. We assume that the origin of every one of these connections is selected at random from a uniform probability distribution over the retina. This means that every possible configuration here may occur, and we have a probability measure over the set of all possible configurations, which is induced by the assumption of a uniform probability of associating any given retinal point with each connection.

Now we have a threshold associated with the  $A$ -units. In general, you will find that some set of association units will respond to any given pattern of the retina. If we displace the pattern, then we get a different set responding, which, in general, has an intersection with the first. If a displacement on the retina is very slight, the intersection will be very large. If we go to disjoint patterns on the retina, there will still, in general, be an intersection in the  $A$ -set, but it may become small. If we raise the threshold sufficiently on our association units, and if we begin introducing constraints on the connections, we can design systems which will guarantee disjoint sets for any non-identical stimuli on the retina, but we don't wish to impose this constraint on the system. So in general we have association units, each of which will respond to a large number of alternative patterns on the retina, and any two patterns, regardless of whether or not they are disjoint on the retina will, in general, activate some  $A$ -units in common.

The statistical problem in all of this is, can we, on the basis of such an organization, discriminate reliably between two classes of patterns on the retina, such as squares and triangles? It turns out we can. This is a function

of the size of the system and its parameters, and we can, by making the system large enough, always discriminate with an error less than  $\epsilon$ , whatever we define  $\epsilon$  to be.

Consider the following special case. Suppose we associate with each  $A$  unit one connection from every point on the retina; let the number of excitatory connections equal the number of points illuminated by a stimulus, and let the threshold of the neuron be equal to the number of excitatory connections. In this case, if we have one association unit for every such configuration, we will have one association unit responding to every possible stimulus and to no other stimulus. We now have a disjoint set of  $A$  units for every possible stimulus. Clearly by following our training procedure we can teach this system anything we want. This is a universal Perceptron, so to speak, in which three layers can form a dichotomy of any set of stimuli. It is not a very interesting system, but it does exist.

COWAN: Up to sets of sets of sets.

ROSENBLATT: Okay. I am talking about simple dichotomies at the moment.

COWAN: I am thinking this seems to be essentially Uttley's conditional probability scheme.

ROSENBLATT: In this special case, it has become similar to Uttley's scheme. I am by no means recommending this as an economical or particularly interesting system, but it is an existence proof that you can, indeed, with a simple perception, discriminate patterns regardless of how uniformly they are spread over the retina. Phenomena of this sort seem to puzzle some people.

COWAN: You can always add associations.

ROSENBLATT: Yes.

PLATT: How many retinal elements have you?

ROSENBLATT: I have not specified the number. We have simulated systems with over five thousand. Our Mark I system, which is currently operating, has four hundred. It uses a  $20 \times 20$  field.

PLATT: What form are these in? Are these photocells?

ROSENBLATT: These are photocells in the Mark I. They are simply logical entities in our simulation programs, where we simply assign one bit in the storage of the digital computer to every sensory point, and the pattern consists of a collection of bits.

COWAN: What is the criterion which determines which is the right response?

ROSENBLATT: This is an arbitrary classification function which I assume has been set up ahead of time, and is known to the reinforcement control system. If we wish to discriminate squares from triangles, then I also assume that the reinforcement control system itself can recognize, on some basis, squares and triangles.

Now it is possible that in a particular environment squares are hot and triangles are cold, so it need not necessarily have to do this by visual means. It might be responding to thermal information, but it has some basis for deciding when something is in class one and when something is in class two.

COWAN: Let me make one comment. I am not sure I understand so far. The classification is based upon the measure of the intersection of two particular inputs?

ROSENBLATT: The classification that we are trying to teach the system is arbitrary.

COWAN: The similarity criterion is merely the measure of the intersections of the two particular stimuli.

ROSENBLATT: The similarity criterion is the measure of the intersection of

the responding sets of  $A$ -units. Now actually, in current notation the intersection between two stimuli  $S_i$  and  $S_j$ , or rather, the expected value of this intersection (if it is normalized) is represented by  $Q_{ij}$ , which is the probability that an association unit constructed with given parameters will respond both to  $S_i$  and  $S_j$ . This is a quantity which is tabulated at the present time.<sup>(2)</sup>

These quantities determine the entire dynamics of these simple systems. The generalization from  $S_1$  to  $S_2$  depends on  $Q_{12}$ . That is to say, if we hit  $S_1$  with one unit of reinforcement, the amount of reinforcement that  $S_2$  gets will be proportional to  $Q_{12}$  in this simple, non-conservative system.

COWAN: If you did not have that random mapping from  $S$  to  $A$ , but if you had the particular association area, then you would have in a sense enough in the system in which you did not need the reinforcement control system; but because you are randomizing from  $S$  to  $A$ , therefore you have to pick out a particular response to get the required behavior. Is that the case?

ROSENBLATT: You will need some sort of reinforcement in any case if you assume you are starting with zero weights.

COWAN: It is not needed in Uttley's system, and all he is basing it on is the measure of the intersection of the sets or intersection of patterns in this case.

ROSENBLATT: Well, Uttley is reinforcing his system.

COWAN: Well, not—so far, the equivalent system you have got is just this. I mean, I am not talking about any decays and conditional probability computations so far. I am in a sense just talking about measures of similarity in the input.

ROSENBLATT: Okay. The measure of similarity, I agree, is equivalent up to this point. But the internal structure of the Perceptron is considerably simpler.

COWAN: Yes. All I am pointing out is that, in his terms, if you would not map by random manner, and leave out the reinforcement control system, you would then be doing exactly the same thing as Uttley does in his system. The point I am making is that if you do not randomize from  $S$  to  $A$ , then you do not need to teach the system.

ROSENBLATT: We do not need to teach the system unless we want to get particular responses for particular sets of stimuli, which is what we want the system to do. This is a learning problem in either machine. I want the freedom of reversing the responses by a learning process.

COWAN: But is that the real problem? It would seem to me that the real problem is to put the squares in  $R$  sub-square, and the triangles in  $R$  sub-triangle. Whether you call them triangles or squares does not appear to be crucial so long as you have got the dichotomy.

ROSENBLATT: Accept for the time being that what we are trying to do is to carry out a *learning experiment* in which we have a response which is indeterminate. This may be an avoidance response or an approach response, if you like, and it is something in which we want the freedom of teaching this system to go either way. We assume this is not something the system knows initially, this is not built in, and for our purposes, this is a learning problem.

VON FOERSTER: I think I would really like to put our symposium to the test at the moment and consider this as a symposium in which we get together in order that something new and interesting should evolve from some of the contributions. I invite Frank Rosenblatt, and all of you, to consider the remarks which came about, if I may use the term, Perceptron and we think it is a very interesting and a very important contribution to the field of our interest. Frank was one of the first people to point out the possibilities of learning, making decisions, and recognizing things. In the last thirty-six hours



I think we have heard not only physiological evidence, but also some other, let us say, analytical evidence and structural evidence, that if you introduce certain pre-structuralization of our systems, then we can see, of course, at once, triangles and squares and things of this sort, I mean, there is no problem there. I would like to enumerate as a point, Mr. Novikoff's paper, Platt's idea. I could enumerate two or three other devices, based on different principles, which are worked out in our lab.; one is the differential geometry point of view; this was worked out by me, and Lars Lofgren has worked out a topological point of view, and Mr. Weston, a dynamic point of view, where problems of this sort could be solved at once for a certain class of figures. No work needs to go into learning, and you can extract right away from what you may call an environment, right away extract certain structural properties, so it is perfectly clear that I think we must not put so much effort in extracting from the property certain simple classes as triangles, squares, and so forth. Those can be done today very easily. However, the Perceptron, of course, remains in its very powerful position, namely, we have already pulled out these classifications, these things can then be fitted into another adaptive system which may learn them, putting names on the things or operating with these devices. Now the real problem which I see at the moment, in which way can the basic principle of Perceptron be employed to these meta-classes which have been extracted from the environment? This is still, for instance, the overlap principle, and the principle of calculating the conditional probability to be employed. For instance, I would see no difficulties in employing a Perceptron behind a set of systems which would work out, for instance, certain measures, before this device and then employ the principle of conditional probabilities to those in order to make further abstractions. I could think of some devices we have in mind where these principles can still be employed, and I really invite the group to aid me in my questions, if this question has been properly put.

COWAN: On the question of Frank's and Uttley's systems, I think I see now the difference. Uttley's system is, to use an aphorism, classifying to learn, whereas Frank's system is learning to classify. If we start looking at it in those terms, it is fine. The second point is, if we have those response systems connected to the world and then to the response, it is really a human observer looking at a black box and choosing his output to be what he wants it to be, and being reinforced by a kind of Skinnerian procedure to produce that type of condition; so that this particular thing is not so interesting. It is when we eliminate the connection from the world and replace it by the connection from the sensory system, then we have something like a reticular formation.

HAYEK: Which is which? Now I really get confused. Which is learning to classify and which is classifying to learn?

COWAN: Uttley's system is classifying to learn and Frank's system is learning to classify. Now when we, in fact, remove that connection and we make a connection from the sensory system to the response control system and from the association area and we look at the response which is still a non-sensory system, in a sense, because it is coming in, now we have something like a reticular formation in terms of performance and then in terms of this model. I think it is becoming very similar now to Beurle's model and I think we could, in fact, synthesize in Frank's model the necessary kind of constraints we have to put on the system and Beurle's model. I think the point Frank made in particular about the stability-instability points in the particular type of system is, in fact, the same point that Beurle made when he talked about



the decay of the wave in his system and the expansion of the wave and that these are quite related. This is the sort of thing we should have to look at.

ROSENBLATT: These are related but not quite identical.

COWAN: May I make one more comment? Uttley's system is somewhat strange in that it falls outside the line of other models. Now I think that is because he interpreted his system as one in which his particular units were units of a particular system and not states of a system. However, if we re-interpret his model in terms of an atomic state description, which normally Uttley's system would not have, then we get much closer to Beurle's model. I think if we look at your *A*-units as states of the model, then maybe we would have a cross-correlation between Dr. Beurle's model and your model, and in those terms I think we do get something. Peter Greene, I am sure, has a lot of things that are relevant to this system, so we can really try to make some kind of closed synthesis.

SHERWOOD: Your initial diagram—I don't know that I quite understood your purpose from your introductory remarks, but it seems to me, should there not be a direct feedback from *R* to *A* and from *R* to *S*, generally?

ROSENBLATT: In a more general case, yes. I was diagramming.

SHERWOOD: Because if you want to be similar to biological substance, I know of no biological substance where there is not a direct feedback.

ROSENBLATT: Yes, there should also be a feedback from *R* to *W*, as far as that goes. I was limiting myself to the network that I was going to discuss rather than a fully general diagram.

SHERWOOD: The feedback could be weighted in such a way that it could be reinforcing.

BEURLE: There have been several references during this symposium to the problem of providing enough storage space to accommodate any possible combination of sensory data that may be received. This problem is serious only if we are considering the complete and unambiguous type of classification machine, in which a storage compartment must be provided for every possible combination. One can overcome this difficulty if one is prepared to use an adaptable classification machine that can utilize a small storage capacity economically, by moulding it to accept just those combinations that do occur regularly. We then avoid the need for astronomical storage at the risk of introducing occasional ambiguity or incompleteness of classification. This principle of adaptability is used very widely in communication equipment. When one has finished making a long-distance phone call, most of the apparatus involved along the route is disconnected, and is then immediately available for use by other people making phone calls between completely different centers.

There seems little doubt that information storage in the brain must be in the latter category. We take advantage of the regularity and uniformity in the world in order to make do with a reasonably modest storage capacity. At the same time it is inevitably true that, as a result of our limited mental capacity, we miss an enormous amount of information available to us because we are in blissful ignorance of its existence.

COWAN: Uttley's system is really a digital computer with a punch card input, and it is the punch card system that does the classifying in this case.

PASK: I entirely agree with you, Dr. Beurle. I wrote a paper, as you know, on these undifferentiated systems, and I agree there is a certain measure of identity between a system such as Uttley's, and, from what I gather, the more elaborate Perceptrons. It is certainly true that such systems are more adaptive

and are indeed able to shape conceptual categories in terms of their peculiar world. Indeed, this is their purpose. But just because of this, and just because I am biased so much in favor of reinforcing sludges, I think it is particularly important to declare that *when* we know the world we want a system to live in, we should give it the kind of percept it needs. Now you can, of course, train it to build up the filters that extract this percept—with more or less tedium depending upon the criteria of similarity entailed—but, gentlemen, such a thing is believable, nay, demonstrable, given enough time, and the only value of the exercise is didactic. So let us adopt the elegant methods Al Novikoff has talked about, and those developed by von Foerster and his colleagues (and those of Jerry Lettvin, in a different language on a different occasion), whenever we are in this happy position of knowing what image of the world the system should have.

The Perceptrons and sludges and trainable whirligigs come into their own when we do not know about the world (not only are we ignorant of *what* image, but also of what *kind* of image to think about). Let me sketch for you a picture not half so stupid as it sounds. The central character can be a Perceptron, if you prefer it, or one of my sludges, though Frank Rosenblatt's machine is far better developed and analysed. Equip it with arms and legs wrought from its response units, so that it can crawl around the world and—this is the important thing—let its interior parts be in intimate contact with the world. The parts are not only *A*-units, but also potentially receptors. When this crawling thing, restricted by its arms and legs, gets reinforced, bits of *A*-unit will specialize to become receptors suited to the world it is in. In a visual world, to light; in an auditory world, to sound. Svoboda, incidentally, suggested such a thing at the Second Congress of the International Association of Cybernetics at Namur. This is within the logical capabilities of the system though, I will admit, the components are unsuitable.

You all know the trick, so I shall not tire you with it. Allow me to set it in context. What happens here when Charlie the crawler develops his special senses is precisely what the embryologists used to get so excited about, that an embryo developing along a *quantitatively* specified gradient and with a fixed rule of evolution (Charlie's rule is entailed in his arms, legs, and *A*-units) will suddenly undergo *qualitative* change. A special kind of differentiation occurs such that we as embryologists experimenting (or observers looking at a form recognizer) have to use a different sort of experimental technique. In conclusion, I have so much faith in the Novikoffs, Lettvins, and von Foerstes of this world that I doubt if it is ever worth making an assembly to learn about invariances under a group of transformations of a figure on a retina or in a delay line. This is a mere hunch, of course, but I think that when the position is reversed, the Perceptrons come into their own, when Charlie the crawler has to act like an embryo. I suspect—again as a hunch—that Charlie is needed in practice more often than appears at first sight.

PLATT: I would like to know about the more sophisticated Perceptrons. In Skinner's experiments he runs into the case of superstitious pigeons, who have been reinforced essentially at random at first or at rare intervals and they come to value the wrong things, but they never lose this evaluation because of random elements when they were first reinforced. Now I would like to ask if the more sophisticated Perceptrons are capable of this kind of superstition associated with random aspects of the pattern which they first get hold of. The second question is somewhat similar. That is, if you distort a pattern, if you learn to distinguish two patterns and you distort one toward the other,

can you cause abnormal delays in the machine or systematic mistakes of the normal patterns, as though it were essentially psychotic?

ROSENBLATT: Let me answer the second question first, because I am a little clearer on it. In the models that I have described, the timing of the system is entirely determined regardless of the conditions. We will, in general, get an increased probability of error if we distort one pattern to resemble another, although there will not (in the type of system that I am discussing) be an increase of delays in response. On the other hand, it would be very easy to think of a system in which there would be increased delays: if we have the model in which the response units, instead of responding instantly to their signal, integrate this signal over a period of time, as a biological neuron would, so that they are really responding to a bias which builds up over a period in which an excitatory input eventually gains dominance over inhibitory input, then you would expect additional time delays as well. Now could you restate the experimental observations you were referring to by Skinner?

PLATT: Well, if a pigeon is reinforced at random by having a peanut drop into the box and it happens to be standing on tiptoe when the first peanut comes, it then tends to stand on tiptoe, let us say, in the belief that this will bring peanuts. And your machine seems to have this possibility of being reinforced, so to speak, for a random element at the very first and thereby making mistakes or becoming superstitious in the same sense.

ROSENBLATT: First of all, this assumes that there is more than one binary response. That is, there must be the possibility of additional output information as well as the identifying response which we are asked to distinguish. Now we have worked with systems with multiple responses. In these cases there certainly may be an extraneous response which happens to ride along with the one that we are reinforcing. These extraneous responses will tend to persevere in later activity.

H. D. CRANE\*

## THE NEURISTOR

Coming from the more prosaic and deterministic field of digital computers, I must say that I have enjoyed listening in on this session on the controversial subject of self-organization. Although this paper will not represent a contribution to the body of information on this subject, I will at least commit the common sin of mentioning the word "neuron".

Neuristor is the term assigned to a class of devices. The conception of this type of device was motivated by a consideration of electronic miniaturization, and an attempt to see if there are some basic problems that can be singled out. One does not have to search very long, however, before becoming troubled by the problem of wiring. One aspect of this problem that is nasty is the high resistance of very tiny "wires". Considerations of this kind were finally boiled down to one general question. Is it possible to build a system (e.g. a digital computer system) in an environment in which good wire is not available for signalling? A recipe for survival in this environment is to consider active channels for signal transmission, in order to overcome the high attenuation of the lines. One of the results that becomes immediately interesting, however, is that although the active channels are initially considered merely as interconnections for conventional circuitry, once you have such active channels, at least of the type that I am going to discuss, you do not need the conventional lumps of circuitry at all. In other words, these active channels are logically far more powerful than indicated by their role of connecting together conventional circuitry.

---

\* Stanford Research Institute, Menlo Park, California. The neuristor is the subject of the author's doctoral thesis, and a detailed report on the subject is available from the Stanford University Electronics Laboratory—Report 1506-2, Neuristor Studies, July 11, 1960.



Let us first consider an ordinary chemical fuse, and talk about its gross properties. You can think of a fuse as a line of "resting" (or potential) energy with a threshold of excitation. When it is excited beyond its threshold, a discharge zone forms that propagates with uniform velocity. Thus, if the burning zone is considered as a signal, we can say that the signal propagates without attenuation. The fact that a fuse is a one-time device, however, makes it clearly of no interest for practical logic realization.

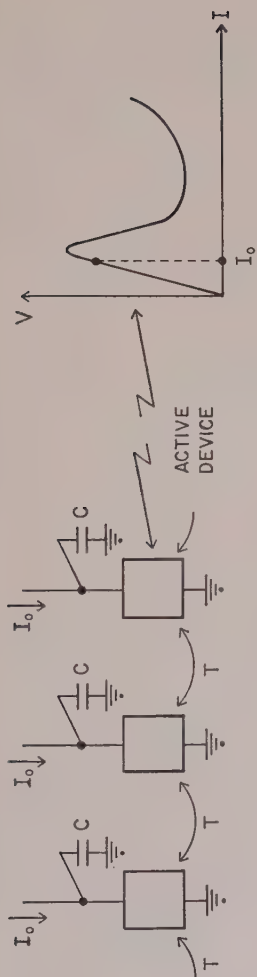
If we now look at the neuron, and pay particular attention to the "all or none" propagation mode along the axon, then we find that the signal propagation mode is similar to that exhibited by a fuse, but that the axon is not a one-time device. The neuron is "self-healing", after each discharge, and may be used again—in fact, an indefinite number of times. If I may steal a phrase from the neurophysiologist, the time of healing is referred to as the refractory period. In this sense, a fuse could be thought of as exhibiting an infinite refractory period. Thus, a neuron exhibits: (1) a threshold, (2) a uniform velocity of signal propagation, (3) attenuationless pulse propagation, and (4) a refractory period.

Neuristor is the term applied to the class of such devices that exhibit these properties. Thus, a neuron is merely an example of a device of this class, realized in an ionic medium. From a device viewpoint it is of interest to consider realization possibilities in other media as well—in particular, electronic forms.

Although the device aspects are interesting, I would rather concentrate here on the logical power of such devices. However, to introduce this material, I must consider some aspects of the device. Let us consider how such a device might be made electronically. I show a simple arrangement, Fig. 1a. Consider a set of classical monostable circuits, arranged into a chain, and coupled so that the firing of one stage triggers the next, and so on, each stage going through exactly the same process as its neighbor, but at a slightly different time. Thus, we have an iterative line with bilateral coupling.

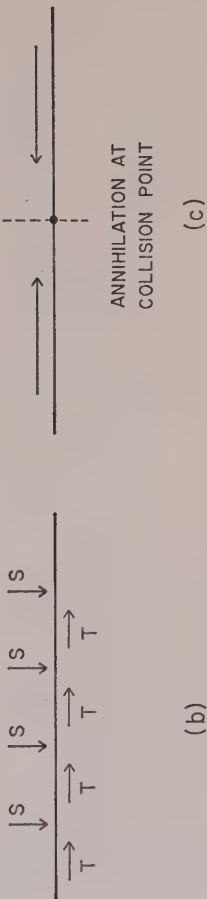
There are three properties necessary for a classical monostable circuit. These are illustrated in Fig. 1 as an active device (say, a thermistor), an energy source  $I_0$ , and an energy storage device, capacitor  $C$ . In the resting state there is a (resting) voltage,  $V_0$ , across  $C$ , so that there is a (resting) energy stored in the capacitance.





BILATERAL TRIGGER COUPLING

(a)



(b)

(c)

FIG. 1

Upon suitable triggering, the stored energy is released, and dissipated in the active device. Thus, the thermistor is heated, and with suitable trigger coupling this local temperature rise can result in the triggering of the next stage, and so on.

Now this thermistor circuit is used merely as an example, and is of no special interest. But it is simple to show that such structures can be made in a totally distributed fashion; i.e. by laying down a layer of "this", and a layer of "that", and so forth. However, I would rather not go into that detail now. Instead, I will appeal to your intuition that such lines could be made.

I will symbolize a line or channel of this type with a line segment, and when I draw a line, I will mean a linearly distributed device with the four properties previously indicated. Then how could we visualize a discharge moving along this line? I think of it this way. We must pay attention to two variables. There is a trigger variable and an energy variable. A local process of triggering puts something into an active region, which in turn causes the release of some energy; the stored energy is converted into the trigger form, which results in triggering of a neighboring region, which causes new energy to be released, and so on (Fig. 1b). It is very much like the propagation of a linear electromagnetic wave, where a change of  $E$  makes  $H$ , a change of  $H$  makes  $E$ . Here, in this non-linear case, a change of  $T$  makes  $S$ , and a change of  $S$  makes  $T$ , where  $S$  is used to represent the generalized energy storage variable, and  $T$  represents the trigger variable.

A very important property to note here is the following. These lines are highly non-linear; superposition does not apply. Thus, if two ends of a line are simultaneously excited so that two discharge zones approach each other, then these two propagating zones are annihilated at the collision point, Fig. 1c. This is so since, at the instant of collision, the energy on either side of the collision point is depleted. That is, the portions of line on either side of the collision point are refractory. Therefore, the net result of the collision is that both propagations are annihilated and the region about the collision point recovers to its resting state. We will see that this collision property is very useful.

Now I said that such devices are logically powerful enough so that they alone can be used to realize all digital logic. Clearly, however, one such device is of no particular interest. Therefore

we must define ways in which we will allow lines to join. Now it turns out that there are two basic junctions, the *T* junction and the *S* junction. The *T* junction is really quite simple. Suppose that we bring several lines together end to end, as indicated in Fig. 2. This can be easily visualized in terms of taking a few fuse wires and making a knot in them. If a discharge process arrives on any line, the discharge process reaches the junction and starts discharges on each of the other lines. In principle, the junction can be made as large as we please. But, three is enough. I mean, three lines coming together is enough for complete digital logic realization. It is a basic type of fan-out mechanism.

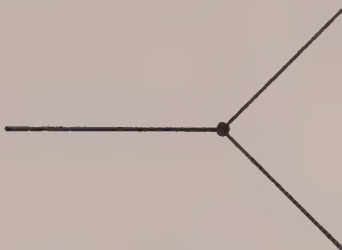


FIG. 2.

Let us now look at an *S* junction. It is entirely different. In this case, lines come together, side by side, and share energy. Consider again the lumped model for a moment. Each stage has its own source of resting energy. However, let us now take two independent stages and make them share a common energy storage, but with no trigger coupling between them, Fig. 3a. Then either stage has access to that energy source. If a particular stage fires, and takes the energy, then it becomes refractive until the energy rebuilds, but, the other stage becomes refractive as well, since it shares the same energy source. Let us now consider coupling a pair of lines such that adjacent stages of each line share energy, Fig. 3b. This is an *S* junction, and it is symbolized in Fig. 3c, where two lines are shown side by side; the cross lines emphasize the common energy coupling.

Now a very important property of an *S* junction is the following. Suppose two pulses are so timed that they collide on the junction,

Fig. 3c. Now I say that even though the pulses are arriving on different lines, they are annihilated just as though they collided on a single line. Again, at the instant of collision, the energy on either side of the collision point is depleted. Since both lines of an *S* junction share energy, a pulse on either line depletes the energy available to both lines. This type of collision situation provides a very powerful logic function. Thus, via an *S* junction, a signal on

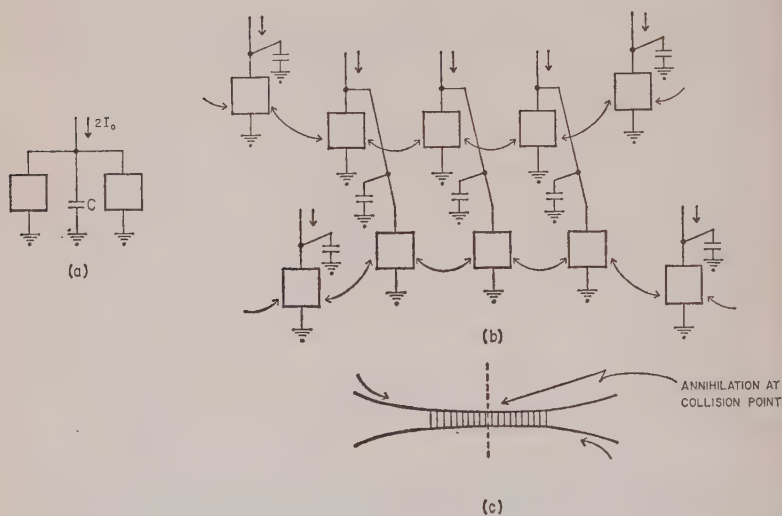


FIG. 3.

one line can control the transmission properties of another line, but without itself generating a signal on that line.

It is convenient to introduce one other junction, a combination of *T* and *S*. This will only take a moment. We take an *S* junction and join two of its lines together at a *T* junction, Fig. 4. A pulse arriving from *A*, passes the *S* junction, and generates two pulses at the *T* junction. Now assuming that the loop is long enough, then by the time that the pulse *B'* reaches the *S* junction, the latter is recovered, and *B'* will pass the junction. But suppose that the loop is shrunk to the point where the *T* junction is placed right at the end of the *S* junction. Now when a pulse from *A* reaches the *T* junction, the pulse *B'* cannot propagate backwards, since at

that instant the  $S$  junction is refractive to the left of the  $T$  junction. This is a handy fan-in structure. Thus, a pulse from  $A$  (or  $B$ ) will excite  $C$ , but not  $B$  (or  $A$ ). This junction is labeled a  $T$ - $S$  junction.

Thus we have two symbols, a  $T$  junction symbol and an  $S$  junction symbol, and everything that we do from now on is really a "game" on these two symbols. With these allowed interconnections of (neuristor) lines, we can make arbitrarily complex



(a)



(b)

FIG. 4.

logic networks. How could we analyse and synthesize such networks? There is one observation to make immediately. A pulse introduced into an arbitrary network composed of these two junction types can only die, or be annihilated, in two ways. A pulse can run off the end of an open line, in which case there is nowhere for it to go, or it can collide destructively with another pulse. These are the only ways. At first one might be horrified even to think of keeping track of all these pulses, but it is interesting that the situation is quite manageable if we study three classes of collision possibilities: (1) collision of isolated pulses, (2) collision of a pulse with a pulse train, and (3) collision of a pulse train with another pulse train. Now we are certainly not going to consider all



of the possibilities here, but I would like to give a few examples of how to get going.

As a sort of warm-up, let us consider some simple examples of pulse-pulse collision. We have been considering a line that is perfectly symmetrical (bidirectional); that is, we can light a fuse on either end. As a first exercise let us consider the possibility of creating a structure between two terminals  $A$  and  $B$ , such that a pulse initiated at  $B$  will reach  $A$ , but a pulse initiated at  $A$  will

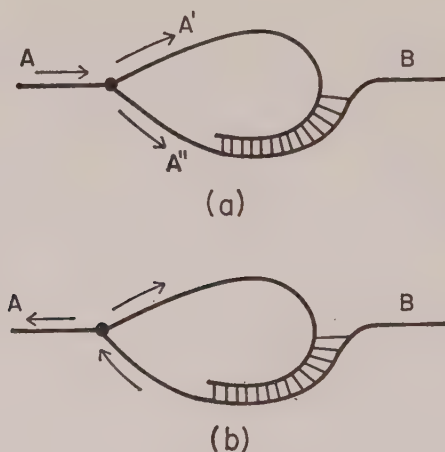


FIG. 5.

not reach  $B$ . The structure of Fig. 5 exhibits this property. This structure involves what I think of as a self-collision process; i.e. a collision between pulses that are very closely related. Thus, the two output pulses simultaneously generated at a  $T$  junction are so related. Now there are many ways that we could make these pulses interact. One such manner of interaction is indicated in Fig. 5. For instance, a pulse from  $A$  generates two pulses,  $A'$  and  $A''$ , at the  $T$  junction. These pulses are then made to collide on an  $S$  junction so that both pulses are annihilated, Fig. 5a. In particular, no pulse reaches  $B$ . Now, however, a pulse coming from  $B$ , passes the  $S$  junction unimpeded, reaches the  $T$  junction, there generating two pulses, one reaching terminal  $A$ , and the other propagating

back to the  $S$  junction. The latter pulse reaches the  $S$  junction too late to affect the original pulse, but we must take account of it, since it puts a limit on how soon again we can use the junction.

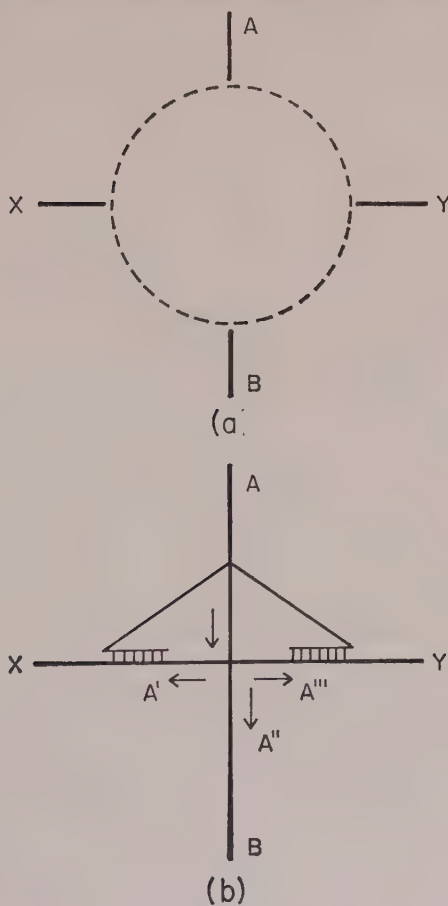


FIG. 6.

We must then think of a refractive period of the structure itself. Thus, the unilateral structure of Fig. 5 is a good example of the use of a self-collision.

Another example of pulse-pulse collision is the following. Consider the problem of realizing non-planar logic, where conven-

tionally one is faced with the necessity of using three-dimensional wiring, i.e. crossing wires are not allowed to touch electrically. Similarly, we could let two (neuristor) lines cross, in three dimensions, but here the result of letting them "touch" upon crossing is not so disastrous. Actually, we would then obtain a  $T$  junction, and now we are motivated to ask whether an appropriate structure could be placed around the junction, so that we could obtain a suitable control crossing structure. That is, in terms of Fig. 6, a structure is desired such that line  $AB$  could be used without affecting line  $XY$ , and vice versa.

I think that there will not be time for me to indicate the nature of the necessary structure, but it is relatively straightforward. (Thus, in Fig. 6b, a pulse from  $A$ , upon reaching the  $T$  junction, generates three pulses. Only the pulse labeled  $A''$ , aiming at  $B$ , should be allowed to survive. From the nature of the structure, it is clear that the other two pulses,  $A'$  and  $A'''$ , are annihilated by self-collision structures.) Let me just say that a structure like this is very similar to a railroad crossing problem. If you cross tracks in a plane, and a train is a thousand feet long, you do not dare let anything come on the other track for a time equivalent to a thousand feet of train. That is, there is a refractive period to the structure, and if you cannot live with that property, then you had better use three-dimensional under-and-overpass structures. Now the same thing applies to the structures considered here. After you use one line, it makes the structure refractive for a period, and you do not want to use the other line during that time. Now it is not obvious, but it can be shown that even with this restriction on timing, that an arbitrarily complex digital computer can be realized in a plane, in this way, with every crossing being used in a safe manner.

One last example of pulse-pulse interaction. If you take a line that is at least one refractive period long, and close it on itself to make a ring, then that ring represents the minimum length ring for storing a pulse which, once started, travels indefinitely. That is, by the time that the pulse arrives back at its original position, the line has recovered. Now this is a very basic structure for storing a binary variable. Thus, a pulse can be circulating on the ring, or not.

Let us just consider how we might put a pulse into a ring.

Assume the ring is quiescent. If we try to excite the ring via a  $T$  junction, Fig. 7a, then we will not be successful, since the two generated pulses will merely destroy each other on the other side of the ring. In fact, this is merely a simple example of a network theorem which states that in an arbitrary  $T$ -network (i.e. a network involving only  $T$  junctions) it is impossible to set up reverberations (i.e. circulating pulses). So then, how could we set up a

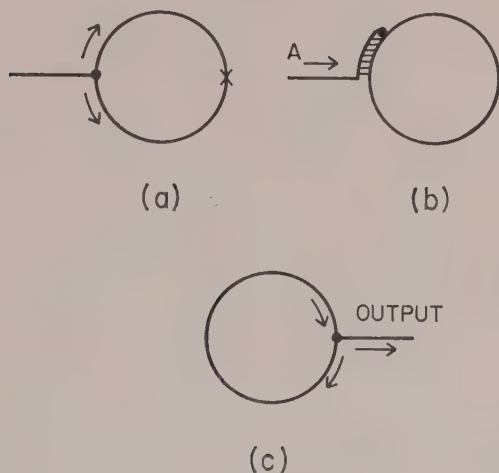


FIG. 7.

circulating pulse on the ring? With a  $T$ - $S$  junction, for instance, it is simple (Fig. 7b). From the  $T$ - $S$  junction properties indicated in connection with Fig. 4, it is clear that a single pulse  $A$  will initiate such a circulation. Similarly, we can kill the pulse very simply once it is circulating, but let us not consider that now. Rather, consider an output line coupled via a  $T$  junction (Fig. 7c). Each time that a circulating pulse reaches the  $T$  junction, an output pulse is generated, as well as another pulse which continues the circulation. Thus, if the ring is circulating a pulse, a uniform pulse train is obtained on the output line. If no pulse circulates on the ring, then the output line is not excited. If the ring circulates a pulse, let us say that it is storing the binary value *one*; otherwise it stores the binary value *zero*.

Finally, let us consider what can be done in the way of pulse, pulse-train interaction, Fig. 8. Suppose that a signal (a discharge zone) arrives along line  $C$ . With no signal being propagated along line  $AB$ , then the pulse from  $C$  will continue to propagate along its line, pass the  $S$  junction, and finally disappear off the end of its line. However, during the time that it takes this pulse to pass the  $S$  junction, it is impossible to pass a signal from  $A$  to  $B$ , since it would destructively collide with pulse  $C$ . Thus we can say that the single pulse  $C$  closes the line for some time  $\tau$  (i.e. inhibits its use). Suppose, however, that we would like to close the line permanently. Then what we would need is a pulse-train along  $C$ . Let the period of the pulse-train be  $T$ . The source of this pulse-train could be a storage ring, indicated by the dashed structure in the figure. If  $T < \tau$ , then it is impossible to pass a pulse from  $A$  to  $B$ , without having it destroyed by collision with the pulse-train. However, if the storage ring were not circulating a pulse, then the line  $C$  would be unexcited and the line  $AB$  could be used at will.

(Incidentally, suppose  $T$  is greater than  $\tau$ . In this case you can see that the pulse-train does not completely close the line; there are "windows" in the closure. If we try to pass a pulse from  $A$ , then there is a probability of the pulse going through. This then becomes an interesting probabilistic gating element. It can be shown that the  $p$  (probability) value of the gate can be made an arbitrary function of a set of logic variables. Thus, the  $p$  value can be altered during operation.)

Returning to the original case where  $T < \tau$ , it is very simple to define a transmission function for the gating arrangement of Fig. 8, which is analogous to the common binary (relay) transmission function. Thus, assume that the ring stores variable  $x$ . If  $x = 1$  (a pulse circulating on the ring) then the line  $AB$  is completely inhibited to pulse transmission. If  $x = 0$  (no pulse circulating on the ring), then signal transmission can be accomplished at will over line  $AB$ . It can be shown that by using the basic gating arrangement of Fig. 8, all of the classical analysis and synthesis techniques developed for relay networks can be directly applied; although it is also possible to synthesize networks that have no simple relay analogues.

The thing that seems of most interest to me in all this, aside from the fun of playing this game of symbols, is that we have a



network technique that involves a great deal of homogeneity. The devices themselves are totally distributed and homogeneous. Further, synthesized networks are homogeneous, in that they involve only a single type of device. Finally, another interesting possibility rests with the use of analog rather than digital junctions. By an analog junction is meant a weakly coupled junction. Thus, for example, with a weakly coupled  $T$  junction, a single pulse may

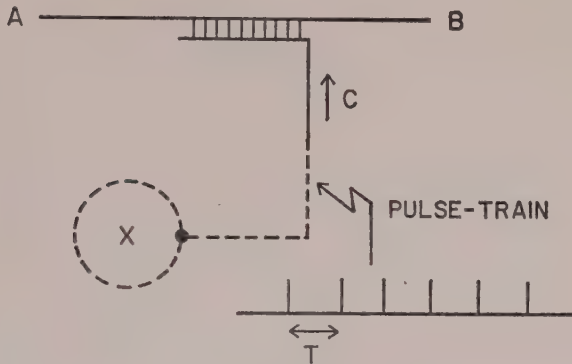


FIG. 8.

not itself be able to stimulate a discharge on the other lines, but only under suitable coincidences of pulses will such discharges be initiated. From what is thus far known, if any analogy at all is in order, we may say that in the human nervous system, digital junctions are essentially non-existent, and that analog junctions are completely dominant. Nevertheless, it is hoped that as a result of this talk, it will be appreciated that digital networks can be synthesized with such devices related only in a digital manner. Whether such an approach will actually become practically useful, depends upon whether the devices themselves can be realized in simple, useful ways. I hope to be able to report at some future conference that they can.



**JOHN R. BOWMAN**

*Technological Institute,  
Northwestern University*

## A NEW TRANSMISSION LINE LEADING TO A SELF-STABILIZING SYSTEM

### PART I

Everyone knows that in a transmission line there is inevitably, so far as we know now, inertia and restoring force. I have had a little fun generalizing this. I will not go into the details, but you can pick out a good many kinds of inertia and a good many kinds of restoring forces. The result is usually interpreted as a frequency, a carrier wave, if you like, even in such a simple case as a physical rod with a restoring spring. If I push on one end, I can transmit a bit to the other, and it comes back. It can be looked on easily as an acoustic wave of infinite wavelength. If I turn one end, the other end responds and we have a torsional rather than a longitudinal acoustic wave of infinite wavelength.

To introduce the specific mechanism I am about to describe, I would like to recall the advice of an old professor who practically held my hand before I faced my first teaching lecture. A student assistant is a very frightened young man; I am sure most people present have been through that experience. He said, "It's very easy. If a student comes to you with a question about theory, work out an example for him, and if he comes stuck on a problem, talk theory."

I think this is good advice, and so to introduce a theory I am going to describe a wax-and-string experiment completed about two years ago. Now the dime stores sell, for about a dime, a magnetic needle compass about the size of a dime. On a hunch, I bought about a pint of them. That was almost the total investment for this research project.

The idea was first to set these nearly touching in a row. The individual needles have a time constant, in pointing somewhere near to the magnetic north pole, of the order of a second. When they are close to one another, however, they interact to an extent that overshadows the field of the earth, and the time constant is of the order of, say, a tenth of a second. Thus they will point north to south, north to south, on down the line (Fig. 1a). The experiment was set up so that north was normal to the axis of the

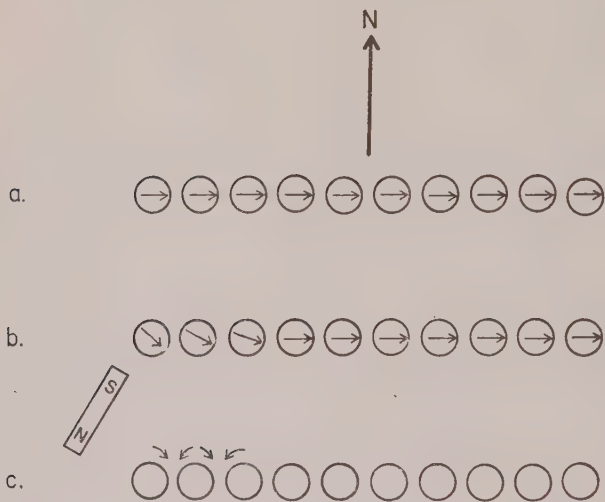


FIG. 1

array, and that gives a very stable sort of array. Now the only additional expenditure for this experiment, other than paper and pencil, was a small bar magnet, and I did not pay for it. I simply borrowed it from the freshman physics laboratory. Now the arrowhead in Fig. 1 indicates north. Bringing up a south pole gives a repulsion that will tend to displace the end needle. You can see, I am sure, that a quasi-static system will result, where we get something as shown in Fig. 1b. The angles of displacement will decrease, so that after the initial impulse a dynamic situation is established and the signal moves along, not too fast,

This is one of the nicest toys I have ever played with. You can bring up the bar magnet slowly and close, and maintain a static situation where equilibrium is propagated, so that the needles assume angles equally. The behavior is an exact analog of a gear train; that is, one turns this way, one that way, and so on (Fig. 1c). It is very much like the bar where you turn one end and observe that the other end turns too. That is not too interesting.

However, if you look upon this as a dynamic rather than a quasi-static system, you can get some extraordinary phenomena that I cannot draw. With a little practice, bringing a south pole up just right, you can make the first compass spin all around and nothing is propagated down the line. The skill in my hand automatically introduces some random numbers, so the experiments were not reproducible. I can tell, nevertheless, of several things that can happen. If you bring the south pole up in a certain fashion, a nice signal goes along, with a complete flip-flop of every needle in the row, and a truly binary, bistable system exists.

On the other hand, if you do not do it in quite the same way, the signal will go down only so far, sometimes apparently even amplified through resonance, and somewhere along the line one of the needles will turn all the way around, and the signal will be reflected and go back again, never getting past a certain point. In other instances—you can run several hundred experiments an hour—you will have a section of several needles that just start spinning in a synchronous fashion until it finally dies out. Eventually it will settle down in one of the two stable states. I repeat, it is one of the nicest toys I have ever played with; total cost under ten dollars.

Now I do not propose this for transmission of information, but I think it is of enough interest to explore as a principle. The mathematical analysis suggests, of course, a very large system of differential equations. You can put about fifty of these in a row on your desk top and you will have fifty dependent angular variables. While you can set down rather easily the differential equations that govern the performance, I think it would slow down even a big IBM machine to get numbers out. Numbers do not give you very much insight anyway as to what really goes on.

Let me generalize this to the case where the number of pivoted dipoles becomes infinite as their size approaches zero, and the



ratio of inertia to dipole moment remains constant. I have been teased a bit about remarks on dimensional analysis here, but I think you can all see quite easily that this is a one-parameter transmission line. The important thing is, of course, the ratio of moment of inertia to the dipole moment. I leave damping out for the moment.

An electrostatic analog is easily conceivable on a molecular scale. Let us have a high-polymer backbone, and introduce an amino acid side group (Fig. 2). Molecularly, that is about the



FIG. 2

biggest dipole moment you can get. Somewhere along the backbone there are going to be other groups just like this one, with free rotation around the bond connecting the group to the backbone. Such a substance could be made. Its dielectric properties ought to be quite remarkable. It would beat, by a factor of perhaps hundreds, any of the barium titanate type of ferroelectrics we have yet found. It would be a storage, or a logical, or a transmission line. It might find use in some of the gadgets that have been talked about here. This is one kind of generalization. I might say that I brought this story first to C. S. Marvel, here at the University of Illinois, and he said that if I really wanted something like it, he would make it, but it would be a terrible job.

## DISCUSSION

ZOFF: I could give you a polymer that I think would satisfy you as an analog to it. It has nicotinic acid side groups as the dipoles.

BOWMAN: It would necessarily be a long linear polymer with these groups spaced at reasonable intervals. The effect might be observable only in solution and perhaps then only under shear, under streaming flow to orient the backbones, but at least there are some possibilities here that I like to dream about.

ZOPF: It would be relatively simple to synthesize such a polymer to have an appropriate average spacing of side groups along the chain; the problem would be to control the lateral interaction between groups on adjacent chains. Polymers like this often show ropy, clotted flow, with rate-dependent elastoviscous properties.

BOWMAN: Now I will give you another generalization. Let us go back to the needles, and this time I am going to leave them out and just draw their housings (Fig. 3). Here is one transmission line, with others put right alongside it. Now you get something really messy.

If you set up an array of this sort, and then poke the thing with a bar magnet, I challenge any IBM machine to compute what will happen. The interactions are now exceedingly complicated. We can make some generalizations: first, if you have an infinite array going out in all directions forever,

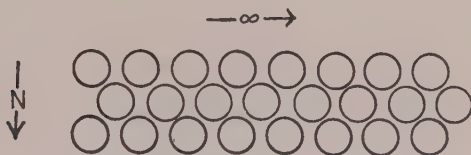


FIG. 3

each needle obviously has as many states as it has closest neighbors in the array, four for a square array, six for a hexagonal array. Hence we have a self-stabilizing, self-organizing system; a suitable perturbation at one point will make it shift to another of its stable states.

If one direction of the array is infinite, and the other direction is  $N$  (Fig. 3), you get a curious dependence on  $N$ . The number of stable states apparently passes through a maximum at about  $N$  equal to fifty to one hundred. This is a very rough estimate. The boundary effect profoundly complicates the infinite case.

Now if you can make a polymer of this type, draw it into fine filaments, nylon style, so that they are oriented, of effectively infinite length so far as the molecular size is concerned, and a few hundred active groups wide, which is the sort of dimensions the textile people speak of, you might get a substance of really extraordinary electrical properties.

ZOPF: In regard to your experiments with the compasses: are you familiar with the paper that Cragg and Temperley<sup>(1)</sup> had in the EEG Journal? They proposed a cortical model by analogy to domain movement in a magnetic material. This was in 1954.

BOWMAN: 1954? They scooped me. They did not take magnets and put them together?

ZOPF: They said this could be done.

BOWMAN: Well, they did not know where the nearest Woolworth store was. It works, and I think it is one of the nicest toys you could play with on a rainy afternoon.

AMAREL: Can you generate any confined effects in a small region of such a system?

BOWMAN: I do not see why not. You could put any constraints you like on the microscopic system or the high-polymer system. I repeat that this is strictly a wax-and-string experiment; bringing a bar magnet up to the end of a linear array of this sort is just not a reproducible perturbation.

ZOFF: In your many-row case, do you find situations where an island of activity moves through the system?

BOWMAN: I once set up a five-by-fifty array on my desk top, and I could not see all that is going on at once; it is just terribly busy.

ZOFF: And a recorder is going to cost more than ten dollars.

BOWMAN: To take this in a really serious way, to have a controlled mechanism that would introduce precise, reproducible perturbations, would run into some big money, but you might get some valuable results from it.

## PART II

### CRYSTALLIZATION

I have tried to give you something, and now I am afraid I have to take something away with an open-ended paper. You all know that a saturated solution of a salt can be supercooled—cooled, in many cases, way below the temperature at which crystallization of the salt should occur. It is then metastable and can be held in many instances at that temperature indefinitely. A sudden jar, a bit of dust, or a seed crystal will induce an apparent self-organizing process that is a pretty spectacular sort of thing. The homogeneous phase separates into two phases, one of which is a crystalline phase, which is just about as organized as matter can get. Here we have a case of perhaps the simplest instance of spontaneous self-organization. It can even be selective.

Figure 4 shows a two-component system, two components that do not form a solid solution. The axes are temperature and the percentage of one component. As you know, we have a eutectic curve separating the liquid and solid regions. Now if we have a liquid of composition  $a$ , at a particular temperature, and cool it down, nothing happens until we hit the eutectic curve, then the process of crystallization sets in and we go down into the solid region. If we hit the singular point of the eutectic curve, neither phase changes in composition. Now if we start at point  $b$  and cool down, one component will crystallize out first, and the composition and temperature will run along the eutectic line until we reach the singularity. That is where we came in in the first case.

Now there is an extension of this that I think I can speak about. A patent was issued recently, wherein a selectivity in the metastable state is going to mean millions of dollars in chemical engineering operations. Suppose we have the eutectic composition and we cool it to point *c* in Fig. 4, assuming it will stay liquid—if we do it very carefully, it *will* stay liquid. If now we add just a speck of either component, that component alone crystallizes

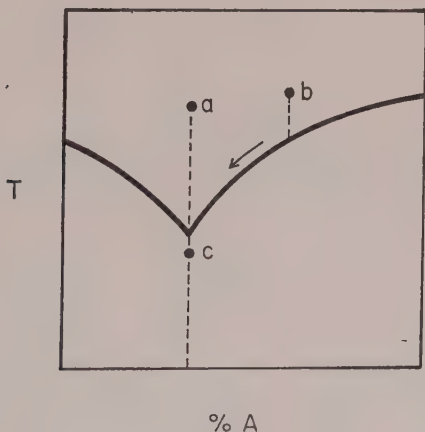


FIG. 4

out. There is apparently a selectivity, provided the crystal forms are sufficiently different, whereby even though you do not have a thermodynamic equilibrium, the kinetics are such that like grows only on like.

Now the main theme of this conference is self-organization. Growth of a crystal, I still believe, is one of the fine examples of self-organization. Let me point out though, that self-organization, at least as far as I know and as far as anyone seems to know that I have talked to here, has no mathematical, quantitative measure. That is why I say this is an open-ended contribution ending on a pessimistic note. Self-organization is certainly not entropy, even though, in the usual conception, we think of low entropy as high ordering, high symmetry, and the second law tells us that Nature, isolated, passes from order to disorder. That is not what we have

been talking about here, if you will agree with me that spontaneous crystallization from a supercooled solution is self-organization.

I leave that question with you, but I should like to repeat a triplet of wisecracks I picked up recently in Alaska. A grand old man who has been teaching chemistry up there for a long time, Bill Wilson, introduces the three laws of thermodynamics with these interpretations: the first law says you cannot win; the second law says you cannot even break even; and the third law says that when it is really cold, you cannot even do business.

#### REFERENCES

1. B. G. CRAGG and H. N. V. TEMPERLEY, The organization of neurones: a co-operative analogy, *EEG Clin. Neurophysiol.* **6**, 84 (1954).



**C. A. ROSEN**

*Stanford Research Institute,  
Menlo Park, California*

## AN APPROACH TO A DISTRIBUTED MEMORY

### INTRODUCTION

Large-scale storage of information presently used in digital computers is in the form of binary bits stored either statically (in magnetic cores, drums or tapes, punched cards, etc.) or dynamically (in the form of pulses recirculated in acoustic delay lines). In these methods each stored bit is associated uniquely, in a one-to-one correspondence, with a binary state of a two-state process, and can in principle always be located precisely in some specified spatial volume or temporal interval. It appears that some forms of biological memory make use of a distributed storage system, that is, one in which information is stored spatially (or perhaps temporally) in a widely dispersed manner. The exact manner in which information is processed, distributed, and stored is not known.

Suppose it is assumed that some body of information to be stored is transformed into binary form, then, in a distributed system, each binary digit will be stored, or leave its trace in more than one site. Read-out of the stored information would then consist of recovering and reconstituting the scattered traces of each binary digit of the complete body of information, preferably in a parallel manner. What properties must the stores have? How should they be interconnected, and what methods of read-in and read-out can be used? Finally, and most importantly: for a given number of binary digits to be stored, what is the minimum number of stores required and how critical to error and failure is such a system? A number of approaches to these problems have been suggested either explicitly or implicitly by Taylor,<sup>(1)</sup> Rosenblatt,<sup>(2)</sup> Uttley,<sup>(3)</sup> Willis,<sup>(4)</sup> and others.

This paper reports initial attempts to devise a distributed memory system which represents one of many possible approaches. It was arrived at by a process of intuition and inference and as yet has no solid mathematical base, but the problems seem tractable, and the system appears capable of extension. This work is now being actively pursued.

### TYPE OF STORES

At the outset, there was no obvious way of making use of binary stores for a distributed memory other than the relatively trivial method of paralleling a sufficient number of such stores for each binary bit to be stored.

The use of a multi-level or analog store appeared to be desirable. This choice was further supported by the fact that a relatively simple and inexpensive analog store was being developed in our laboratories by A. E. Brain\* making use of multi-aperture magnetic cores. In essence, the following functions can be performed:

- (1) At least 25 levels of storage are available within a linearity tolerance of 10 per cent.
- (2) Store levels can be read out non-destructively.
- (3) Each store can be signal-gated to provide a means for absolute inhibition, i.e. the store may be made inactive.
- (4) The stores may be incremented or decremented, such changes being summed algebraically in a simple manner.
- (5) Read-out signals from a number of stores may be summed algebraically to yield a single output.

### INTERCONNECTIONS, METHODS OF STORAGE, AND READ-OUT

With the type of stores and their functional properties described as above, a paper study was initiated, using cut-and-try methods in an effort to devise a structure which could be analysed to answer some of the questions posed originally. Initially the random interconnection and "forced teaching" scheme introduced by

---

\* Valuable support was given by H. D. Crane, D. Bennion and D. Englebart. This work will be reported in the near future,

Rosenblatt was explored. Referring to Fig. 1, which illustrates a portion of a system, the stores are labeled numerically, the outputs are labeled alphabetically, and the numbers opposite each output represent the connections from the stores to that particular output. For simplicity in illustration, each such connection represents two pathways:

- (1) Operating from right to left, increments or decrements of one unit are summed algebraically and are read in to each analog store, which is connected to several output blocks.
- (2) Operating from left to right, the outputs of several stores are summed algebraically at each output block and represent a read-out signal.

Restricting this system to the storage and read-out of binary patterns, a choice of a 1 for a particular input block implies that increments of +1 units will be sent to each store connected to that input. Alternatively, a choice of a 0 causes a decrement of one unit to be imposed on each connected store. In the illustration one of eight possible binary patterns has been chosen (namely, 1 0 0) and the sums of increments and decrements have been entered into each store. The sum of the read-out signals at each store is shown separately (these represent sequential operations).

The simplest manner in which the read-out sums (+1, -2, -4) may be manipulated to yield the original stored or forced pattern, is to compare each output with the algebraic average of all outputs; in this case, the average is  $-5/3$ , the output for *A* is greater than  $-5/3$  and thus becomes a binary 1; the outputs for *B* and *C* are smaller and become binary 0s, thus recovering the original stored binary pattern, 1 0 0. With the same interconnection scheme for the second example shown, with a chosen input pattern of 1 0 1 it can be seen that the same procedure yields an incorrect result, namely 1 0 0. In general, it was found that for arbitrarily chosen interconnection schemes there usually could be found some patterns which could not be read out correctly.

## REGULAR SETS\*

After a considerable number of interconnection schemes were tested, altered, and retested it became evident that, at least for a relatively small number of stores and binary outputs, specific constraints and symmetry conditions had to be imposed on the so-called random wiring if all possible patterns were to be stored and read out correctly. Figure 2 illustrates one scheme (of possibly many) which was found having interesting properties and which was apparently amenable to further orderly development. For convenience this wiring arrangement will be called a regular set.

The following rules have been observed in forming this set:

- (1) Each store and each output block have an equal number of connecting leads. Thus the number of stores equals the number of outputs.
- (2) At each output the numbered store leads form combinations of two; each such possible combination appears at least once and only once. For examples in Store A, the combinations (1, 2), (1, 3), (2, 3) appear; in Store B, there are the combinations (1, 4), (1, 5), and (4, 5), and so on. The identity combinations (1, 1), (2, 2), (3, 3), etc., are excluded. Observing these conditions, the relation between the number of stores,  $N$ , and the number of leads,  $r$ , is shown in Appendix A, Case 1, to be

$$N = r^2 - r + 1. \quad (1)$$

Equation (1) represents a necessary condition to be satisfied, but does not prove that it can be satisfied, or how to go about it.

---

\* It was recently pointed out by W. H. Kautz of SRI that such sets of numbers have had a long history in the field of Combinatorial Analysis. In particular, similar sets are used in the form of Graeco-Latin squares, serving as a means for design of experiments. The set shown in Fig. 2 is one of many sets known as Steiner's triples. See Kaplansky, I. M. Hall, Jr., *et al.*, *Some Aspects of Analysis and Probability* (John Wiley and Sons, New York City (1958)).

Shown in Appendix B are regular sets for  $r = 2, 3, 4,$  and  $5.$

Referring again to Fig. 2, there are shown the results for four different patterns of binary outputs, using the operating procedures outlined previously. By symmetry arguments, the patterns chosen can be shown to represent all the  $2^7$  patterns possible with 7 outputs. With the exception of the unique patterns 1111111 and 0000000, this procedure will yield a correct result in each case. Even the two unique patterns may always be recognized by adding a simple additional procedure. Also of interest are the following properties which have been tested for regular sets up to  $r = 5,$  and by inference may hold for all  $r.$  (The following properties hold for all binary patterns except the two unique patterns noted above.)

- (1) For any chosen binary input pattern, the algebraic sum of read-out signals is the same for all 1s and the same for all 0s. Thus in Trial (b), the algebraic read-out sums are  $-1$  for each output  $A$  and  $B,$  which were both chosen to be 1s, and  $-5$  for each output  $C, D, E, F, G,$  which were all chosen to be 0s.
- (2) The algebraic difference,  $D,$  between a read-out sum at a chosen 1 input and a chosen 0 input is related to the number of leads,  $r,$  by

$$D = 2r - 2$$

Thus, for Trial (a),  $D = (-3) - (-7) = 4$  for outputs  $A$  &  $B$

Trial (b),  $D = (-1) - (-5) = 4$  for outputs  $A$  &  $C$

Trial (c),  $D = (1) - (<3) = 4$  for outputs  $A$  &  $D$

- (3) The maximum number of levels required for each store is  $2r.$  If these levels are not precisely separated by equal increments, one may expect that, as  $r$  increases, positive and negative errors will tend to cancel in the output summation process. On the other hand, as  $r$  increases, the read-out difference  $D,$  between a 1 and 0 increases approximately



as  $2r$ ; it thus appears that accurate discrimination may improve with system size.

- (4) It is to be noted that the number of stores required equal the number of binary outputs. To date, no wiring arrangement has been found for a system using a smaller number of stores than the number of desired binary outputs, which will permit all patterns to be read out correctly, using the simple read-in, read-out procedure outlined above. As shown later, a more complex "teaching" procedure shows promise.

#### A SYMMETRICAL BUT NONREGULAR SET

An example of a symmetrical wiring scheme with less stores than outputs is shown in Fig. 3. In this set the previously stated rule regarding the combinations of two is observed, that is, each doublet is present, and present only once. Identity doublets are again excluded. Again, by symmetry, the binary read-in patterns shown in Trials (a) to (f) inclusive represent all the possible  $2^6$  patterns. If the previous procedure of comparing each output with the average is used, the only patterns which will be read out correctly will be those represented by Trials (b) and (c). These represent a total of 32 read out correctly out of 64, the rest being read out either wrongly or indeterminately.

#### FAILURE OF ONE STORE IN A REGULAR SET

Suppose it is assumed that one store has failed in a regular set. The type of failure chosen is that one store will yield zero read-out signals, whether or not it has been properly incremented or decremented on read-in. This is, incidentally, the equivalent of disconnecting that store completely and thus this test will reveal the effect of using a number of stores smaller than the number of desired outputs.

Referring to Fig. 4, Store 1 has been disabled (or equivalently, is assumed to read out zero). The digital pattern selected for storage is: 0 0 0 0 1 1. In Trial (a), the output sums are  $-2, -2, -2, -5, -5, -1, -1$ . The average is  $-2\frac{4}{7}$  and therefore

using the simple comparison with average scheme, the binary equivalent would be 1110011, obviously quite wrong. With this procedure, it is found that 30 of the 128 possible patterns are incorrect or indeterminate, two of these 30 being the unique patterns 1111111 and 0000000.

### AN ITERATIVE "TEACHING" SCHEME

More powerful methods of iterative read-in and read-out can be devised to reduce the number of incorrect patterns read out of the stores. One such method follows.

The algebraic sums at each output are each compared with a fixed threshold. At all outputs where the absolute value of the sum is less than the threshold, the original incrementing and decrementing procedure is repeated; at those outputs where the sums are equal to or greater than the threshold, no incrementing or decrementing is performed. This procedure will in general change the stored value in most if not all the stores. On read out, the sum of each and every output is again compared with the threshold, and this procedure is repeated until all output sums (in absolute value) exceed or equal the threshold. If this procedure converges, it will be found that those output sums which have positive values represent binary 1s; the negative sums represent binary 0s.

Referring back to Fig. 4, we find that the read-out sums for Trial (1) at outputs *A*, *B*, *C*, *D*, *E*, *F*, *G* were  $-2$ ,  $-2$ ,  $-2$ ,  $-5$ ,  $-5$ ,  $-1$ ,  $-1$  respectively. A threshold of 4 was arbitrarily chosen. The sums for outputs *A*, *B*, *C*, *F*, *G* are less in absolute value than 4, and outputs *D* and *E* are greater than 4. In Trial (2), decrements of 1 were applied to those stores connected to outputs *A*, *B*, and *C*, and increments of 1 applied to stores connected to outputs *F* and *G* since the original read-in binary pattern selected had binary 0s for outputs *A*, *B*, and *C*, and binary 1s for outputs *F* and *G*. The read-out sums in Trial (2) are now  $-2$ ,  $-2$ ,  $-2$ ,  $-6$ ,  $-6$ ,  $0$ ,  $0$ . The outputs *A*, *B*, *C*, *F*, and *G* are still below threshold, and the above incrementing and decrementing process is repeated. This whole procedure is iterated for a total of thirteen trials until all outputs exceed threshold. At the end of Trial (13), the output

sums are  $-4$ ,  $-4$ ,  $-4$ ,  $-16$ ,  $-16$ ,  $4$ ,  $4$ . The negative quantities are interpreted as binary 0s, the positive, as binary 1s, and thus the original binary pattern 000011 is recovered from the stores correctly.

Using this iterative procedure, it can be shown (see Appendix C) that the number of incorrectly read-out binary patterns is 2 out of 128, reduced from 30 out of 128 for the simple uniterated procedure. If an additional step is employed, even these two are resolved and all will be read out correctly. Time has not permitted a complete testing of systems larger than those for  $r = 3$ , or for investigating the effects of more drastic changes in the symmetry of regular sets. These results, however, seem to indicate that with sufficiently sophisticated methods of read-in and read-out, it may be possible to reduce the number of required stores, improve performance if component failure occurs, or both.

#### SUMMARY

Premised on the use of multistate storage elements, there has been introduced a class of wiring schemes for a distributed memory which appears useful and conceptually illuminating. Interconnection schemes, methods of read-in and read-out are described, as applied to small systems; extension to larger systems warrants further investigation.

#### APPENDIX A

##### RELATION BETWEEN NUMBER OF STORES TO INTERCONNECTIONS PER STORE

Number of stores =  $N$

Number of leads per store =  $R$

Number of outputs =  $n$

Number of leads per output =  $r$

*Case I*

$$N = n; \quad R = r.$$

The number of combinations of  $N$  stores taken two at a time [with no identity combinations such as (1, 1), (2, 2), etc.] is  ${}^N C_2$ .

For each output having  $r$  connected leads, the number of combinations of  $r$  leads taken two at a time, is  ${}^r C_2$ . The total number of such combinations is  $n({}^r C_2)$ .

Thus

$${}^N C_2 = n({}^r C_2).$$

But,

$$N = n.$$

Expanding and simplifying,

$$N = r^2 - r + 1.$$

*Case II*

$$N = n.$$

But  $(N)(R) = (n)(r)$  where  $N, R, n, r$  are integers. This follows from the fact that the total number of connecting leads emanating from the stores must equal the total number of leads connected to the outputs, and each lead is assumed non-divisible.

Proceeding as in Case I,

$${}^N C_2 = n({}^r C_2)$$

or

$$N(N-1) = n(r)(r-1).$$

Thus two relations must be satisfied, where all quantities are integers:

$$(1) \quad (N)(R) = (n)(r),$$

$$(2) \quad N(N-1) = n(r)(r-1).$$

Using the trial-and-error technique, the following table lists some of the low-order solutions possible.

<i>r</i>	<i>R</i>	<i>N</i>	<i>n</i>
2	1	2	1
2	2	3	3
2	3	4	6
2	4	5	10
2	5	6	15
etc.			

3	1	3	1
3	3	7	7
3	4	9	12
3	6	13	26
etc.			

4	4	13	13
4	5	16	20
4	8	25	50
4	9	28	63
etc.			

## APPENDIX B

## REGULAR SETS—INTERCONNECTION SCHEMES

$$N = r^2 - r + 1$$

$$r = 2, \quad N = 3$$

Output	Connected to Stores
A	1, 2
B	1, 3
C	2, 3

---

$$r = 3, \quad N = 7$$

Output	Connected to Stores
A	1, 2, 3
B	1, 4, 5
C	1, 6, 7
D	2, 4, 6
E	2, 5, 7
F	3, 4, 7
G	3, 5, 6



$$r = 4, \quad N = 13$$

Output	Connected to stores
A	1, 2, 3, 4
B	1, 5, 6, 7
C	1, 8, 9, 10
D	1, 11, 12, 13
E	2, 5, 8, 11
F	2, 6, 10, 13
G	2, 7, 9, 12
H	3, 5, 9, 13
I	3, 6, 8, 12
J	3, 7, 10, 11
K	4, 5, 10, 12
L	4, 6, 9, 11
M	4, 7, 8, 13

$$r = 5, \quad N = 21$$

Output	Connected to stores
A	1, 2, 3, 4, 5
B	1, 6, 7, 8, 9
C	1, 10, 11, 12, 13
D	1, 14, 15, 16, 17
E	1, 18, 19, 20, 21
F	2, 6, 10, 14, 18
G	2, 7, 11, 15, 19
H	2, 8, 12, 16, 20
I	2, 9, 13, 17, 21
J	3, 6, 11, 16, 21
K	3, 7, 10, 17, 20
L	3, 8, 13, 14, 19
M	3, 9, 12, 15, 18
N	4, 6, 12, 17, 19
O	4, 7, 13, 13, 18
P	4, 8, 10, 15, 21
Q	4, 9, 11, 14, 20
R	5, 6, 13, 15, 20
S	5, 7, 12, 14, 21
T	5, 8, 11, 17, 18
U	5, 9, 10, 16, 19

## APPENDIX C

ITERATIVE "TEACHING" FOR REGULAR SET  
WITH ONE STORE INACTIVE

$$r = 3, N = 7$$

Output or Input	Wiring to Stores
A	1, 2, 3
B	1, 4, 5
C	1, 6, 7
D	2, 4, 6
E	2, 5, 7
F	3, 4, 7
G	3, 5, 6

Assume Store 1 fails, i.e. the output from Store 1 is zero.

The output (or input) blocks can be divided into two groups: the first group, *A, B, C*, all share a connection with the failed Store 1; the second group, *D, E, F, G*, have no connections with Store 1.

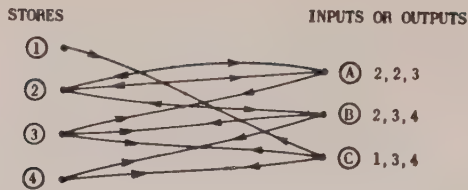
It has been found that due to symmetry the 128 possible digital patterns can be reduced to ten prototype patterns which are tabled below. The right-hand column indicates the combinatorial calculation yielding the number of equivalent patterns for each prototype.

Prototype patterns							Number of equivalent patterns
A	B	C	D	E	F	G	
1	0	0	0	0	0	0	$2({}^3C_1 \times 1) = 6$
1	1	0	0	0	0	0	$2({}^3C_2 \times 1) = 6$
1	1	1	0	0	0	0	$2(1 \times 1) = 2$
1	0	0	0	0	0	1	$2({}^3C_1 \times {}^4C_1) = 24$
1	0	0	0	0	1	1	$2({}^3C_1 \times {}^4C_2) = 36$
1	1	0	0	0	0	1	$2({}^3C_2 \times {}^4C_1) = 24$
0	0	0	0	0	0	1	$2(1 \times {}^4C_1) = 8$
0	0	0	0	0	1	1	$2(1 \times {}^4C_2) = 12$
0	0	0	0	1	1	1	$2(1 \times {}^4C_3) = 8$
0	0	0	0	0	0	0	$2(1 \times 1) = 2$

Total = 128 =  $2^7$  patterns.

Each of these ten prototype patterns has been tested, using the method of iterative "teaching". The following table summarizes the results:

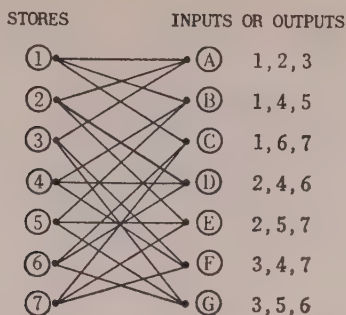
Operation	Output or Input							Input or output	Wiring to stores
	A	B	C	D	E	F	G		
Read-in (Binary)	1	0	0	0	0	0	0	A	1, 2, 3
Read-out 1st Trial	-2	-6	-6	-7	-7	-7	-7	B	1, 4, 5
Read-out 3rd Trial	4	-6	-6	-4	-4	-4	-4	C	1, 6, 7
								D	2, 4, 6
Read-in (Binary)	1	1	0	0	0	0	0	E	2, 5, 7
Read-out 1st Trial	-2	-2	-6	-5	-5	-5	-5	F	3, 4, 7
Read-out 11th Trial	4	4	-18	-5	-5	-5	-5	G	3, 5, 6
Read-in (Binary)	1	1	1	0	0	0	0	Store 1 assumed to have "failed". Threshold level set at 4.	
Read-out 1st Trial	-2	-2	-2	-3	-3	-3	-3		
Read-out 2nd Trial	-4	-4	-4	-6	-6	-6	-6		
Read-in (Binary)	1	0	0	0	0	0	1		
Read-out 1st Trial	0	-4	-4	-5	-5	-5	-1		
Read-out 4th Trial	6	-6	-6	-6	-6	-6	-6		
Read-in (Binary)	1	0	0	0	0	1	1		
Read-out 1st Trial	2	-2	-2	-3	-3	1	1		
Read-out 4th Trial	6	-4	-4	-6	-6	4	4		
Read-in (Binary)	1	1	0	0	0	0	1		
Read-out 1st Trial	0	0	-4	-3	-3	-3	1		
Read-out 8th Trial	5	5	-11	-4	-4	-4	10		
Read-in (Binary)	0	0	0	0	0	0	1		
Read-out 1st Trial	-4	-4	-4	-7	-7	-7	-3		
Read-out 8th Trial	-4	-4	-4	-10	-10	-10	6		
Read-in (Binary)	0	0	0	0	0	1	1		
Read-out 1st Trial	-2	-2	-2	-5	-5	-1	-1		
Read-out 13th Trial	-4	-4	-4	-16	-16	4	4		
Read-in (Binary)	0	0	0	0	1	1	1		
Read-out 1st Trial	0	0	0	-3	1	1	1		
Read-out 33rd Trial	-4	-4	-4	-36	4	4	4		
Read-in (Binary)	0	0	0	0	0	0	0		
Read-out 1st Trial	-6	-6	-6	-9	-9	-9	-9		



	Output or input			Stores no.	$\Sigma$ increments and decrements
	A	B	C		
Read-in (binary pattern)	1	0	0	1	-1
Read-out (algebraic sums)				2	+1
	+1	-2	-4	3	-1
				4	-2
Average output	$= \frac{-5}{3}$				
Read-out (binary)	1	0	0		

	A	B	C	Stores no.	$\Sigma$ increments and decrements
Read-in (binary pattern)	1	0	1	1	+1
Read-out (algebraic sums)				2	+1
	3	2	2	3	+1
				4	0
Average output	$+\frac{7}{3}$				
Read-out (binary)	1	0	0 X		

FIG. 1. The behaviour of part of a storage system with 4 multi-level stores, three binary input channels and three binary outputs.



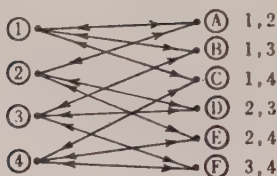
Trial	Operation	Input or output						
		A	B	C	D	E	F	G
(a)	Read-in (binary)	1	0	0	0	0	0	0
	Read-out (sum)	-3	-7	-7	-7	-7	-7	-7
(b)	Read-in (binary)	1	1	0	0	0	0	0
	Read-out (sum)	-1	-1	-5	-5	-5	-5	-5
(c)	Read-in (binary)	1	1	1	0	0	0	0
	Read-out (sum)	+1	+1	+1	-3	-3	-3	-3
(d)	Read-in (binary)	1	1	1	1	1	1	1
	Read-out (sum)	+9	+9	+9	+9	+9	+9	+9

Store no.	Sum of increments and decrements for trial no.			
	(a)	(b)	(c)	(d)
1	-1	+1	+3	+3
2	-1	-1	-1	+3
3	-1	-1	-1	+3
4	-3	-1	-1	+3
5	-3	-1	-1	+3
6	-3	-3	-1	+3
7	-3	-3	-1	+3

FIG. 2. Wiring diagram of distributed memory.



STORES                      INPUTS OR OUTPUTS



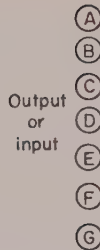
Trial no.	Operation	Input or output						
		A	B	C	D	E	F	
(a)	Read-in (binary) Read-out (sum)	1 -2	0 -4	0 -4	0 -4	0 -4	0 -6	?
(b)	Read-in (binary) Read-out (sum)	1 0	1 0	0 -2	0 -2	0 -4	0 -4	✓
(c)	Read-in (binary) Read-out (sum)	1 +2	1 +2	1 +2	0 -2	0 -2	0 -2	✓
(d)	Read-in (binary) Read-out (sum)	1 -2	0 -2	0 -2	0 -2	0 -2	1 -2	×
(e)	Read-in (binary) Read-out (sum)	1 0	1 +2	0 0	0 0	0 -2	1 0	?
(f)	Read-in (binary) Read-out (sum)	1 +6	1 +6	1 +6	1 +6	1 +6	1 +6	?

Store no.	Sum of increments and decrements for trial no.					
	(a)	(b)	(c)	(d)	(e)	(f)
1	-1	+1	+3	-1	-1	+3
2	-1	-1	-1	-1	-1	+3
3	-3	-1	-1	-1	+3	+3
4	-3	-3	-1	-1	-1	+3

FIG. 3. Symmetrical but not regular set.

Wiring to  
store nos.  
1, 2, 3  
1, 4, 5  
1, 6, 7  
2, 4, 6  
2, 5, 7  
3, 4, 7  
3, 5, 6

Assume Store No. 1 fails; i.e.  
the output from Store 1 is zero



Trial No.	Operation	Input or output						
		A	B	C	D	E	F	G
(1)	Read-in (Binary)	0	0	0	0	0	1	1
	Read-out (Sum)	(-2)	(-2)	(-2)	-5	-5	(-1)	(-1)
(2)	"	(-2)	(-2)	(-2)	-6	-6	(0)	(0)
(3)	"	(-2)	(-2)	(-2)	-7	-7	(1)	(1)
(4)	"	(-2)	(-2)	(-2)	-8	-8	(2)	(2)
(5)	"	(-2)	(-2)	(-2)	-9	-9	(3)	(3)
(6)	"	(-2)	(-2)	(-2)	-10	-10	4	4
(7)	"	-4	-4	-4	-13	-13	(1)	(1)
(8)	"	(-2)	(-2)	(-2)	-11	-11	5	5
(9)	"	-4	-4	-4	-14	-14	(2)	(2)
(10)	"	(-2)	(-2)	(-2)	-12	-12	6	6
(11)	"	-4	-4	-4	-15	-15	(3)	(3)
(12)	"	(-2)	(-2)	(-2)	-13	-13	7	7
(13)	"	-4	-4	-4	-16	16	4	4

Store no.	Sum of increments and decrements for trial no.												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	-3	-4	-5	-6	-7	-8	-9	-9	-10	-10	-11	-11	-12
3	1	2	3	4	5	6	5	7	6	8	7	9	8
4	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-2	-1	-2
5	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-2	-1	-2
6	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-2	-1	-2
7	-1	-1	-1	-1	-1	-1	-2	-1	-2	-1	-2	-1	-2

FIG. 4. Failure of one store—regular set.

## REFERENCES

1. W. K. TAYLOR, Pattern recognition by means of automatic analog apparatus *Proc. IEE (London) Pt. B* **106** (6), 198 (March, 1959).
2. F. ROSENBLATT, The perceptron—a theory of statistical separability in cognitive systems. *Cornell Aeronautical Laboratory Report VG-1196-G-1* (1958).
3. A. M. UTTLEY, Conditional probability computing in a nervous system. *Mechanization of Thought Processes*, (N.P.L. Symposium No. 10) H.M.S.O., London (1959).
4. D. G. WILLIS, Plastic neurons as memory elements. *Lockheed Missile Systems Division—48432, International Conference on Information Processing, Paris* (June, 1959).

**SAUL AMAREL**

*RCA Laboratories, Princeton, N.J.*

## AN APPROACH TO AUTOMATIC THEORY FORMATION

### INTRODUCTION

I would like to discuss here an approach and certain preliminary results of an investigation on automatic theory formation processes. I am assuming that such processes are embedded in information processing systems that “learn” through direct interaction with their task environment, and are capable of evolving their own procedures for executing required tasks. If a man would be assigned the learning function in the information processing system, he most likely would attempt to develop a theory, a “mental model” about the work environment, on basis of the limited amount of information that he would gather, relevant to that environment; this would provide him with an economical and versatile way of storing and using his experience. I am proposing that the mechanized learning scheme take a similar approach; if it does, then the learning mechanism is equivalent to an automatic theory formation process.

A “theory” formed inside a machine is an information structure that efficiently encodes the limited sample of information relevant to a certain class of the environment, and that can be used for reliable predictions over that class. The machine theory of a class represents in machine language an *invariant* property shared by all members of the class; it can be regarded as the machine meaning of the class as a distinct entity. The process of automatically *forming* a theory about a class of the environment is a self-organizing cognitive task; it is equivalent to the process of seeking a suitable design for a pattern recognizer over the class. This general problem has been mostly considered by different

investigators in connection with automatic recognition of sensory patterns; however, attention in the areas of language patterns<sup>(1)</sup> and mathematical patterns<sup>(2)</sup> is recently growing.

I believe that processes of self-organization and learning can be studied more advantageously at the present state of knowledge in the context of relatively simple and familiar formal environments; this way it is possible to concentrate attention on the properties of the self-organization schemes, under controlled conditions. Specifically, theoretical problems of *realizability* can be handled with relative ease; this way one can know *a priori* what is feasible in principle, and thus prevent long sterile explorations. This is the reason for choosing a simple area in mathematical logic as the cognitive environment for my present investigation.

I think that *research methodology* is one of the most significant problems in the study of artificial self-organizing systems. These systems have to be first postulated and then they have to be explored experimentally and theoretically. The exploration of a postulated system is not substantially different in character than the study of an organism by a natural scientist. It is important to carry on significant experimentation with different types of artificial organisms, so that (a) given a partial knowledge of expected tasks it would be possible to select such an artificial organism that is likely to realize them, and (b) given an artificial organism it would be possible to have *a priori* knowledge of what is a feasible task for that organism. I think that we have to go into an area which I might call *experimental engineering* so that we might gain a fundamental understanding of the types of behaviors that can be extracted from different structures. There is no doubt that whatever theoretical progress can be made in this area of characterizing self-organizing systems will be significant in many ways; especially it can prove a welcome guide in the *postulation* of self-organizing systems. The freedom of postulating different artificial organisms distinguishes the engineer from the natural scientist. However, since evaluation of a given postulation is far from simple, it becomes vital to have heuristic guides for constraining postulation within some narrow, reasonable, limits. It is natural that most heuristic guides be based on analogies from biological organisms. Yet, it is important to maintain a certain degree of freedom within the guiding analogies.



The major aim of our investigation is to study how provisional machine structures or hypotheses—that compete for assuming the status of working theory—are *created* in the machine, and how they evolve (till extinction or stabilization) during the cognitive processes. We reserve to ourselves the freedom to postulate and then evaluate various machine processes for theory formation, taking into consideration only certain general premises and assumptions, some of which are induced from psychological analogies. While I will cover later in more detail certain of these assumptions, I am giving next some of the general premises.

The form of any hypothesis capable of emerging in a machine at any time depends on the language available to the machine at that time—primitive and compound symbols and processes. However, economy constraints within the machine (limitation on available processing time and complexity) cause the actual number of hypotheses generated and evaluated at a certain stage of processing to be a small subset of the set of all hypotheses that can be formulated in the language available to the machine. The specific choice of the hypothesis subset that is actually generated in a machine depends on heuristic rules and generating tendencies that are *determined by previous system activity*. A certain stage of language development in the machine, together with a set of hypothesis generation tendencies determine its state, or “stage of development”. Different sequences of problems handled by the machine result in different states; the ability of the machine to theorize on a certain class depends on its “intellectual history”.

In the course of previous work I have experimented with several theory formation schemes, over the testing ground of the simple logic environment (to be described in detail later). At present the investigation centers on a cognitive system  $M$ . In the following I will describe the organization and operation of  $M$ , as well as the concepts and research problems that come out from the investigation of  $M$  to date.

#### GENERAL NATURE OF THE INPUTS TO THE SYSTEM $M$

We assume that a system of type  $M$  will admit input information in the form of pairs ( $I:O$ )—called the *input items*. The  $I$  of an input item represents an  $n$ -tuple of elements,  $I = (x_1, x_2, \dots, x_n)$ , and

the  $O$  represents a *corresponding*  $m$ -tuple of elements,

$$O = (y_1, y_2, \dots, y_m)$$

the elements of  $I$  and  $O$  are taken from finite sets (alphabets) that are familiar to  $M$ . In every practical situation, there is a certain "range of interest" for the  $n$ -tuplets  $I$ , to which there corresponds a range of  $m$ -tuplets  $O$ . In physiological terms,  $I$  and  $O$  represent sensory stimuli or elementary concepts in terms of which a situation comes to the attention of man.

We further assume that a fundamental *supposition* has already been made, to the effect that certain collections of input items belong to certain distinct classes, arbitrarily named  $D_1, D_2$ , etc., the supposition can be regarded as an exploratory one, that  $M$  is asked to confirm. The system  $M$  will attempt to justify this supposition through the generation of appropriate distinct *class-belongedness rules* for each class.

The initial assignment of collections of input items to classes could have been made in  $M$  at a previous stage of operation, or it could have been produced in  $M$ 's environment. Similarly, the input items could have been either results of previous processes in  $M$ , or direct observations from the environment. For our purpose here, I assume that input items and class assignments come from the environment of  $M$ .

The form of an *input message* to  $M$ , during theory formation will be as follows:

$$D:(I:O)$$

indicating that the input item  $(I:O)$  belongs to the class  $D$ .

The set of input messages:

$$[D_j:(I_1:O_1), D_j:(I_2:O_2), \dots, D_j:(I_n:O_n)]$$

where the "range of interest" of  $I$  is defined by the finite set  $[I_1, I_2, \dots, I_n]$ , provides an extensive definition of the class  $D_j$ .

I would like to illustrate by the following examples the type of messages that might appear at the input of an  $M$ -type machine: (a) Correspondences between input data and output data that specify a certain desired *process on data*. (b) Examples of *problem statements and their solution* such as: theorem and proof, logical expression and simplest equivalent logical expression, a game of

strategy and a set of moves to win. (c) Instances of morphological and syntactical *analysis of language* and instances of *sentence translation* between two languages. In the case of morphological analysis, the class  $D$  would include instances of correspondences between stems and modified stems that exemplify a certain morphological rule. (d) Associated sets of observables in physical phenomena such as: observations on cloud cover related to observations on other weather parameters, observations on behavioral relations of elementary particles in a nuclear experiment.

### SPECIFIC CHOICE OF INPUTS USED IN THE PRESENT STUDY

We have used for our study of  $M$  input messages taken from a simple area in mathematical logic. Specifically, the input items that we have fed to  $M$  are instances from sixteen types of mappings between propositional functions of two arguments. For the purpose of a more explicit discussion of the processes in  $M$ , I find it necessary briefly to review certain definitions and concepts involved in the formulation of our input messages.

We express a propositional function of two binary arguments,  $a, b$ , by  $s$ . This is a mapping from the set  $[1, 0]^2$  into the set  $[1, 0]$ . There are sixteen such mappings or propositional functions, forming a set  $\sigma$ .

$$\sigma = [s_1, s_2, \dots, s_{16}]. \quad (1)$$

(For correspondence between our notation and other notations, see Fig. 1.) We express the computation of a function  $s_i$  of  $a, b$  as:

$$s_i ab = c. \quad (2)$$

(We are using here the Lukasiewicz parenthesisless notation.) As an example, the propositional expression for conjunction,  $a$  and  $b$ , corresponds with our  $s_2 ab$ .

Given then  $s_i ab = c$  and  $s_j ab = d$ . We further form  $s_k cd = e$ ;  $c, d, e$  take values from the set  $[1, 0]$  while  $s_i, s_j$  and  $s_k$  take values from  $\sigma$ . We can write then:

$$s_k cd = s_k s_i a b s_j a b = e. \quad (3)$$

We consider next a function  $s_1$ , such that  $s_1ab = e$ , where  $s_1$  is also from  $\sigma$  and we write the functional equation:

$$s_ks_1abs_jab = s_1ab. \quad (4)$$

This equation expresses the functional equivalence between a compound triplet of propositional functions and a single propositional function; this is illustrated in Fig. 1.

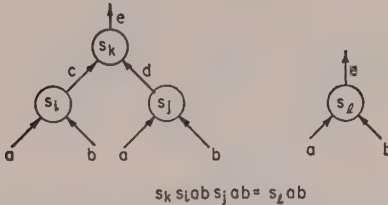
Our notation: Set $\sigma$	Expressions in propositional calculus	McCulloch-Pierce's notation (see Ref. (4))
$s_1$	$F$	$\times$
$s_2$	$a \& b$	$\times \cdot$
$s_3$	$\bar{a} \& b$	$\times \cdot$
$s_4$	$\bar{a} \& \bar{b}$	$\times \cdot$
$s_5$	$a \& \bar{b}$	$\cdot \times$
$s_6$	$a \neq b$	$\cdot \times$
$s_7$	$a$	$\cdot \times$
$s_8$	$b$	$\cdot \times$
$s_9$	$\bar{b}$	$\cdot \times$
$s_{10}$	$\bar{a}$	$\cdot \times$
$s_{11}$	$a \equiv b$	$\cdot \times$
$s_{12}$	$a \supset b$	$\cdot \times$
$s_{13}$	$a \vee b$	$\cdot \times$
$s_{14}$	$a \subset b$	$\cdot \times$
$s_{15}$	$\bar{a} \vee \bar{b}$	$\cdot \times$
$s_{16}$	$T$	$\cdot \times$

FIG. 1. Corresponding notations for propositional functions of two arguments.

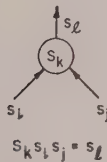
We now make a distinction between functions of  $a, b$  and functions of functions. If we represent the latter by capital letters and we omit the arguments  $a, b$  from the former, we can write (4) as follows:

$$S_k s_i s_j = s_1. \tag{5}$$

In analogy with expression (2), we regard (5) as expressing the computation of a function  $S_k$  of two arguments that have values



(a) GRAPHIC REPRESENTATION OF A TRIPLET OF PROPOSITIONAL FUNCTIONS AND AN EQUIVALENT SINGLE PROPOSITIONAL FUNCTION



(b) DESCRIPTION OF THE EQUIVALENCE IN (a) ABOVE AS A SINGLE MAPPING OF PROPOSITIONAL FUNCTIONS

FIG. 2

$s_i, s_j$ ; the computation results in a value  $s_1$ . Thus, we consider  $S_k$  as a mapping from the set  $\sigma^2$  of  $16^2$  elements, into the set  $\sigma$  of 16 elements, and express this as follows:

$$S_k \alpha \beta = \gamma. \tag{6}$$

The 2-tuplet  $\alpha\beta$  takes values from  $\sigma^2$  and  $\gamma$  takes values from  $\sigma$ .<sup>(6)</sup> There are sixteen such mappings, forming a set  $\Sigma$ .

$$\Sigma = [S_1, S_2, \dots, S_{16}]. \tag{7}$$

Each mapping  $S_k$  determines a class, called  $D_k$ , whose members



are instances of the mapping. A general member of  $D_k$  can be expressed in the form:  $((\alpha, \beta):\gamma)$ .

On the basis of my previous definitions I can now write an expression for the form of the input messages that we are feeding to  $M$  for theory-formation:

$$D_k:((\alpha, \beta):\gamma). \quad (8)$$

$D_k$  stands for any of the classes  $D_1, D_2, \dots, D_{16}$ ; these correspond with the sixteen mappings represented by the elements of the set  $\Sigma$ . Comparing (8) with the previously discussed general form  $D:(I:O)$ , we see that in our present case we have:

$$I = (\alpha, \beta), \quad O = \gamma.$$

There are several reasons for making this particular choice of input messages for our investigation. First, the messages come from a well-known formal area and they present a relatively simple problem to  $M$ . Second, in our exploration of  $M$ 's previous "intellectual" history we have assumed a development paralleling the path of mental development of children proposed by Piaget.<sup>(3)</sup> Specifically, we have explored a development of  $M$  that starts with the classification of concrete objects into sets, goes to the manipulation of sets, and then moves to the formation of relations between sets and the manipulation of related sets. As suggested from Piaget's theory of mental development, the next interval of development in  $M$  can lead to the formation of a structure that represents the propositional calculus: we are assigning then to  $M$  a cognitive task that can lead from a certain assumed initial capacity (sets, relations) to an internal structure of relationships between propositional functions. I do not intend to go at present in more detail into the history of  $M$ , since it has only heuristic value that concerns the choice of a realizable task assigned to  $M$  and the choice of  $M$ 's state at the time it starts working on the task.

Another reason for choosing this particular area of logic as a task for  $M$  was our current interest in functional equations of the propositional calculus that emerged from our study of the stable switching circuits that were introduced by McCulloch.<sup>(4)</sup>

THE COGNITIVE TASK ASSIGNED TO  
M. OVERALL OPERATION

A geometric representation of a mapping  $S_k$  is shown in Fig. 3. Each input item of  $D_k$  corresponds with a directed line segment in the mapping; a specific pattern of  $16^2$  directed line segments specifies  $D_k$ . It is interesting to note that the total number of possible distinct line patterns, or mappings, is  $16^{16^2} = 2^{1024}$ .

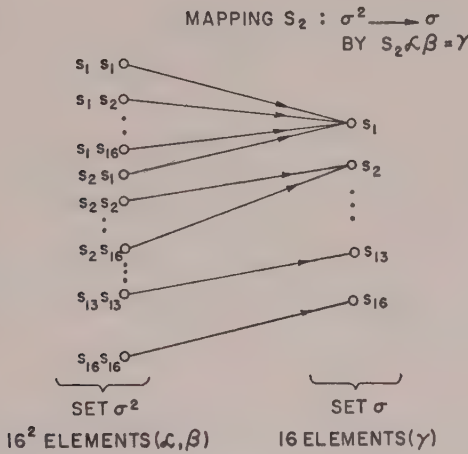


FIG. 3

From this enormous number of possible patterns, sixteen patterns, corresponding to the classes  $D_1, D_2, \dots, D_{16}$ , come gradually to the attention of  $M$ , in the form of sequences of mapping instances—the input items. The general objective of the cognitive activity in  $M$  is to form its own representation for each one of these sixteen distinct patterns, or in other words to attain a “machine theory” for each one of the sixteen classes of items. We regard this theory as an  $M$ -explanation of each class, achieved by  $M$  through an active search for order in the class; here again “order” in terms of constructs available in  $M$ .

In analogy with our expectations from human theories, a “machine theory” should provide predictive power and it should

provide perceptive power for future situations in addition to satisfying immediate cognitive drive.

An important question is: At what stage of the cognitive process should a machine theory appear? After the admission of the total evidence on a class, or after consideration of a limited evidence? I feel that machine theories of various degrees of tentativeness should be generated in  $M$  for any amount of evidence available; however, they should be accompanied by an index of confidence that can be used in decisions involving their application to specific situations. Continuous theory construction provides a more efficient way of summarizing-storing quantities of raw information and in addition it offers, in my opinion, a convenient scheme for the dynamic evolution of a mechanized theory formation process.

Note also that in most cases one cannot have complete evidence; indeed it is hard to think of any real life case where a complete specification of a complex situation is available at the time a decision is needed.

In accordance with our previous discussion on the desired cognitive performance of  $M$ , we have assigned to  $M$  the following general first-level goal:

### Goal I

(a) Generate a machine representation for a class  $D_k$  in the form of a *procedure*\* that prescribes how to produce for all valuations of  $(\alpha, \beta)$  from the "range of interest"  $\sigma^2$  the correct values of  $\gamma$ . Thus,  $M$  should attempt to generate a procedure  $h_{D_k}$  for the class  $D_k$  such that:

$$h_{D_k}(\alpha, \beta) = \gamma \quad (9)$$

for all  $((\alpha, \beta): \gamma) \in D_k$ . (The previous expression indicates that the processing of any inputs  $(\alpha, \beta)$  according to the procedure  $h_{D_k}$  produces the desired corresponding  $\gamma$ .)

(b) Attain the *simplest* possible  $h_{D_k}$ .

We regard  $h_{D_k}$  as a *machine-theory* over the items of the class  $D_k$ . Since in no practical situation all the members of  $D_k$  will be

---

\* A machine procedure corresponds to an algorithm that could be realized either by a computer program or by a network design.

available, the power of a given  $h_{D_k}$  can be tested only over that part of the membership of  $D_k$  that  $M$  has admitted. Therefore a decision to the effect that a certain  $h_{D_k}$  satisfies the Goal I can only be made on basis of an *induction*, based on the existing evidence of performance of the procedure  $h_{D_k}$ . Clearly, it is desirable to construct in  $M$  "hypotheses"  $h_{D_{k_p}}$  that "justify" (in the sense of expression (10)) almost all (preferably all) the admitted input items.

We will clarify later our interpretation of simplicity. Generally, simplicity of  $h_{D_k}$  implies, in machine terms, a faster and more economic processing of *future problems* in areas involving  $D_k$ ; as a corollary, a simple  $h_{D_k}$  has a better chance of becoming a useful building block in theories of a higher level.

Together with Goal I, we assign to  $M$  a *limit of maximum processing power* that represents the maximum energy that  $M$  is permitted to dissipate in its activity toward goal attainment.

While limitation of processing power is concerned with economy *at present*, simplicity of machine theory formed in  $M$  is concerned with *economy in the future*. It is interesting to study how different balances between these two factors affect the processes of  $M$ . However, since in most practical situations the factor of "economy at present" is better known and, more appreciated than the hard-to-evaluate "economy in the future", it is reasonable to consider it as a fixed restriction for the construction of theory formation schemes.

An additional argument concerning limitation of processing effort comes from the basic problem of realizability. There is no guarantee that the task set by Goal I, for a given arbitrary class (with members that are instances of a mapping from  $\sigma^2$  to  $\sigma$ ), will be realizable by a certain  $M$ , within any length of time. Furthermore, most of the times we cannot tell *a priori* whether a given cognitive task is  $M$ -realizable (unless there is a basic theoretical understanding of the system), and the only reasonable operational procedure we have is to let  $M$  try to attain the assigned Goal I, under a limitation of maximum processing effort. If  $M$  fails we still do not know whether this is due to the fact that the solution is not  $M$ -realizable or whether  $M$  is unable to find a solution that is  $M$ -realizable in principle.

We assume that  $M$  will consider one by one all the classes in

its environment, taken in a certain order. The *order of treating the classes* has great significance on both the efficiency of theory formation and the form of the machine theories attained; this is due to the nature of the cognitive process of  $M$ , where procedures attained in the course of the cognitive activity with previous classes are used in the construction of procedures for classes that are subsequently treated. The control of the order in which material from different classes is presented to  $M$  has some analogy with the devising of programs of instruction for people—a task which is known to have great effect on human cognition and learning. Our plan is to examine the theory forming process of  $M$  for different sequences of class presentations. We have experimented so far with sequences starting with  $D_2$ ,  $D_{13}$ , as these classes correspond respectively with the familiar propositional function of conjunction and disjunction.

Information about a class  $D_k$  comes to  $M$  through input messages, where individual members of  $D_k$  (instances of the mapping  $S_k$ ) are given. We assume that  $M$  does not have control over the order in which it admits input items from  $D_k$ . As long as Goal I is not satisfied for a certain class  $D_k$ , and  $M$  did not exceed the maximum level of processing effort that it can dissipate on  $D_k$ ,  $M$  will keep requesting instances of  $D_k$  from its environment.

For each input item of  $D_k$  admitted, say a pair  $((s_i, s_j):s_1)$ ,  $M$  has the following second-level goal:

### Goal II

Generate a machine procedure  $h_{D_{k_p}}$  such that:

$$h_{D_{k_p}}(s_i, s_j) = s_1. \quad (10)$$

Note that there are  $16^{(16^2-1)}$  mappings that satisfy the single correspondence pair  $((s_i, s_j):s_1)$ , and it is possible that a sizable fraction of these mappings is  $M$ -realizable, meaning that  $M$  can attain in principle machine procedures for them. The problem is to find among those procedures the one that satisfies as many other input pairs as possible from  $D_k$ ; preferably, all the other input pairs fed to  $M$ . The process of attaining such a machine procedure without complete exhaustion of alternatives (over a space, say, that is constrained by a maximum “complexity” of procedures) is at the crux of any automatic theory formation



problem. I will outline later one of the approaches that we are taking in treating this problem, after having completed the general description of the operation in  $M$ .

$M$  will actively seek a procedure  $h_{D_{k_p}}$  for each input item, within a certain limitation of processing effort. Only if  $h_{D_{k_p}}$  is attained or if the allowed processing effort is exceeded, will  $M$ 's cognitive activity over the input item stop, and a new instance from  $D_k$  will be admitted.

We are orienting the cognitive activity of  $M$  in such a way that the "justification" of a new input item from  $D_k$  will be first attempted by machine procedures that were attained from previous activity with  $D_k$ ; if this fails, a new machine procedure "justifying" the input item will be constructed on basis of other procedures, elementary and/or compound, that are available to  $M$  at that time. There is an analogy between these processes in  $M$  and human cognition processes, where a new experience has to be "constructed mentally" in terms of existing concepts, or transformations of existing concepts, so that it can be absorbed.

In the course of its cognitive activity over items from  $D_k$ ,  $M$  gradually organizes assemblies of machine procedures in a special type of associative memory,  $\mu$  (to be described later in some detail). The procedures are stored in  $\mu$ , broken down in their elementary parts; the parts are coupled through the intermediary of associative links that are different for different  $D_k$ 's. In addition to the structural aspects of the information that  $\mu$  stores about the  $D_k$ 's,  $\mu$  also stores numbers specifying *strengths of association* between parts of stored procedures; these numbers depend on the degree of success of  $M$  toward attaining a consistent machine procedure "explaining" a class. From the numbers for strength of association,  $M$  derives *confidence numbers*,  $c(h_{D_{k_p}})$ , for each generated overall procedure (or, tentative hypothesis). The confidence numbers provide a basis for making inductions on  $h_{D_k}$ .

If, as a result of a successful cognitive process,  $M$  will consistently offer  $h_{D_{k_p}}$  as a procedure that "explains" long sequences of instances from  $D_k$ , and if as a result the confidence  $c(h_{D_{k_p}})$  of the induction

$$[h_{D_k} = h_{D_{k_p}}]$$

exceeds a certain predetermined level, then  $M$  will consider its

Goal I attained, for the class  $D_k$ . The theoretical considerations on the validity of the induction are in the domain of inductive logic.<sup>(5)</sup> In this domain many problems remain open and controversial; I think however that at the present state of knowledge on mechanizations of theory formation much work of invention, experimentation, and analysis remains yet to be done before being forced to consider critically those problems.

The cognitive activity of  $M$  over a class  $D_k$  stops if Goal I is attained for  $D_k$ , or if the limit of processing power assigned to the theory formation over  $D_k$  is exceeded, or if no more instances of the  $D_k$  can be admitted. After the activity over  $D_k$  stops,  $M$  shifts its attention to another class, and if no other class exists in the environment of  $M$ , the cognitive activity stops.

#### PREDICTIVE AND CATEGORIZATION TASKS BASED ON THE MACHINE THEORIES ATTAINED IN $M$

After having attained a procedure  $h_{D_k}$  satisfying the Goal I,  $M$  can use it directly in the performance of either predictive or categorization tasks.

*Predictive tasks.* The request for such a task is expressed by an input message in the form:

$$D_k:((\alpha, \beta):?)$$

which means: given a valuation of  $\alpha, \beta$  from  $D_k$ , find the value of  $\gamma$ . In response to this input message,  $M$  uses the procedure  $h_{D_k}$  in order to execute the process:

$$h_{D_k}(\alpha, \beta) = \gamma$$

which produces the required value of  $\gamma$ .

*Categorization tasks.* The form of the input message requesting such a task will be:

$$?:((\alpha, \beta):\gamma)$$

which means: given an input item  $(\alpha, \beta, \gamma)$ , find to what class, among the classes having machine theories (or tentative hypotheses), the input item belongs. In response to this input message,  $M$  performs successive processes:

$$h_{D_1}(\alpha, \beta) = \gamma_1$$

$$h_{D_2}(\alpha, \beta) = \gamma_2$$

for all  $h_{D_k}$  in  $M$ . Comparing the  $\gamma$  of the input message with the internally constructed  $\gamma$ 's,  $M$  obtains a coincidence, say  $\gamma = \gamma_t$ . From this, it deduces that the input item belongs to  $D_t$ , or, in machine terms, that it obeys the rule  $h_{D_t}$ . It is possible that more than a single coincidence occurs; in other words, the particular instance given is common to more-than-one mappings, represented by the  $h_{D_1}, h_{D_2}, \dots$  in the machine. This leads to ambiguity of categorization, which is usually undesirable. However, the burden of avoiding these occurrences rests on the choice of suitable disjunct mappings and not on the choice of a theory formation process; the latter attempts to mechanize specified mappings.

Clearly, the above schemes for performing predictive and categorization tasks can be applied to a wide variety of different areas. The scheme for carrying out the categorization task may prove to be an especially powerful one for problems of automatic *pattern recognition* and of automatic *linguistic analysis*. Halle<sup>(6)</sup> has already applied a scheme of this type to the morphological analysis of language; he called the scheme "*analysis by synthesis*". In Halle's approach the list of relevant stems in a linguistic form corresponds to our  $I$ , the linguistic form itself to our  $O$  and the various rules of flecional morphology to our various stored  $h_{D_i}$ 's. To analyse a particular form  $O$ , i.e. to determine the special meaning carried by the flecional state of the form, the various rules  $h_{D_i}$  are applied to the  $I$  till the  $O$  is obtained. The procedure  $h_{D_i}$  whereby the form was generated is then the desired analysis. If there is more than one procedure for generating the particular form, the analysis is ambiguous. To avoid ambiguous analysis, it might be necessary to modify the way in which a set  $I$  is chosen in association with a form  $O$ , before starting analysis; new procedure rules  $h_{D_i}$  have to be developed then, either by a linguist or by a cognitive machine process of the type explored with our  $M$ .

In addition to the "analysis by synthesis" method of recognizing patterns (which might prove promising in other than linguistic areas), mechanical pattern recognition can also be performed in the form of the predictive task; in fact, most investigations in that area have taken that direction. Here, a class  $D_j$  is extensively defined by correspondent pairs ( $I:O$ ), where  $I = (x_1, x_2, \dots, x_n)$  is an  $n$ -tuple of measurements on a sensory event, and its associated

$O \doteq y$  is a binary variable indicating whether this sensory event is an exemplar of a certain pattern  $P_j$  or not. If a procedure  $h_D$ , is attained somehow (say, through a cognition process in an  $M$ -like machine), then  $h_{D_j}$  is a formula for a pattern recognizer that, given a certain  $I$ , will produce a decision (binary value of  $y$ ) on whether  $I$  is an exemplar of  $P_j$  or not. One of the common network realizations of an  $h_{D_j}$  used in such schemes of pattern recognition is the threshold switching cell or formal neuron; in this case, the general process

$$h_{D_j}(I) = 0$$

takes the explicit form:

If

$$\sum_{r=0}^n a_{jr}x_r \geq 0, \text{ then } y = 1$$

If

$$\sum_{r=0}^n a_{jr}x_r \leq 0, \text{ then } y = -1$$

where  $x_0 = 1$  and  $x_1, x_2, \dots, x_n$  are input variables (sensory inputs) taking value 1 or  $-1$ . The elements of the  $(n+1)$ -tuple  $(a_{j0}, a_{j1}, \dots, a_{jn})$  represent numbers that are adjusted to suitable values for each specific machine procedure that corresponds to a class  $D_j$ .

Note that in both modes of pattern recognition, the "analysis by synthesis" mode (where  $h_D$  is used to construct the input form from some of its elements) and the "predictive" mode (where  $h_D$  is used to compute directly from measurements on the input form whether the form is an instance of a certain pattern or not), decisions are made on basis of an appropriate machine theory of a class  $D$ , i.e. on basis of a procedure  $h_D$  with a certain associated confidence number.

A schematic representation of the overall theory formation process in  $M$  is shown in Fig. 4. The inner feedback loop,  $F_1$ , shown in that figure, produces procedures that establish a "path" (in the language of  $M$ ) between the  $I$  complex and the  $O$  complex in the input items. The loop,  $F_2$ , produces new schemes for combining procedures (so that the synthesis capability is extended) and stores the new schemes in the associative memory  $\mu$ . The

outer loop evaluated the produced procedures and maintains production on by continuously feeding problems (the input items) to the inner loops, till either the cognitive goal is satisfied or the effort spent on it exceeds the allowable level.

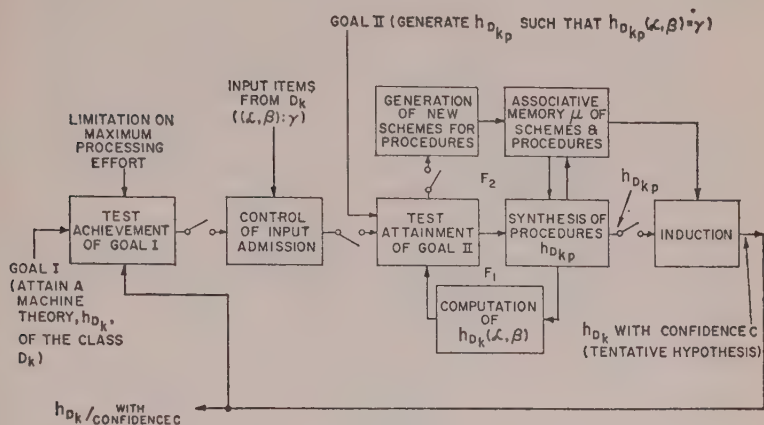


FIG. 4

### A MECHANISM FOR FORMING AND CONTROLLING MACHINE PROCEDURES (HYPOTHESES)

Let us consider a mechanism (now under study) that governs the generation of successive procedures in  $M$ , in response to input items admitted from the classes  $D_1, D_2, \dots, D_{16}$ .

We assume that  $M$  possesses a store of elementary procedures at the start of the present cognitive activity. These initial procedures resulted from assumed previous activity of  $M$  on sets and relations; they are:

(a) Procedures for realizing *combinatorial operation* on subsets of  $\sigma$ . Specifically:  $\cap, \cup, C$ .

We represent the execution of these procedures as follows:

$\cap (A_1, A_2, \dots, A_n) = B_1$ ; for the operation of obtaining the *intersection* set  $B_1$  of the sets

$$A_1, A_2, \dots, A_n.$$

$\cup (A_1, A_2, \dots, A_n) = B_2$ ; for the operation of obtaining the *union* set  $B_2$  of the sets  $A_1, A_2, \dots, A_n$ .



$C(A_1, A_2) = B_3$ ; for the operation of obtaining the complementary set  $B_3$  of  $A_2$  with respect to  $A_1$ .

and  $(A_1, A_2, \dots, A_n, B_1, B_2, B_3)$  represent subsets of  $\sigma$ .

(b) Procedures for realizing *mappings from the set  $\sigma$  into the set  $2^\sigma$* ;  $2^\sigma$  stands for the set of all subsets of  $\sigma$ . Each procedure mechanizes a correspondence ( $\epsilon: A$ ) between elements of  $\sigma$  (denoted by  $\epsilon$ ) and elements of  $2^\sigma$  (that are sets denoted by  $A$ , and  $A \subset \sigma$ ). A set  $A$  might have a single element; we represent  $A$  then by  $[\zeta]$ , where  $\zeta$  is the name of the element from  $\sigma$  that forms the point set. This is an interesting case and  $M$  has two elementary procedures,  $I$ ,  $N$ , that have this property.

The procedure  $I$  establishes the *identity relation* and maps each element of  $\sigma$  on a point set with a single element equal to itself (we have ( $\epsilon: [\epsilon]$ )). The procedure  $N$  establishes a "negation" relation and maps the elements of  $\sigma$  as follows:

$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$
$[s_{16}]$	$[s_{15}]$	$[s_{14}]$	$[s_{13}]$	$[s_{12}]$	$[s_{11}]$	$[s_{10}]$	$[s_9]$
$s_9$	$s_{10}$	$s_{11}$	$s_{12}$	$s_{13}$	$s_{14}$	$s_{15}$	$s_{16}$
$[s_8]$	$[s_7]$	$[s_6]$	$[s_5]$	$[s_4]$	$[s_3]$	$[s_2]$	$[s_1]$

(each element in the upper line is mapped on the point set standing right below it in the lower line). We represent by  $I(\epsilon) = (\zeta)$  and  $N(\epsilon) = (\zeta)$  the processes of carrying out these two procedures.

$M$  has, in addition to the previous two procedures, two other procedures that realize mappings. They are  $>$  and  $<$ ; for them, the sets  $A$  appearing in the correspondence ( $\epsilon: A$ ) have usually more-than-one elements. The procedure,  $>$ , establishes the "*inclusion*" relation and,  $<$ , establishes the relation "*converse to inclusion*". It is convenient to represent these relations by the structure of Fig. 5; in this structure each node represents an element of  $\sigma$  and it is designated by a symbol in McCulloch-Pierce's notation. (See Fig. 1 for definitions of notation.) A node-element "includes" another node if it stands higher in a chain on which both the nodes stand; a node "converse includes", or is "included" by another node if it stands lower in a common chain. Note that a node-element stands in the "negation" relation

with a node located symmetrically to it with respect to the center  $c$  of the structure.

We represent by  $>(\epsilon) = A$  and  $<(\epsilon) = A$  the processes of carrying out the procedures  $>$  and  $<$  respectively.

The distinction that we have made between elements and point sets is useful since we can now consider all mapping procedures as

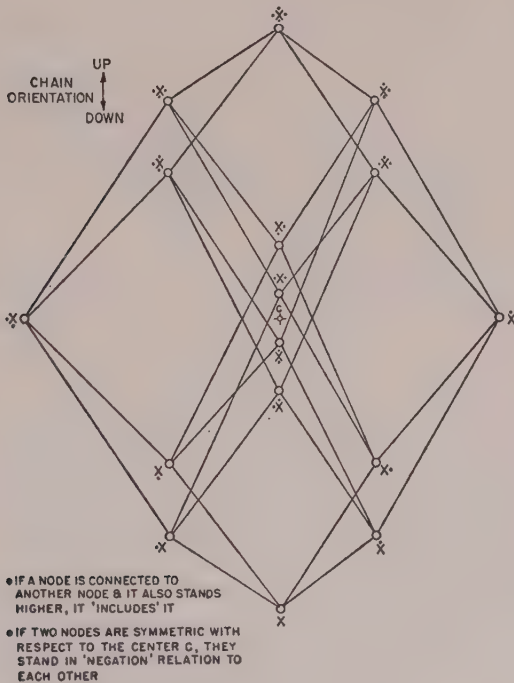


FIG. 5

prescriptions on how to “move” from an *element* of a certain set to another *set*. We can then assume that, on basis of the previous point of view,  $M$  can specify a procedure for “moving” from a collection of elements (a subset of  $\sigma$ ) to a collection of sets (a subset of  $2^\sigma$ ); this way  $M$  has a specification for a slightly more powerful procedure that consists of a repetition of any of the existing element-to-set procedures, and provides a way to “move” from a set to more-than-one sets. If the repetitive scheme utilizes

one of the previously discussed elementary procedures, then we consider it as an elementary procedure (with different terminal properties). As an example, if the procedure  $>$  is used in the repetitive scheme, then  $>_n(A) = (B_1, B_2, \dots B_n)$  represents the operation of obtaining by repetition of  $>$  over the  $n$  elements of the set  $A$  the  $n$ -tuple of sets  $(B_1, B_2, \dots B_n)$ ; we consider this as an elementary operation.

A graphical representation of the procedures initially available to  $M$  is shown in Fig. 6.

We assume that  $M$  can apply the above procedures on elements of  $\sigma$  that appear in input items, as well as on elements of  $\sigma$  and subsets of  $\sigma$  that are produced by intermediate machine procedures. Thus,  $M$  can string available procedures and construct an enormous number of new compound procedures, that can be potential candidates for the desired  $h_D$ 's.

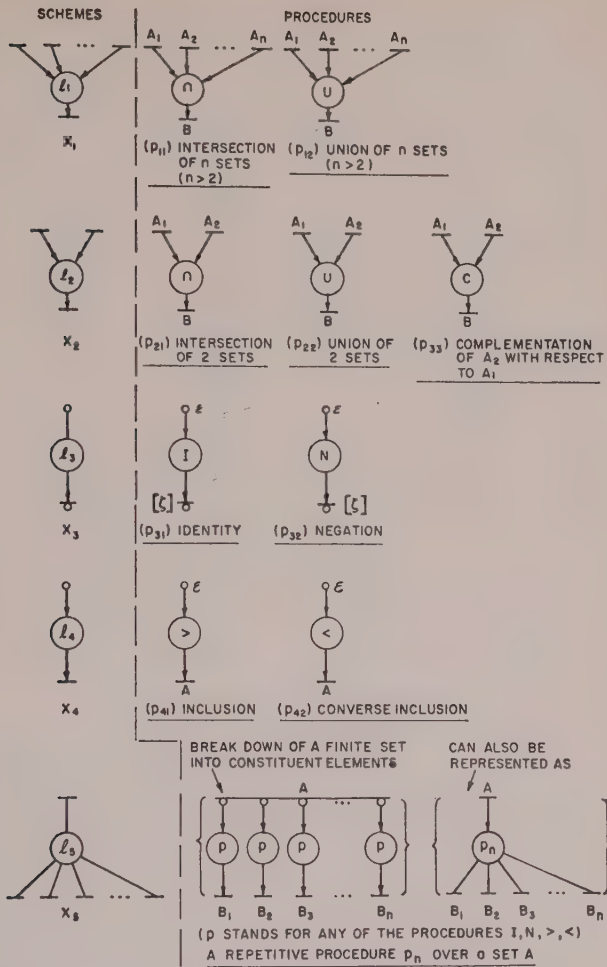
Given an input item,  $((\alpha, \beta):\gamma)$ , from  $D_k$ , the task of constructing a procedure  $h_{D_{kp}}$  such that:

$$h_{D_{kp}}(\alpha, \beta) = \gamma$$

through the utilization of available procedures and modes of construction, is a typical problem-solving task. Formal tasks of this type, especially in the area of theorem-proving, have received considerable attention in the last few years; much work on automated "heuristics" has already been done in this area.<sup>(7)</sup>

In our case, each input item presents a problem for solution; the solution attained is  $h_{D_{kp}}$ , which is analogous to the statement of a proof procedure in the case of theorem-proving.  $M$  will use heuristic procedures, or search *strategies* in order to solve the problem presented by each input item. The special problem of automatic theory formation is to find mechanisms that will *control the heuristics* on basis of the generation of previous solution-procedures, so that a single procedure can be attained that applies to all input items. An analogous problem will arise in the area of theorem-proving if the objective would be to automatically attain a systematic procedure, an algorithm, for the proof of all the true theorems in a certain area, given a sample of true theorems.

We consider the elementary procedures shown in Fig. 6 as the basic building blocks available to  $M$  at the start of its cognitive activity; they can be regarded as the elementary *moves* that  $M$  can



LEGEND:

- REPRESENTS A SET (A SUBSET OF  $\sigma$ ).  $A, B, A_1, A_2, \dots, A_n$  ARE NAMES FOR SETS.
- o REPRESENTS AN ELEMENT OF A SET,  $\epsilon$  IS A NAME FOR AN ELEMENT.
- $\sigma$  REPRESENTS A SET HAVING A SINGLE ELEMENT,  $[\zeta]$  REPRESENTS A SET FORMED OFF A SINGLE ELEMENT  $\zeta$
- $(p)$  REPRESENTS A PROCEDURE  $p$

FIG. 6

make in its combinatorial game. The system  $M$  classifies available building blocks (procedures) according to the kinds of objects between which they can be placed.\* This implies an analysis of the objects appearing in an input item before deciding on what move to take. Clearly, such analysis is limited by the analytical capability existing in the system. At the initial state of  $M$ 's operation, we assume that  $M$  can just decide whether an object is an element of  $\sigma$  or a subset of  $\sigma$ . It follows that  $M$  needs a description of its own procedures in terms of terminal properties (configuration of terminals, form of information negotiable at a terminal, i.e. element or set) in order to decide whether a procedure should be summoned and applied on an input item or not. We assume then that  $M$  has stored a network diagram in association with each one of its procedures; we call such a network diagram a *scheme*.

In general, a scheme  $X_i$ , is a network of connected blocks each of which has certain terminal properties; the subscript  $i$  denotes network types. We assign to each block of a scheme a symbol-name  $l_j$  ( $j = 1, 2, \dots$ ); an  $l_j$  stands either for a procedure or for a scheme. If all  $l$ 's of a scheme stand for procedures, then the resulting well-specified network represents a *procedure*. The schemes containing a single block are the *elementary schemes*  $X_1, X_2, X_3, X_4, X_5$ , and they are shown in Fig. 6; the  $l$  of an elementary scheme can only stand for a procedure. It is clear that for each elementary scheme there are several procedures.

One of the vital properties of  $M$  is the ability to *generate compound schemes*, made up from existing ones. We assume that  $M$  is equipped with techniques for joining together existing schemes in such a way that their joins are compatible. A few compound schemes of two and three constituent blocks are shown in Fig. 7(a). We measure the *complexity of a scheme* by the initial number of unspecified blocks it possesses; so the complexity of  $X_2$  is 1, of  $X_6$  is 2 and of  $X_8$  is 3. It is clear that there is no limit to the possible complexity of a compound scheme.

A block belonging to a compound scheme (represented by its  $l$ ) is specified in terms of a scheme.

The  $l$ 's of a compound scheme can be specified in stages, up to

---

\* This is an elementary kind of problem solving strategy: "abstract some of the properties in a problem statement so that at least completely unreasonable, incompatible, solution procedures will not be tried."



the point that all the  $l$ 's stand for procedures; at that point we have the specification of a *compound procedure*. The simplest specification of a compound scheme is obtained in two stages: first, all the  $l$ 's of the compound scheme are specified by elementary schemes, then each elementary scheme is assigned an elementary procedure; this type of specification is illustrated in

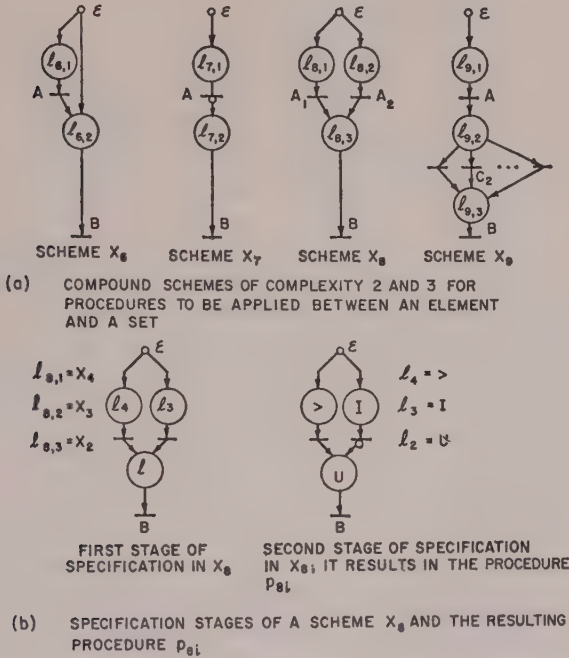


FIG. 7

Fig. 7(b). If in the first stage of specification an  $l$  is specified by a compound scheme, then the number of stages of specification is at least three, and the complexity of the fully specified scheme (the compound machine procedure) increases.

There are therefore two dimensions along which we can change the type and complexity of machine procedures: the one is the structure of the basic unspecified compound scheme; the other is the number of stages of specification and their character.

Thus, a small repertoire of elementary procedures and of basic network schemes, combined with the ability to perform hierarchical constructions to any level, enables  $M$  to generate a very large number of procedures. I do not know presently what are the bounds on this number, under given limitations of time and computing hardware; this problem, and its extension—the characterization of the *space of constructable procedures*—are worth studying in order to establish a theoretical basis for the evaluation of machine theory formation processes.

We assume that  $M$  stores basic schemes and elementary procedures in a memory  $\mu$ . At the start of the present process  $\mu$  contains the schemes  $X_1, X_2, \dots, X_5$ , and the corresponding procedures  $p_{11}, p_{12}, \dots, p_{51,n}, p_{52,n}$  (see Fig. 6 for definition); subsequently, as new compound schemes, i.e.  $X_6, X_7, \dots$ , are produced by  $M$  they are also stored in  $\mu$ .

Let  $m$  stand for any scheme  $X_i$  or procedure  $p_j$ ; we regard an  $m$  as the description of machine *means* to achieve a goal. In addition, let  $e$  denote the terminal specifications for a scheme—either the specification for a master scheme in terms of characteristics on input messages (of the type  $D : (I : O)$ ), or the specification for a block in a scheme (an  $l$ ); we regard an  $e$  as the statement of structural goals, *ends*, that the terminal properties of a scheme or procedure should satisfy.

$M$  establishes in  $\mu$  *associative links* between  $m$ 's and  $e$ 's. The decision for selecting an associative link in a given situation is made on basis of the distribution of *strengths of association* of the links. An unconditional strength of association of a link  $(m_i : e_j)$ , called  $a(m_i : e_j)$ , is a number stored in  $\mu$ , that indicates the relative *a priori* preference that  $M$  has in general for a link  $(m_i : e_j)$ . Given a certain  $e_j$ , the distribution of association strengths  $a(m : e_j)$  is composed of a component indicating relative participation of the link in the construction of previous successful syntheses, superimposed on a bias component that favors simpler schemes over the more complex ones; the bias component is a way of injecting in  $M$  a general preference for *simple* procedures. In addition to the number  $a(m_i : e_j)$ , each link that has been previously used in synthesis has associated with it several numbers,  $a(m_i : e_j)/E_k$ , that represent conditional strengths of association for different condition  $E_k$ ; an  $E_k$  represents a chain of structural

goals, starting from the highest goal (derived directly from the input message) and ending at the goal directly "covering"  $e_j$ . Thus, an  $a(m_i : e_j)/E_k$  indicates the relative *a priori* preference of  $M$  for the link  $(m_i : e_j)$ , if the link is considered at a certain specific stage of a synthesis; that stage is well defined by the sequence of previous goals. If a goal  $e_j$  appears in  $\mu$ , following a sequence of goals  $E_k$ , then the selection of an associative link is made on basis of a distribution of association strengths composed of the numbers  $a(m : e_j)/E_k$ , for all the links that have strengths conditional to  $E_k$ , and the numbers  $a(m : e_j)$  for the remaining links; the maximum of this distribution determines the choice of an association link. A link chosen under these conditions will carry from that point on a number for an association strength conditioned to  $E_k$  (if it did not have such a number previously). If a conditional strength drops below a certain limit, it is withdrawn from  $\mu$ ; this way storage will not grow wildly and  $M$  will keep only "interesting" information on associative paths.

The following example (see Fig. 8) illustrates the associative activity in  $\mu$ : an input message  $D_i : (\epsilon : A)$  is presented to  $M$ . After initial analysis in  $M$ , the goal  $e_0$  is stated in terms of class name (here  $D_i$ ), terminal characteristics of  $I$  (here  $\epsilon$ ) and terminal characteristics of  $O$  (here a set  $A$ ); then,  $e_0$  is applied to the input of  $\mu$ . In the next step, the values of the association strengths  $a(m : e_0)$  are compared for all relevant  $m$  (all existing schemes). Association is established through the strongest associative link to a scheme; assume that this is  $X_8$  in the present example. The network parameters  $l_{8,1}$ ,  $l_{8,2}$  and  $l_{8,3}$  of the scheme  $X_8$  are retrieved; their respective terminal properties are established and they are used for the statement of three new structural goals  $e_{8,1}$ ,  $e_{8,2}$ ,  $e_{8,3}$ . These goals are next applied to the input of  $\mu$ . Let us follow  $e_{8,1}$ . From the distribution of association strengths (composed of numbers  $a((m : e_{8,1})/e_0)$  for the links  $(m : e_{8,1})$  that have strengths conditioned to  $e_0$ , and of numbers  $a(m : e_{8,1})$  for the remaining links), the strongest link is found to lead to the scheme  $X_4$ . Now, the network parameters  $l_4$  of  $X_4$  is retrieved, it produces the structural goal  $e_4$ , which leads to the elementary procedure  $p_{4,1}$  on basis of the superiority of the strength  $a(p_{4,1} : e_4)$ ; note that this last choice was based on an unconditional association strength (because say  $p_{4,1}$  was not previously linked to the sequence of goals

$e_{01}, e_{8,1}, e_4$ ). Completing the process for the two remaining branches we get the procedure shown in Fig. 8. We show in Fig. 9 the tree of association that grew in this simple example.

Our system of  $(m : e)$  associations has some similarity with Tolman's<sup>(8)</sup> idea of a store of *means-end readinesses* governing

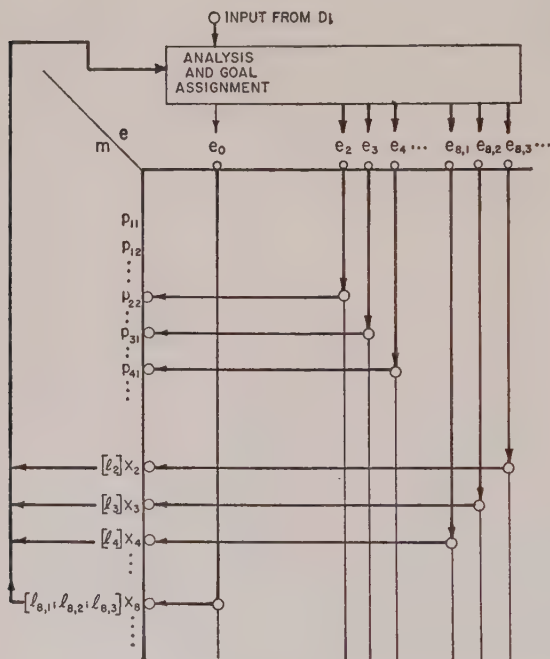


FIG. 8

human purposeful behavior; our concept of association strength would correspond to Tolman's "confidence on Beliefs".

For a given input item ( $I : O$ ) from a class  $D_j$ , a tree of associations,  $\Delta i$ , grows in  $M$  and, when completed, it represents a certain procedure  $p_i$  proposed by  $M$ .  $p_i$  is applied to the input item, and if successful (i.e. if  $p_i(I) = O$ ), then all the associative links in  $\Delta i$  are strengthened; if unsuccessful, then these associative links are weakened.

It is of extreme importance not to wait for the growth of a complete tree in order to check if the entire procedure is satisfactory. Measurements of *partial success* are necessary in order to restrict the number of trees that should be grown before a satisfactory procedure is attained. Here comes the crucial role of *heuristic rules* of strategy that determine how to assign *subgoals* during the synthesis of procedures (growth of association trees).

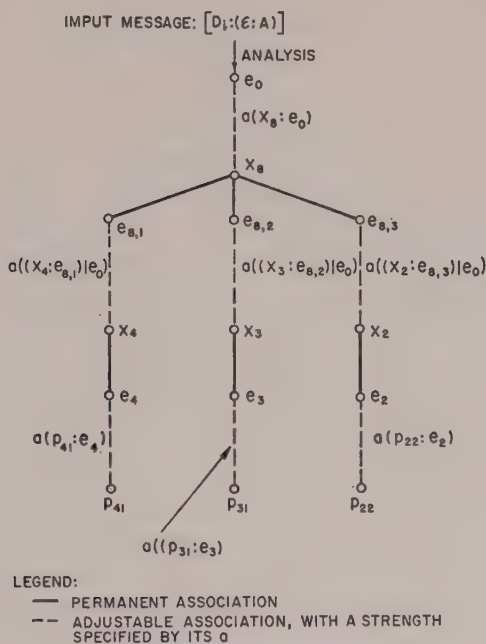


FIG. 9

The way in which heuristic rules control the tree growth process is described next: consider a scheme  $X$  that is proposed in response to the input message  $D_j : (I : O)$ . At this point, the goal is to find a procedure (which is a specification of  $X$ , denoted  $[X]$ ) that will satisfy the input item, i.e. such that  $[X](I) = O$ .  $X$  might contain subblocks  $l_j$  that are structurally compatible, but they do not have specific goals attached to them, i.e. there is no specific request in



the form  $[l_j](I_1) = O_1$  for  $l_j$  (where  $I_1 : O_1$  is a correspondence that  $[l_j]$  should "justify"). Specific subgoals are assigned to the  $l_j$ 's of  $X$  by heuristic rules;\* clearly, the assignment of subgoals must be consistent with  $X$ 's goal. Now, each  $l_j$  with its subgoal might originate the proposal for another scheme, whose subblocks are assigned new compatible subgoals by heuristic rules, and so on. Consider the smallest partial procedure (the smallest subtree,  $\Delta_j$ , ending at the bottom of the main inverted tree—the subtree "hangs" from a certain  $e_k$ ) that has an assigned subgoal. If by testing this partial procedure over its subgoal, the procedure is unsuccessful, then all association links of  $\Delta_j$  are weakened. Now, since a new distribution of association strengths emerges, and some of the links that previously had maximum strength in  $\Delta_j$  might drop to secondary positions, a different tree  $\Delta_1$  is likely to grow starting from  $e_k$ . A different partial procedure will be now generated and then tested over the subgoal of  $e_k$ . Eventually, either a successful procedure is attained, or a certain effort limit is exceeded. When a successful procedure appears, all associative links of the tree emanating from  $e_k$  are strengthened; this partial procedure is then a candidate to be combined with other partial procedures, in order to be tested over higher subgoals. Finally, a collection of partial procedures forms a total procedure that should be tested over the main goal. If unsuccessful, then the association links of the entire tree hanging from  $X$  are weakened, and the tree growth process starts all over till a certain effort limit is reached; if a total procedure is successful, then all the association links of its corresponding tree are strengthened.

Different reinforcement policies lead to different "tree growing" processes, and, to different *convergence* properties of the entire cognitive process. We have assumed so far a policy whereby the reinforcement of a higher level association is determined by the reinforcements of all the lower level association that it "covers" in the association tree.

If the entire distribution of association strengths for a certain

---

\* Each choice of heuristic rules is expected to result in different kinds of constraints for the tree growth process. We have used so far certain simple, arbitrarily chosen, heuristic rules (see the illustrative example in the next section); but we don't have yet a way of assessing how such a choice affects the formation process in  $M$ .

goal drops below a certain minimum value, and that goal cannot be modified by changing a higher goal (suppose for instance that this is the first or main goal), then  $M$  generates a new scheme to be tried for that goal; the new scheme enters permanently the repertoire of schemes in  $M$ . The introduction of new overall schemes in  $\mu$  is an important factor in the versatility and capacity for evolution of  $M$ .

During the tree formation process, a goal  $e$  considers particular associations to schemes in a regular sequential *scan*, going mostly from simpler schemes to more complex ones, and from *a priori* more promising schemes to *a priori* less promising ones.

After a successful procedure is formed for an input item of  $D_j$ , a new input item from  $D_j$  is admitted.  $M$  will tend to propose first the procedure that was successful for the previous item. If successful again, this procedure will be reinforced. If unsuccessful, a new procedure, usually more complex than the previous one, will be found. After treatment of a large number of input items (picked, without special choice, from the class  $D_j$ ), we expect that a single configuration of association links (that corresponds to a certain procedure  $h_{D_j, \nu}$ ) will become dominant and hopefully will remain *stable*;  $h_{D_j, \nu}$  is then the working hypothesis for a machine-theory of  $D_j$ . The problem of understanding the relationship between reinforcement policies and the dynamic behavior—especially the *stability*—of the synthesis process (the tree growth process) is, I feel, a fundamental one, and needs much further study.

If there is no scheme in  $\mu$  that represents in a single block the general terminal properties of a newly attained stable procedure, then such scheme is formulated and introduced in the repertoire of schemes; the corresponding new procedure enters the repertoire of elementary procedures. This way, building blocks of higher power are available to  $M$  for the construction of procedures in future cognitive tasks. Future machine representation produced by  $M$  tend to be expressed in a language that uses the more powerful "primitives" (the newly acquired elementary procedures) or, if this is difficult, in a *mixed language* (using both the initial elementary procedures and the acquired ones).

To summarize then, an input item from a class  $D_j$  initiates a scan in the space of association trees. The scan is restricted by heuristic rules and it moves from simple and *a priori* more

promising trees to more complex and *a priori* less promising trees. Each input item leaves a trace in  $\mu$  by strengthening the association links of that tree that has resulted in a procedure "satisfying" the item. The association tree that satisfies a large number of input items from  $D_j$  becomes increasingly "solidified" and, when solid enough, it represents a stable synthesis method that  $M$  uses to assemble a procedure which is the machine-theory of  $D_j$ . If at a later time, a predictive task of the type  $D_j : (I : ?)$  is assigned to  $M$ , then a synthesis process will be initiated in response to the name  $D_j$ ; this process will be governed by association strengths that have been conditioned to  $D_j$  during the previous cognitive activity. The synthesis will result in the specifications of the procedure  $h_{D_j, \nu}$  that  $M$  had previously formed for  $D_j$ ;  $h_{D_j, \nu}$  will then be used in order to compute  $O$  from the given  $I$ .

GENERAL DESCRIPTION OF THE COGNITIVE  
ACTIVITY IN  $M$  OVER A SEQUENCE OF CLASSES  
FROM THE SET  $D_1, D_2, \dots, D_{16}$

In order to illustrate the character of the cognitive activity in  $M$ , I will describe certain highlights of the theory forming processes over the sequence of classes  $D_2, D_{13}, D_5, D_6$  (see eq. 7 for definitions).

The processes to be outlined come from preliminary hand simulation experiments with the proposed  $M$ .<sup>\*</sup> Before going into the operation of  $M$  over  $D_2, D_{13}, D_5, D_6$  it should be repeated that  $M$  starts with the elementary procedures and schemes shown in Fig. 6, with a certain distribution of association strengths that is biased toward simple and *a priori* useful schemes, and with a set of heuristic rules for synthesis.

*Class  $D_2$ :* An input Message  $D_2 : (\alpha, \beta : \gamma)$  enters  $M$ .

---

<sup>\*</sup> It is interesting to note that our research roughly parallels the operation of the  $M$  under study, as it can be summarized as follows: choose classes  $D_i$  which known class belongedness rules—attempt to attain from extensive definitions of  $D_i$  the class belongedness rules of  $D_i$ , through a postulated "procedure"  $M$  (the theory formation mechanism)—according to results of tests (simulations) adjust  $M$ —test the adjusted  $M$ , and so on, till an  $M$  with a high degree of confidence is attained. We are presently at the stage of testing an adjusted  $M$  over the set of classes  $D_1, D_2, \dots, D_{16}$ .

The terminal properties of the input item are analysed. A structural goal  $e_0$  is formulated for the terminal properties of a required scheme;  $e_0$  is in terms of the terminal properties of the input item.

No compatible scheme is found in  $M$ .

A simple new scheme,  $X_a$ , is formed that satisfies  $e_0$ .  $X_a$  is shown in Fig. 10; it has three blocks,  $l_{a1}$ ,  $l_{a2}$ ,  $l_{a3}$ .

On basis of the input message and the scheme  $X_a$ , an assignment of specific subgoals to the blocks of  $X_a$  is sought. (Assignment of "reasonable" subgoals (based on existing heuristic rules), even a partial assignment, substantially simplifies the process of searching for a suitable procedure).

$l_{a3}$  is considered first (from the heuristic rule: "try to work your way back from the results to the premises") and a structural goal  $e_{a3}$ , abstracted from the terminal properties of  $l_{a3}$ , is formulated.

The association ( $X_2 : e_{a3}$ ) is proposed (on basis of the system's tendency to first explore the simplest schemes).

$X_2$  has one block  $l_2$  (see Fig. 6).  $l_2$  is specified by the procedure  $\cap$  (according to an assumed initial distribution of association strengths).

On basis of the given input message, the scheme  $X_a$ , and the specification of  $e_{a3}$  by  $\cap$  (through the association chain  $e_{a3} \rightarrow X_2 \rightarrow e_2 \rightarrow \cap$ ), an assignment of specific subgoals to  $l_{a2}$  and  $l_{a3}$  is made as follows:

$g_{l_{a1}}$ : The procedure  $p_{l_{a1}}$  that specifies the scheme  $l_{a1}$  should satisfy:

$$p_{l_{a1}}(\alpha) = A_1, \text{ where } \gamma \in A_1.$$

$g_{l_{a2}}$ : The procedure  $p_{l_{a2}}$  that specifies the scheme  $l_{a2}$  should satisfy:

$$p_{l_{a2}}(\alpha) = A_2, \text{ where } \gamma \in A_2.$$

The associations ( $X_4 : e_{a1}$ ), ( $X_4 : e_{a2}$ ) are proposed (on basis of  $M$ 's tendency to first explore the simplest schemes, compatible with the desired terminal properties).  $X_4$  has a single block  $l_4$ . According to the existing strengths of association, a compatible elementary procedure,  $p_{4l}$ , is proposed as specification of  $l_{a1}$ ; similarly for  $l_{a2}$ .



The goal attainment of  $l_{a1}$  is tested by performing the operation  $p_{4i}(\alpha) = A_1$  and examining whether  $\gamma \in A_1$ ; similarly, the results of an operation  $p_{4j}(u)$  is examined for  $l_{a2}$ .

If  $g_{a1}$  or  $g_{a2}$  or both are not satisfied, then  $l_{a1}, l_{a2}$  try different compatible specification of increasing complexity. Here the compound schemes  $X_6, X_7, X_8, X_9$  are tried in sequence. For a new proposed scheme (say an association  $(X_8 : e_{a1})$  that produces again three sub-blocks  $l_{8,1}, l_{8,2}, l_{8,3}$ ), new specific subgoals are assigned to the sub-blocks, the sub-blocks get initial specification by simple procedures, and so on, as was previously done at the immediately higher hierarchical level in the association tree (namely at  $X_a$ ).

When the specification of  $l_{a1}, l_{a2}$  reaches a certain complexity, then construction along these lines, stops, and the specification of  $l_{a3}$  is reconsidered; a new specification for  $l_{a3}$  is made and the process of searching for suitable specifications for  $l_{a1}, l_{a2}$  repeats again along lines similar to the ones described above.

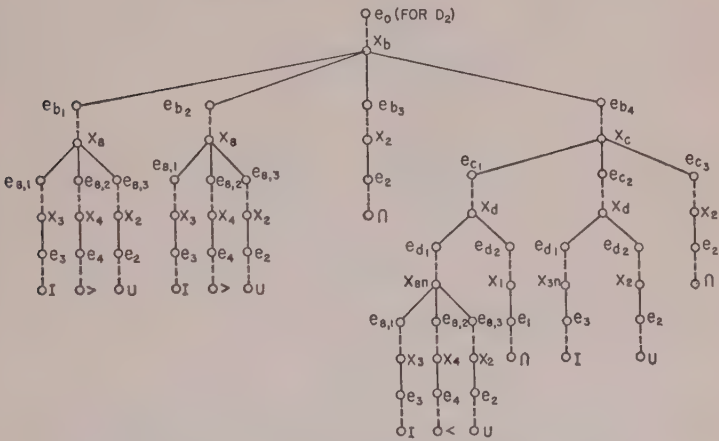
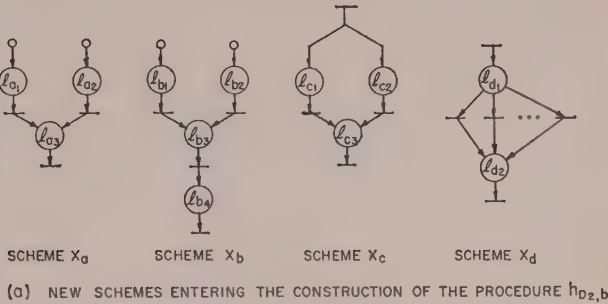
If, during the previous, process, all subgoals are satisfied, for certain compound procedure  $p_{X_{a,i}}$ , this procedure is tested over the input message; if it satisfied the input message, i.e. if  $p_{X_{a,i}}(\alpha, \beta) = \gamma$ , then activity over that message stops, otherwise the specification for  $l_{a3}$  is reconsidered and a new process for specifying  $l_{a1}, l_{a2}$  is carried on. Eventually a compound procedure is found for an input message. However, no single procedure is found for a number of input messages. This reduces in time the strength of association  $(X_a : e_0)$  below the acceptable minimum level.

$M$  forms then the slightly more complex scheme  $X_b$  (see Fig. 10), and proposed the association  $(X_b : e_0)$ . Starting from this association, a tree of associations is constructed that finally results in a compound procedure  $h_{D_{2,b}}$  that consistently satisfies a long sequence of input items entering  $M$  from the class  $D_2$ . The association tree as well as the resulting compound procedure are shown in Figs. 10 and 11.

With the attainment of  $h_{D_{2,b}}$ , the Goal  $I$  of  $M$  over the class  $D_2$  is satisfied; i.e.  $h_{D_{2,b}}(\alpha, \beta) = \gamma$  for all input items  $((\alpha, \beta) : \gamma)$  that enter  $M$  from the class  $D_2$ . Following this,  $M$  stores  $h_{D_{2,b}}$  in  $\mu$  as a new elementary procedure. In correspondence with  $h_{D_{2,b}}$  a new scheme  $X_e$  is stored in  $\mu$ ;  $X_e$  is shown in Fig. 11.



Class  $D_{13}$ : Following the cognitive activity over  $D_2$ ,  $M$  proposes to “justify” an input item from  $D_{13}$  with the procedure  $h_{D_2,b}$ , after choosing an association to the new scheme  $X_e$ ; it fails, and then  $M$  generates new compound schemes containing



(b) ASSOCIATION TREE FOR SPECIFYING THE PROCEDURE  $h_{D_2,b}$

FIG. 10

$X_e$  (see Fig. 12), on basis of which it specifies—tries new procedures, and fails again. Then  $M$  regresses to consideration of the scheme  $X_b$ , that is specifiable in terms of more elementary procedures. Starting from  $X_b$ ,  $M$  constructs a tree of associations that finally results in a compound procedure  $h_{D_{13},b}$  that satisfies all the items entering from  $D_{13}$ . The structure of  $h_{D_{13},b}$  is the same as

$h_{D_{2,b}}$ . The only difference between the two procedures is in the blocks specified by inclusion and inverse inclusion. An “>” in  $h_{D_{2,b}}$  appears as an “<” in  $h_{D_{13,b}}$ , an “<” in  $h_{D_{2,b}}$  appears as an “>” in  $h_{D_{13,b}}$ . As in the case of  $h_{D_{2,b}}$ ,  $h_{D_{13,b}}$  is stored in  $\mu$  as an elementary procedure; the scheme  $X_e$  represents now both the procedures  $h_{D_{13,b}}$  and  $h_{D_{2,b}}$ .

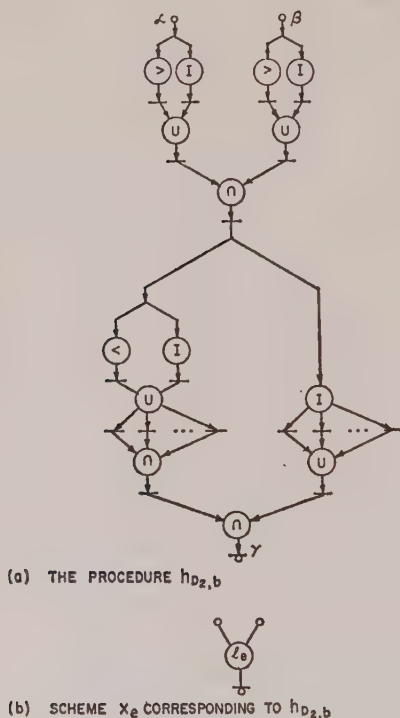


FIG. 11

*Class  $D_5$* :  $M$  proposes first an association to the scheme  $X_e$ ; it fails and it then tries the compound schemes  $X_f, X_g, \dots$  shown in Fig. 12. A specification of the scheme  $X_g$  constitutes a satisfactory procedure  $h_{D_{5,g}}$ ; this procedure is shown in Fig. 13.

*Class  $D_6$* :  $M$  considers first the schemes  $X_g, X_e$  and fails; it tries then other compound schemes (from Fig. 12). A specification

of the scheme  $X_j$  results in a satisfactory procedure,  $h_{D_{6,j}}$ ; this procedure, together with its corresponding association tree are shown in Fig. 14.

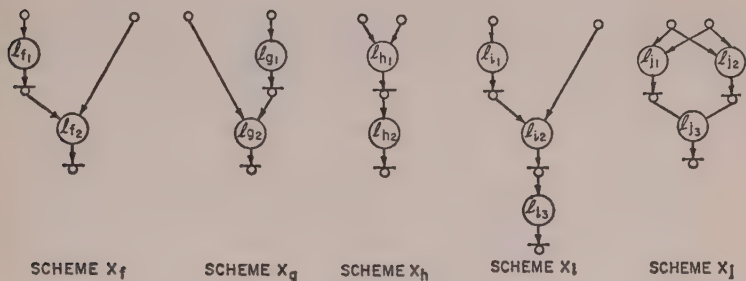


FIG. 12

Extending the experimentation with  $M$  over all the remaining classes (in the sequence  $D_7, D_4, D_9, D_{10}, D_{14}, D_{15}, D_3, D_8, D_{11}, D_{12}, D_{16}$  and  $D_1$ ), the procedures  $h_{D_{7,p}}, h_{D_{8,p}}, \dots, h_{D_{1,p}}$  are also attained.

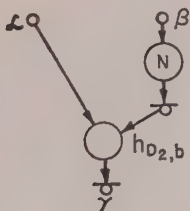


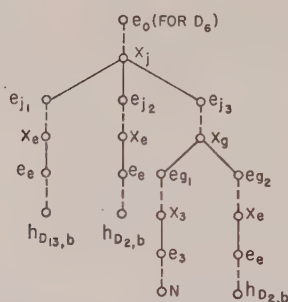
FIG. 13

INTERPRETATION AND APPRAISAL OF THE MACHINE-THEORIES ATTAINED IN  $M$  FOR THE CLASSES  $D_1, D_2, \dots, D_{16}$

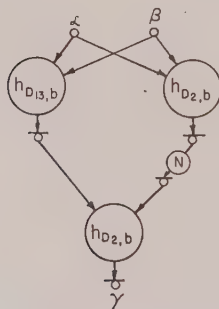
If we consider the structure of Fig. 5, where the relation of inclusion and negation for each element of the set  $\sigma$  are diagrammed, we can interpret any procedure  $h_{D_{1,p}}$  as a prescription for moving from two nodes of the structure to a third node (the nodes are not necessarily distinct).

Thus,  $h_{D_{2,b}}(\alpha, \beta)$  produces the "greatest lower bound" or "Infimum" (abbreviated "Inf") of  $\alpha$  and  $\beta$ . If  $\alpha$  and  $\beta$  stand on

different chains, the  $\text{Inf}(\alpha, \beta)$  coincides with  $\alpha$  or  $\beta$ , whatever stands lower in the chain. Note that the “lower bound” part of “greatest lower bound” is obtained by the part of the procedure  $h_{D_{2,b}}$  specifying the structural goals  $e_{b1}, e_{b2}, e_{b3}$ ; the additional restriction “greatest” is obtained by the part of  $h_{D_{2,b}}$  that specifies  $e_{b9}$  (see Figs. 10 and 11 for definitions). Since the members of  $D_2$



(a) ASSOCIATION TREE FOR SPECIFYING THE PROCEDURE  $h_{D_{e,l}}$



(b) THE PROCEDURE  $h_{D_{e,l}}$

FIG. 14

are instances of the mapping  $S_2\alpha\beta = \gamma$  (which corresponds to a conjunction of two-argument proposition functions), then  $\text{Inf}$  is the operational meaning of the mapping  $S_2$  in  $M$ .

The operation  $h_{D_{13,b}}(\alpha, \beta)$  produces the “least upper bound” or “Supremum” (abbreviated “Sup”) of  $\alpha$  and  $\beta$ . This is similar to the interpretation of  $h_{D_{2,b}}$  on the structure of Fig. 5, with polarity reversed; i.e. “down” is changed by “up” and “lowest”

by "highest". Here the machine meaning of the mapping  $S_{13}$  (which corresponds to a disjunction of two-argument propositional functions) is the operation Sup over elements from  $\sigma$ .

An interpretation of all the attained procedures,  $h_{D_{i,p}}$ , as operations over the structure of Fig. 5, is given in Fig. 15; we regard the operations, or "laws of motion", described in this last figure as a machine theory for the sixteen classes  $D_1, \dots, D_{16}$ . The

Classes, $D_i$ (In the order of their presentation to $M$ )	Operations over the partly ordered Set $\sigma$ , defining the classes $D_i$
$D_2$	$\text{Inf}(\alpha, \beta) = \gamma$
$D_{13}$	$\text{Sup}(\alpha, \beta) = \gamma$
$D_5$	$\text{Inf}(\alpha, N\beta) = \gamma$
$D_6$	$\text{Inf}[\text{Sup}(\alpha, \beta), N \text{Inf}(\alpha, \beta)] = \gamma$
$D_7$	$\text{Sup}[\text{Inf}(\alpha, \beta), \text{Inf}(\alpha, N\beta)]$
$D_4$	$N \text{Sup}(\alpha, \beta) = \gamma$
$D_9$	$\text{Sup}[\text{Inf}(\alpha, N\beta), N \text{Sup}(\alpha, \beta)]$
$D_{10}$	$\text{Sup}[N \text{Sup}(\alpha, N\beta), N \text{Sup}(\alpha, \beta)]$
$D_{14}$	$\text{Sup}(\alpha, N\beta) = \gamma$
$D_{15}$	$N \text{Inf}(\alpha, \beta) = \gamma$
$D_3$	$N \text{Sup}(\alpha, N\beta) = \gamma$
$D_8$	$\text{Sup}[\text{Inf}(\alpha, \beta), N \text{Sup}(\alpha, N\beta)]$
$D_{11}$	$\text{Sup}[\text{Inf}(\alpha, \beta), N \text{Sup}(\alpha, \beta)] = \gamma$
$D_{12}$	$N \text{Inf}(\alpha, N\beta) = \gamma$
$D_{16}$	$T(\alpha, \beta) = \gamma = s_{16}$
$D_1$	$F(\alpha, \beta) = NT(\alpha, b) = \gamma = s_1$

Definition of  $T(\alpha, \beta)$ :

$$T(\alpha, b) = \text{Sup}[\text{Sup}[\text{Inf}(\alpha, \beta), \text{Inf}(\alpha, N\beta)], \text{Sup}[N \text{Sup}(\alpha, N\beta), N \text{Sup}(\alpha, \beta)]]$$

FIG. 15. Machine definitions of the classes  $D_1, D_2, \dots, D_{16}$ , interpreted as operations over the partly ordered set  $\sigma$ , whose structure is shown in Fig. 5.

procedure  $h_{D_{16}}$  attained by  $M$  (interpreted as a  $T$  operator) is not a simple one; however the complexity of the formation process for  $h_{D_{16}}$  (based on the previously formed and stored schemes  $X_e, X_f, \dots, X_j$ ), is not more serious than the complexity of the formation process for  $h_{D_{2,b}}$  (based on the simpler schemes available at that earlier stage). Note that each class is defined in terms of Inf, Sup and  $N$ .

It is easy to see that the form of the theory attained by  $M$  is a *lattice theoretical* version of propositional calculus.<sup>(9)</sup> The partly



ordered set  $\sigma$ , under the inclusion relation, forms a distributive lattice with unique complementation; any two elements of  $\sigma$  have an Inf and a Sup, as we have indicated above. This form of theory might prove useful for investigations of problems in adjustable switching networks (such as the problem of logical stability and reliability in switching networks).

The formation of an algebraic structure for propositional calculus inside a machine that started from a "stage of development" where it could manipulate sets and a few relations, has some correspondence with Piaget's view<sup>(10)</sup> of the development of a "*structure d'ensemble*" of logical operations in a child's brain, subsequent to the stage of familiarity with classes and relations. It can be expected that further specification and simulation of  $M$  over the interval of development that I have described here will contribute to the evaluation and advancement of such theories on human mental development.

There is also functional similarity between certain models of human perception<sup>(11)</sup> (where an information-hungry organism seeks a congruence between its own internal operational representation of a sensory event and the many altering "figures" of the event as it moves with respect to the sensors) and the  $M$  that I have outlined here; this is another area where further study of  $M$  can provide theoretical backing for psychological research.

From an *engineering standpoint*, the process of theory formation in a machine provides a design approach whereby the storage of a limited amount of input data from a class is organized in such a way that subsequent questions pertinent to the entire class can be processed in a versatile way fast and *economically*—without need for table lookup in a high-speed memory filled with an enormous number of specific instances from the class.

The  $M$ -representation of the classes  $D_1, D_2, \dots, D_{16}$  (the machine-theory of these classes) provides an example of such organization of storage. It consists of: (a) a collection of initial elementary procedures, mechanized either as stored correspondence tables, stored routines or switching networks (these correspond to present computer microinstructions); also a set of compound procedures that assume the status of elementary procedures during the operation of  $M$  (corresponding to computer macroinstructions or stored microprograms), (b) an association matrix, with controllable

stored association strengths, that functions as an assembler of compound procedures from elementary procedures (it can be interpreted as a *controllable program generator* producing a distinct program for each class), (c) a processor that manipulates elements from  $\sigma$  according to prescribed procedures and computes specific answers in response to retrieval questions. (It corresponds to the control and arithmetic units of present computers.)

The mechanization of the program generator in the previous *active storage* scheme poses a challenging technological problem; it seems at present that flexible content addressed memories will provide a promising answer to this problem.

An active storage organization of the type described previously is expected to be, in most practical cases, more *economical* than complete storage of a large set of instances from a class. In addition, it provides predictive power over unknown parts of a class, and it promises an enormous capacity for assimilating and hierarchically structuring more complex information, through the construction of new broader machine procedures in terms of all the previously attained ones. However, a price must be paid, in the form of a substantial processing effort expanded during the "write in" into storage; this effort is that of theory formation (learning) where data is suitably encoded "in machine language" before becoming assimilated. As in the case of rationalizing the utility of science, the price paid is counterbalanced by the expected gains in the performance of future useful tasks.

#### STATE OF INVESTIGATION. FUTURE WORK

In summary, the investigation that I have described here has resulted to date in the formulation of general features for a theory-formation machine,  $M$ , the selection of a suitable formal cognitive environment for  $M$ , the determination of the feasibility of the cognitive task, and the extraction of several focal problems for further study.

The next stage of the investigation will concentrate on the detailed specification and evaluation (both experimental and theoretical) of the mechanization of certain key functions in  $M$ . These are: generation of new schemes for compound procedures, control of the generation of association trees, control of the

association strengths and its effect on the convergence properties of the theory-formation process, and control of processing power allocated in various partial processes of *M*.

I believe that a more general understanding of theory-forming-learning, mechanisms will come from a detailed study of the problems that emerge from the mechanization of our relatively simple, cognitive task in *M*. Some of these problems do not seem easy, but I believe that they are not insurmountable.

#### ACKNOWLEDGMENT

I am indebted to L. Kovendy for his interesting comments and discussions.

#### REFERENCES

1. N. CHOMSKY and G. A. MILLER, Pattern conception, Report No. AFCRC-TN-57-57 (August 7, 1957) (ASTIA Doc. No. AD110076).
2. S. WATANABY, Information theoretical aspects of inductive and deductive inference, *IBM Journal of Research and Development* 4, No. 2 (April, 1960).
3. J. PIAGET, *The Origins of Intelligence in Children*. (1936), trans. by M. COOK, New York; International University Press (2952).
4. W. S. McCULLOCH, Agatha Tyche of nervous nets—the lucky reckoners, *Mechanization of Thought Processes*, Vol. II, HMSO, London, pp. 611–25 (1959). (Presented at the Symposium held in November, 1959, at the National Physical Laboratory, Teddington, England.)
5. R. CARNAP, *Logical Foundations of Probability*, University of Chicago Press (1950).
6. See M. HALLE and K. N. STEVENS, Analysis by synthesis, W. WATHEN-DUNN and L. E. WOOD (eds.) *Proceedings of the Seminar on Speech Compression and Processing* (Air Force Cambridge Research Center Technical Report 59-198, Bedford, Mass.) (December, 1959).
7. See for instance: H. GELERTNER, Realization of a geometry theorem proving machine, *Proceedings of the First International Conference on Information Processing* (held in Paris, June, 1959).
8. E. C. TOLMAN, Principles of purposive behavior, from *Psychology: A Study of a Science* (edited by S. KOCH) McGraw Hill, N.Y. (1959).
9. G. S. BIRKHOFF, Lattice theory, *American Mathematical Society*, New York (1948).
10. J. PIAGET, La logistique axiomatique on “pure”, la logistique operatoire on psychologique et les Realites auxquelles elles correspondent, *Methodos*, Vol. IV (1952).
11. See the discussion on the posterior intrinsic system of the brain in KARL H. PRIBRAM, Theory in physiological psychology, *Annual Review of Psychology*, Vol. II (1960).

## DISCUSSION

COWAN: I must make some comments on Dr. Amarel's paper. First of all, turning to some of the problems of obtaining solutions in some of these two-valued systems, which is related to something he talked about, I would like to point out that in a recent thesis at London University, a former colleague of mine, Peter Goswami, has invented an algorithm for obtaining all solutions of any two-valued logical problems in a minimum number of steps, and it completely reduces any search problem to a very short number of trials. The second point, that of obtaining solutions, is related to the many-valued case. If you cannot separate your particular two-valued logical systems, then there is some stronger ordering system, and you do not need to use this algorithm. It does become rather long for the many-valued systems. You can usually show that a search procedure is better, more efficient, in the many-valued systems for certain cases.

The other point which I would like to touch on is the last point, when you talk about the recognition of invariance in these systems. I think this problem has been completely solved already by Mautner in a paper called "Logic and Invariant Theory", in which he sets up the conditions for obtaining invariants by using the tensor formulation of Boolean logic. This may be of interest to you.

AMAREL: I would like to clarify one thing. I didn't choose this problem in logic because I want to find invariances in the equivalences between triplets of propositional functions and single propositional functions. I chose it because it is a relatively simple problem and there is hope that it can be mechanized. I am interested to find in what way, starting with a small number of concepts and tools, a machine can evolve a stable invariant representation of an environment which is useful for later operation of the machine. I am interested in the properties of hypothesis formation processes.

COWAN: It is not obvious to me why you need the particular apparatus that you constructed here.

AMAREL: Because it rests on manipulation of sets and relations (with which I have assumed that the machine is familiar).

COWAN: But do you need it? It seems to me that the particular calculus developed by McCulloch is ideal here.

AMAREL: Again, I am not trying to develop an algorithm in order to obtain solutions to various logical problems; such algorithms can be developed in many different ways. What I would like to do is to use this problem as an experimental research tool by which I can see how machine hypotheses are formed, how they die in case they are not found suitable and how they propagate to stable configurations.

COWAN: But, as I say, if that is what you want, then the contributions of Schutzenberger and Grau are quite relevant to this, because Schutzenberger characterizes what you can do when you look at arbitrary compositions of these things, and Grau adds to that the particular invariants that you can recognize in this kind of system. It seems to me that this kind of work would be of interest to you.

AMAREL: The problem of *what* one can do with a certain number of primitive things is less interesting to me at present than the problem of *how* feasible things can be done in a machine.





**PETER H. GREENE**

*Committee on Mathematical Biology,  
The University of Chicago*

## NETWORKS WHICH REALIZE A MODEL FOR INFORMATION REPRESENTATION\*

The recognition and discrimination of meaningful perceptual stimuli presupposes the active formation of stable perceptual elements to be recognized and discriminated. A person lacking this process would combine all sorts of stimuli into meaningless groups. In like fashion, purposeful activity must presuppose a formative stage in which meaningful patterns are synthesized from the actions of whatever set of muscles may chance to be employed. Mathematical investigations have heretofore not generally dealt with these formative stages, because so little is known about the means of representation of information in the nervous system. Equally little of value is known about the representation of percepts, ideas and actions in artificial systems. Most of the concepts which have been developed in this area of inquiry deal primarily with external relations among perceptual elements which are supposed to be given in a form which does not represent their internal structure. These elements serve merely as labels for our ideas. Thus the theories account for sorting and combining of perceptual elements and of concepts, but do not deal with the meanings of these entities.

I have elsewhere<sup>(1)</sup> reviewed arguments that one of the main requirements is that each element of information contain partial representations of many other elements and schemata for their interconnection. These arguments, based upon the work of a number of philosophers, psychologists and logicians, led to the

---

\* This research was supported by the Office of Naval Research under Contract No. Nonr 2121(17) NR 049-148. Reproduction in whole or in part is permitted for any purpose of the United States Government.

suggestion<sup>(2)</sup> that some of these requirements might be met if information could be thought of as being represented in the form of vectors in a function space (such as modes of oscillation of a complicated network) which may be resolved into components in various coordinate systems. These systems represent various points of view from which the information may be regarded, and some of the information each system may be elicited by a probabilistic mechanism for further stages of processing.

Some of the mathematical properties of that model seemed to resemble certain features of higher mental processes. Analogues seemed to exist for the clarification of an idea as it comes into awareness, associative and learning processes, the representation of potentialities, and in general, the important schemata for interconnection of impressions, the perception of good Gestalten, formal properties of instinctive behavior, formal properties of primary process thinking<sup>(3)</sup> in the genesis of thoughts, and certain other features which seem characteristically mental. Although there is no way at present to tell whether such formal properties of the model would appear in the same situations in which their mental analogues appear, the existence of so many analogies to interesting forms of behavior makes it appear worthwhile to continue studying the model in the hope of making further identifications with mental processes. The present discussion will not reexplain the previously presented analogies, but will confine itself to two separate topics, the first part considering the motivation for the vector-space model, and the second part indicating a possible physical realization of the abstract structures of that model.

## I

This section consists of some general observations and examples concerning behavior which are intended to supplement the arguments which led to the previously cited model. These observations on the active role of the nervous system in perception and other behavior help motivate the fundamental assumption of the abstract model that information is represented in the form of linear superpositions of certain basic functions in a function space.

Review of the literature of developmental and comparative psychology reveals the impressive extent to which the animal reacts

to its surroundings by fitting to the surroundings certain patterns of activity which it is inherently able to perform, or which it has developed through previous activity. It does not seem to start with a homogeneous network upon which it depicts experiences; rather, experiences seem to excite certain patterns inherent in the network, which are then reshaped to fit better. On the microscopic level, this is certainly how the nerve impulse is used, and on a super-organismic level, there is evidence<sup>(4)</sup> that the complex structures emerging in biological evolution result from the improving effects of natural selection acting upon structures which were already present and which became important, for example, when the animal moved to a new habitat. However, the best examples for the present purpose come from the level of the ordinary behavior of the animal.

As an example, the sea anemone has a complicated pattern of activity which may be changed by stimulation. A brief exposure to food leads to a prolonged series of changes associated with feeding, digestion and elimination. But, according to Pantin,<sup>(5)</sup> the animal will sometimes run through the whole sequence without any evident external stimulation. Thus the complex feeding behavior of the anemone appears to result from the triggering of a pattern which it can perform independently of any food, but which becomes useful in the right situations. In general, according to Pantin, stimuli may best be regarded as causing the anemone to shift from one of its intrinsic patterns to another.

The work of Lorenz, Tinbergen and others has shown that instinctive behavior in general appears to depend upon the excitation of analogous innate mechanisms by specific "releasing" stimuli; and the role of learning in instinctive behavior, according to Thorpe,<sup>(6)</sup> is to make fine adjustments in the set of releasers so as to enable the animal better to meet the complexities of its environment.

Another example of the active role of the nervous system in constructing the units of perception prior to their fitting and adjustment to the environment comes from the way children perceive and represent shapes by drawing them. For instance,<sup>(7,8,9)</sup> a young child may respond to visual patterns by "drawing" a circle, square and cross all as unintelligible scribbles. Later, the circle and square come to be represented by roundish scrawls,

while the cross is represented by slashing scribbles. Next the swirling that represents circle and square becomes sharpened into a single circular outline for each. This same outline serves to represent any closed figure. At this stage, the child is good at representing closure or openness of curves and relations of contiguity. At a still later age, the circle is still represented by a circular outline, while the square is often represented by a circular outline with four short lines intersecting the outline at one place or another.

This behavior is what one might expect if the child's perceptual and motor networks have naturally occurring swirling and slashing modes all over them. These might be built in from the start, or they might arise from motor activity and from the passage of light over the retina during natural movements of the child. Then the circle and square are more likely to excite the swirling modes than the slashing modes, while the cross excites the slashing modes. When the child is a little older he fits these modes to the pattern by suppressing most of the overlaid swirls or slashes. When he has developed still further and, let us say, can perceive the finer aspects of things, he is sensitive enough for the square (which has straight sides and relatively sharp corners) to excite a second component, the slashes, so he puts them in the picture. He puts them close to the swirl because he is sensitive to contiguity (a natural property of neural nets), but he does not integrate both modes into the same structure (closed *and* sharp) as he must if he is accurately to represent the square.

This independence of modes is also seen in the independence of innate releasing stimuli in the theory of instinct. These specific characteristics, which when perceived set off a behavior pattern, are generally completely independently acting stimuli which may easily be described in words as independent characteristics of the environment.<sup>(10)</sup>

We may say that it appears that the child or animal has certain representational or motor patterns which are used, if they happen to be useful, in reacting to the environment. The child seems to draw something by letting it excite the closest inherent patterns which he happens to possess. Then he begins to sharpen the result to make it fit better. Ideas like these suggest that a perceptual machine might work better if it followed a stage of processing that



contains such excitable structures. Then a circle would be perceived anywhere in the network because anywhere in the network such a swirling mode could be excited. The big problem is to find the proper compromise between a network which has nothing built in, which may not be able to develop very far because it has too little to build upon, and a network with all sorts of specific detectors built in, which may react to specialized stimuli but not perceive in any reasonable sense of the word. A very relevant investigation in this connection is that of Lettvin *et al.*,<sup>(11)</sup> who showed that the message transmitted through the optic nerve of a frog is already organized into such "perceptual" information as the location of convex objects in the visual field.

Now in the formal model it was pointed out that the natural thing to consider in relation to coding of perceptual information as modes in a network, as motivated by the above behavioral examples, is the superposition of patterns or, conversely, the resolution of complicated patterns into linear combinations of basic patterns. The second part of this paper will suggest a possible answer to the problem of where these basic modes might come from. Here, a few general remarks may indicate the kind of behavior which served to motivate the ideas of superposition and resolution of perceptual representations. The idea of basic patterns to be superposed came from the examples of instinctive behavior and perceptual-motor representation discussed above and, at least formally, the paper on frog vision<sup>(11)</sup> suggests that the sensory receptors may encode and transmit certain aspects of the incoming signal which may be represented as an ordered  $n$ -tuple of independent intensities. At the behavioral level, a response such as flight from an object perceived as a predator may depend, as we have seen, upon the perception of independent characteristics.<sup>(10)</sup> An action may be a superposition of more elementary actions, as in the commonly superposed tendencies to approach and to flee in the fighting and mating behavior of birds and fishes.<sup>(12,13,14,15)</sup> Facial expressions in canines may sometimes be analysed into the combinations of two or more independent components present in various intensities.<sup>(15,16)</sup> Finally, a dream image may be a condensation of two or more images.<sup>(3)</sup> Whether or not these phenomena might be subsumed under linear superposition in a technical, nonmetaphorical sense is at present unknown, but the



possibility seems worth investigating. It is, in any case, worthwhile to learn just how much such mechanisms *could* account for, and in what way they might do so, for the purpose of suggesting experiments and interpreting experimental results. The formal model based upon these ideas predicts some unsuspected possibilities for these mechanisms, and for the functions of such simple things as purely random excitation. When a physiological effect in the brain is newly discovered, one would like to be able to relate it to behavior, and the theoretical exploration of possibilities is an aid to knowing what to look for in investigating physiological effects.

## II

The formal model was an outgrowth of a cell assembly viewpoint about brain function, although it was previously presented in connection with its possibilities for the construction of intelligent machines. People talk about cell assemblies. These are complicated patterns of oscillation. That is all people can say. Instead of saying that an idea is coded as a "complicated pattern", the formal model supposes that it is a superposition of "modes of oscillation" (or, as we shall suggest, other types of basic vectors), which is at the least no more vague, and carries with it certain associations about what these modes can do. Supposing that to be the case, I considered the vector space of these modes, or orthonormal functions, and said that an idea was coded as a vector, with its components in various coordinate systems (specified by various orthonormal sets of vectors) representing various points of view. Information was to be elicited by a process which split the idea into its components from one point of view and selected one or more of the components according to a probabilistic process which involved the correlation of the mode with a sample of shot noise. The process was motivated by psychological considerations, and led in a way which will not be described here, to properties which, as mentioned above, seemed identifiable with interesting psychological phenomena involving the integration of complex patterns in perception and thought.

Two of the main questions in regard to this model were: (1) Where might these orthonormal systems come from? Some physical process must be capable of generating orthogonal sets of

vectors in such a way that the coefficients in the transformation equations between these sets had useful properties related to the model. (2) The vector space probably must be complex-valued in order to obtain a number of interesting properties; and the shot noise must be complex-valued with independent real and imaginary parts for the selection process to work at all. How might these requirements be realized in a physical system?

The desirability of the complex vector space, together with the considerations regarding networks which follow, suggested the following idea. Periodic phenomena may be described by complex numbers, but this is commonly done with the convention that only the real part is interpreted as having physical meaning, so that one does not retain two independent components, and moreover, linear operations alone give answers consistent with the convention. However, the transfer function of a component (the ratio, as a function of frequency, of the output at a given frequency to the input at that frequency) may be represented by a complex number which not only has independently significant real and imaginary parts, but which may legitimately be used in multiplicative relationships. For instance, transfer functions multiply along paths in a network: the transfer function of a path is the product of the transfer functions of its segments. Is there then some way in which the complex vectors required by the formal model might be realized not by the inputs to the network, but by transfer functions in the network? In vague terms, which are considered as no more than possibly suggestive, the Gestalt-like properties, etc., lie not in the inputs, but in the way the net handles the inputs.

Now it happens that information-processing models of other authors sometimes employ networks having transfer-function matrices which can satisfy the formal requirements of the present model. The remainder of this paper will show how one such type of network might be regarded as a realization of the present abstract model. It is not known at present whether such a realization is useful; both the formal model and the processes to be discussed here might very well turn out to be useful individually, while they nevertheless yield no new advantage in combination. The value of the discussion lies chiefly in the fact that it provides something more concrete to think about in attempting to identify

the model with psychological behavior. The abstract model in its present state is not claimed to be a contribution to our knowledge of behavior. It is proposed solely as an example to show that a mathematical model can have properties which may be discussed in the same terms as a number of fundamental problems in the synthesis of thought which have not yet been statable in a form which could be adequately understood or rigorously investigated. Any instance of a concrete system satisfying the postulates of the model is potentially useful because it gives us a chance to learn how the model works in specific situations, and perhaps to clarify our hazy notions of the mind by seeing in which ways our system falls short of what we would call true thinking. By learning the kinds of things that may be said about the model in these situations, we may hope to develop concepts with which to investigate these elusive topics.

The remainder of the paper will discuss a realization of the model which utilizes a network which can perform a mathematical operation known in multivariate statistical analysis as the transformation to principal components.<sup>(17)</sup> This transformation is often useful in reducing the amount of data which must be considered in a complicated situation, and can sometimes be used in communication systems to reduce the number of channels required to carry a number of messages. A natural way of reducing the number of statistical variables which need to be considered as contributing significantly to a given problem is to transform linearly from the coordinate system in which these variables lie to another coordinate system in which the transformed variables are independent, and in which perhaps only a few of the transformed variables have large variances. Then if one is interested in individual differences in the population, only those few linear combinations vary enough from one member of the population to the next to require notice, while all the rest may be ignored for some purposes. To accomplish this, one applies the unitary transformation which diagonalizes the covariance matrix of the original variables, so that the columns of the conjugate of the transformation matrix will be the eigenvectors of the covariance matrix. If these columns have been ordered so that their associated eigenvalues are arranged in order of decreasing magnitude, then it is well known that the first transformed component will be the

normalized linear combination with maximum variance, and each succeeding component will have maximum variance of all normalized linear combinations uncorrelated with each previous component. The new variances will be equal to the eigenvalues, which are the diagonal terms of the transformed covariance matrix.

In transmitting a set of correlated signals, one may take advantage of the redundancy entailed by the correlations by performing a transformation to principal components, discarding any output channels which are relatively inactive, transmitting the remaining reduced number of messages, and finally transforming back to the original coordinates to reconstitute the message.<sup>(18)</sup> As a corollary of the preceding paragraph, one sees that for a given number of discarded channels, a smaller mean square discrepancy between transmitted and received signals is achieved by the transformation to principal components than by any other linear coding system. Moreover, if the signals are Gaussian, the total channel capacity, in the information-theoretical sense, required to meet a given mean square error criterion is also minimized by the transformation to principal components.

The rank of the covariance matrix is diminished by the number of independent linear relations among the variables, so that inactive variables in the principal component system serve as detectors of linear relations.

It is clear that any process which involves a unitary transformation to a coordinate system in which some matrix is diagonal has something in common with the behavioral model under consideration, because this model is based entirely upon transformations between various bases with respect to which various operators are diagonal. That this mathematically trivial connection may be of value in the exploratory discussion which has been advocated above is suggested by some investigations of M. C. Goodall.<sup>(19,20,21,22)</sup> He considers, as a model for some rudimentary cognitive functions, the possibility of constructing networks which can automatically perform the transformation to principal components or to some less specific orthogonal coordinate system, transform correlated inputs into independent outputs, and in some cases unscramble messages which have been mixed together. Such nets detect the presence of linear relations, as explained above, and



might be applicable to the recognition of group invariants by finding invariant forms, that is, by detecting the presence of linear relations among certain monomials fed into the net. Two examples of Goodall's approach which are relevant to the present discussion will be briefly described in part.

First we shall see how a network can transform a set of correlated inputs into a set of independent outputs, or what is equivalent, express a set of inputs as a set of linear combinations of orthonormal output functions. If the  $n$  input functions of time are  $x_i(t)$  and the  $n$  output functions are  $y_k(t)$ , where

$$\langle y_k(t)y_j^*(t) \rangle \equiv \int y_k(t)y_j^*(t) dt = \delta_{kj}$$

and stars denote complex conjugates, and if

$$x_i(t) = \sum_k A_{ik}y_k(t)$$

then  $A_{ik} = \langle x_i(t)y_k^*(t) \rangle$ . The transfer matrix of the net will be  $A^{-1}$ , where  $A$  is the matrix  $(A_{ik})$ . If  $C$  denotes the covariance matrix of the inputs,  $C_{ik} = \langle x_i(t)x_k^*(t) \rangle$ , then it is clear that  $C = AA^*$ , where  $A^*$  is the hermitian conjugate of  $A$ . It is also clear that given a set of inputs  $x_i(t)$ , if we can find a matrix  $A$  such that  $AA^* = C$ , and if we set the transfer matrix equal to  $A^{-1}$ , then the outputs  $y_k(t)$  will be orthogonal, and we shall have  $A_{ik} = \langle x_i(t)y_k^*(t) \rangle$ . Goodall proposes to approach this solution through a relaxation process in a network having variable gains. He starts with an arbitrary transfer matrix  $A(0)^{-1}$ , and lets  $A_{ik}$  vary in the direction of  $\langle x_i(t)y_k^*(t) \rangle$  according to an equation like

$$\tau dA_{ik}(t)/dt = \langle x_i y_k^* \rangle - A_{ik}(t)$$

in which the  $x_i(t)$  are supposed to be approximately stationary over a time interval considerably larger than the averaging time. Replacing  $y$  by  $A^{-1}x$ , and using the definition of  $C$ , we have in matrix notation,

$$\tau dA(t)/dt = CA^{*-1}(t) - A(t).$$

Multiplying this equation on the right by  $A^*(t)$ , multiplying the



conjugate of this equation on the left by  $A(t)$ , and adding the two resulting equations, we have

$$\tau(d/dt)[A(t)A^*(t)] = 2C - A(t)A^*(t)$$

the solution of which is

$$A(t)A^*(t) = C - e^{-2t/\tau}[C - A(t)A^*(t)].$$

Thus an arbitrary nonsingular initial transfer matrix always converges to one which yields independent outputs.

In order to obtain a transfer matrix which is the inverse of the matrix which is varied, Goodall employs negative feedback. Suppose that a linear network has a forward transfer matrix  $F$ , and that the outputs are fed back through a transfer matrix  $G$  and the resulting values subtracted from the inputs to  $F$ . Then the closed loop transfer matrix is  $F(I+GF)^{-1}$ , where  $I$  is the identity matrix. Since we wish the transfer matrix to be  $A^{-1}$ , where  $A$  is made to approach the matrix  $\langle x_i y_k^* \rangle$ , we may let  $F = I$  and take  $G = A - I$ , so that  $G_{ik}$  is made to approach  $\langle x_i y_k^* \rangle - \delta_{ik}$ .

It may be noted in passing that Milner,<sup>(23)</sup> in his discussion of hypothetical cell assemblies, speculates that the inhibitory feedback synapses grow in strength with use, and one version of this growth could depend, as here, upon the correlation between input and output. In addition, he speculates that in the case of the one-to-one correspondence of inputs and outputs which we have here, the inhibitory path from an output to its corresponding input is weaker than the inhibitory cross-connection to other inputs. Milner supposes that this is the case in order that an assembly may not shut itself off; the preceding paragraph derives such a requirement on the inhibitory strengths from a not entirely equivalent starting point.

If instead of taking the forward transfer matrix  $F = I$ , we take  $F$  equal to a very large scalar multiple of  $I$ , then the closed loop transfer matrix is approximately equal to  $G^{-1}$ , and in this case we may let  $G_{ik}$  approach  $\langle x_i y_k^* \rangle$ . Since for our purposes there is no fundamental significance in the more complicated expression of the preceding discussion, and since there is no difficulty in changing from one expression to the other in any result we may obtain, we shall hereafter consider the forward transfer matrix to be the inverse of the feedback transfer matrix. This will make

it possible in what follows to have inverses for transformations between coordinate systems.

When the network described above has transformed a given set of inputs into a set of orthogonal inputs, the feedback transfer matrix consists of the coefficients of the inputs expressed as linear combinations of the outputs; thus we may say that the net expands the inputs in terms of an orthonormal set determined by those inputs and the initial state of the net. We may modify this situation by forcing the outputs to assume the values of some independently chosen orthogonal set of functions. Then the feedback transfer matrix will converge to the expansion coefficients of the input functions in terms of the orthogonal set clamped on the outputs (if these coefficients exist). Thus we may express vectors in various orthonormal coordinate systems, as required by the abstract behavioral model.

Since a matrix  $A$  satisfying  $AA^* = C$  is determined only up to unitary transformations, we see that the possible orthogonal output sets corresponding to a given input set differ by unitary transformations, and are determined by the initial state of the net. These orthogonal sets may be used for the expansion of other sets of vectors, and thus provide a link with the premise of the abstract model that the particular coordinate system in which a given vector is expressed will depend upon some sort of point of view, which means, as occurs here in one form, upon the initial state of the observer.

A coordinate system of particular interest is, as we have seen in the discussion of principal components, that in which the covariance matrix ( $\langle x_i x_k^* \rangle$ ) becomes diagonal. To perform the transformation to this system, Goodall introduces the second mechanism which will be mentioned here. Suppose that  $C = V^* \Lambda V$ , where  $V$  is unitary and  $\Lambda$  is diagonal. Then if in vector notation  $y = Vx$ , we have, as we expect from the discussion of principal components,

$$\langle yy^* \rangle = \langle Vxx^*V^* \rangle = VCV^* = \Lambda$$

and furthermore, since

$$\Lambda V = VC = \langle Vxx^* \rangle = yx^*$$

we have

$$V = \Lambda^{-1} \langle yx^* \rangle$$

so that

$$V_{ik} = \langle y_i x_k^* \rangle / \langle |y_i|^2 \rangle.$$

Conversely, if  $V$  is a transfer matrix which approaches the last expression, then provided that  $V$  is unitary, the network will come to produce the transformation to principal components. Goodall<sup>(21)</sup> describes a feedback relaxation process intended to accomplish this, although it is not clear that  $V$  will necessarily become unitary, and the stated definition becomes meaningless whenever  $C$  is singular. However, since our present purpose is to learn what could be done with such a network (and since we could assume for this purely conceptual purpose that one component of the network was a computer which diagonalized matrices) we shall assume that the network will actually perform the desired transformation.

After this lengthy introduction to Part II, which had to outline what was originally to have been discussed by Mr. Goodall in this symposium, we may proceed to the main business of linking these networks to the abstract behavioral model. We recall<sup>(2)</sup> that we would like to find some way to define vector spaces in which we may refer vectors to any of a number of orthonormal bases. The vector space may in some cases have to be complex-valued (for reasons which will not be discussed here) in order to obtain interesting results. The calculation of expected values of variables which identify various states of the system which we should like to define and be able to detect led to the introduction of operators defined on these vector spaces, and we were interested in coordinate systems in which various operators were diagonal. The unitary matrices which transferred from one coordinate system to another had the significance that the squared magnitudes of their elements represented certain conditional probabilities of detecting states of the system. Finally, the way in which these probabilities were derived from the matrix elements involved taking the scalar product of the information vectors with a random vector (which had to be complex-valued for the procedure to work). To date there has been no suggestion as to a conceivable origin of the coordinate systems, and this has left the model completely up in the air. Now we shall show that the networks described above could provide a realization of any of the elements of this model.

One way in which we may obtain orthonormal coordinate systems is through the use of Goodall's network which expresses a set of input functions as linear combinations of orthogonal output functions. We have seen that the particular orthogonal set which comes into existence depends both upon the inputs and upon the initial state of the network, which latter we might hope to identify with the "points of view" in the model. If this assignment of meaning to orthogonal sets can be done in a significant way, then the previous discussion about expressing inputs in terms of independently chosen orthogonal sets clamped on the output terminals will mean that we may examine stimuli by expressing them in terms of any of a number of meaningful orthogonal sets which are stored in the memory (or which can be produced by something stored in the memory). Since the effect of these networks is to undo the correlation between inputs, expressing one input set in the coordinate system determined by the action of the network on another input set reveals in a crude sense how the two sets differ by showing what kind of dependence is left in the first set after one has cancelled out the kinds of dependence existing in the second set.

There is nothing more to add in this case to the purely mechanical task constituting this paper, because we have vectors (inputs and outputs), orthogonal bases (outputs) and operators (network transfer matrices), although no situation is specified in which some significant operator becomes diagonal. Unfortunately, as previously pointed out, no simple way is immediately suggested to make the vector space complex, and this may be a drawback. However, the random noise by which the components of the vectors must be multiplied could be realized in amplifiers with random gains, and these gains may be represented in a natural way by complex numbers.

We shall discuss more fully the case in which vectors are given by the rows and columns of the transfer-function matrix. In order to provide reciprocal bases we shall require the inverses of these matrices. Inverting a unitary matrix requires only transposition and conjugation, but we can invert an arbitrary matrix by using a feedback network with large forward gains. The overall transfer matrix will, as explained earlier, be approximately the inverse of the feedback matrix. This technique is used in the first type of



relaxation network, and may possibly be used in the second type too. Since the whole program of investigation of the behavioral model at present may be regarded as a conceptual exercise in seeing what kinds of things may be extracted from a small set of ingredients, we shall assume in the remainder of the discussion that each transfer matrix occurs in the feedback path of a network which will invert it. This discussion will apply to any such network which performs unitary transformations, including in particular the second type of relaxation feedback network, which performs unitary transformations diagonalizing input covariance matrices. All such networks will be called unitary networks. We are interested in unitary networks because we would like to define our vectors in terms of transfer functions so they may be complex, and if the vector components are defined as the rows or columns of the unitary transfer matrix, the vectors will be orthonormal. The ways in which unitary matrices may be obtained include letting an arbitrary matrix  $A$  vary according to  $\tau dA/dt = A^{*-1} - A$ , by letting the first type of relaxation network expand one orthogonal basis in terms of another, or by letting the second type of relaxation network diagonalize the input covariance. We already know that the second and third of these ways are useful in their own right, and they may both be used in the behavioral model, if we include covariance matrices among the matrices which are to be diagonalized according to the model.

The rows of the unitary transfer matrix are orthonormal and may be taken as the basis of a coordinate system required by the behavioral model. The elements of the  $k$ th row are the transfer functions of the forward paths leading to the  $k$ th output. Under our assumptions there are feedback paths having a transfer matrix which is the inverse of the forward matrix. The orthonormal columns of this matrix are considered to form vectors in the dual space of the space spanned by the forward vectors, as will be explained in greater detail.

We may think of the same set of input elements connected to a number of unitary networks. The orthonormal vectors in each network will have been specified in some useful way. For instance they may have been permanently determined by the covariance matrices of inputs at some time in the past, or else we may consider some of the nets to be constantly changing, so that signals



are analysed in the principal axis system of what has gone just before.

The identification of unitary nets with the behavioral model is simply a matter of finding notation in which to make a description of what the network does in terms of unitary space. We need the following: a basis, vectors expressed in terms of this basis, a conjugate space of functionals on the given space, an inner product, and operators on the space. Following standard notation and terminology,<sup>(24)</sup> the inner product of vectors  $x$  and  $y$  will be written  $(x, y)$ , where  $(y, x) = \overline{(x, y)}$  and  $(ax, y) = a(x, y)$ , so that  $(x, ay) = \overline{a}(x, y)$ . If  $y'$  is a functional, then  $y'(x)$ , regarded either as a functional of  $x$  or of  $y'$ , will be written  $[x, y']$ . The adjoint  $A'$  of an operator  $A$ , is defined by  $[Ax, y'] = [x, A'y']$ , and if  $y$  is the vector in the unitary space such that  $[x, y'] = (x, y)$ , then corresponding to the adjoint  $A'$  we have  $A^*$ , such that  $(Ax, y) = (x, A^*y)$ . We know that if we are given the matrix of  $A$  with respect to a basis, then the matrix of  $A'$  with respect to the dual basis will be its transpose, while the matrix of  $A^*$  correspondingly becomes the hermitian conjugate of the matrix of  $A$ . We have  $(aA)' = aA'$ , while  $(aA)^* = \overline{a}A^*$ . Finally, an inner product is defined upon the dual space in the customary way: if to  $y_1'$  and  $y_2'$  in the dual space correspond  $y_1$  and  $y_2$  in the original space, then  $(y_1', y_2') = (y_2, y_1)$  by definition. In the notation of the formal model, namely the Dirac notation<sup>(25)</sup> for the same things, vectors are  $|x\rangle$ , their conjugates  $\langle y|$ , and inner products  $\langle y|x\rangle$ , which means the same as  $(x, y)$  above.  $(Ax, y)$  is written  $\langle y|A|x\rangle$ . Thus, the  $ij$ th element of the matrix of  $A$  in the  $x_k$  basis is  $\langle x_i|A|x_k\rangle$ . The  $k$ th coordinate of a vector  $v$  in the  $x$  system is  $\langle x_k|v\rangle$ . We transform from one coordinate system to another by using the operator identity

$$\sum_k |x_k\rangle\langle x_k| = 1$$

where the summation is over the elements of any orthonormal basis.

Thus

$$|v\rangle = \sum_k |x_k\rangle\langle x_k|v\rangle = \sum_{k,m} |u_k\rangle\langle u_k|x_m\rangle\langle x_m|v\rangle, \text{ etc.}$$

and

$$\langle x_i|A|x_j\rangle = \sum_{m,n} \langle x_i|u_m\rangle\langle u_m|A|u_n\rangle\langle u_n|x_j\rangle, \text{ etc.}$$

In general, a vector  $v$  is an entity defined independently of any coordinate system; what we observe as numerical coordinates (or "representatives") will be its expansion coefficients ( $x_k|v$ ). Similarly, an operator is independent of a basis, but we use one of its matrices  $\langle x_i|A|x_j\rangle$ . The above discussion just recalls the standard definitions and notation, but is presented in such detail partly to explain the Dirac notation in the hope that it will prove convenient for the reader who may desire to explore the behavioral model, but mainly because the following identification of unitary feedback networks with the model, while mathematically trivial, is confusing unless the notation is kept straight.

We shall be saying nothing physical about the net (with one exception)—we merely say that out of the many things we can say about the network, one of them is in the language of the formal model. There are a number of ways of trying to define the necessary entities. For instance, the vectors could be either all paths to a particular output, or all paths from a particular input. The feedback paths can likewise be looked at two ways. If a vector is represented by all forward paths to a point, its dual might be all feedback paths back from that point, or all forward paths from some input point. Out of these possibilities we must find a combination that satisfies the definitions reviewed above, and such a combination will be described. Once you have waded through such dull reading, you will see how the elements of the formal model may be represented by network diagrams, and thereafter, the diagrams, which are easy to draw, can substitute for all the notational details.

In the diagrams which follow solid lines indicate forward paths and broken lines indicate feedback paths. All forward paths go roughly left to right, and all backward paths right to left. (This says nothing about the networks, but only about the direction on the page in which they will be drawn.) A lower-case Roman letter will designate the transfer function of a forward path, and the corresponding Greek letter the transfer function of the reverse path between the same pair of terminals. A superscript indexes a destination of a path, and a subscript indexes an origin. If a letter has both sub- and superscript following it, then the left(right)-hand one indexes the left(right)-hand end of the path in the conventional diagram just described.

With these conventions we can construct diagrams for each element of the model:

1. *The general model*: As an example, Fig. 1A shows three unitary nets connected to the same input terminals, without attempting to indicate all the paths in each net.

2. *Basis*: A typical basic vector is indicated by  $t^k$  in Diagram B. This vector is composed of all forward paths toward the  $k$ th output terminal. The  $j$ th coordinate of vector  $t^k$  is  $t_j^k$  (note conventions for placement of  $j$  and  $k$ ). The whole basis is the set of all such vectors as in Diagram C.

3. *Scalar multiple and linear combination*: In Diagram D the vector  $at^k$ , where  $a$  is a scalar gain represented by the triangle, corresponds to the set of all transfer functions from the left-hand terminals indexed by  $j$  to point  $P$ . The linear combination

$$\sum_k a_k t^k$$

corresponds to the similar set of transfer functions in Diagram E.

4. *Linear functional*: A set of backward paths, such as the broken lines originating from point  $A$  in diagram F, which may be connected to the paths of a forward vector (or linear combination), such as the solid lines terminating at point  $B$ . The value of the functional in the case illustrated is the transfer function from  $A$  to  $B$ . We see that the functional may be regarded either as a functional of  $x$  or of  $y'$ .

5. *Dual basis*: A vector in the dual basis is a set of feedback paths from a point. If we have a basis composed of vectors such as  $t^k$  in Diagram G, then a typical vector in the dual basis is  $\tau_i$ , having components  $\tau_i^j$  (again note conventions for placement of  $i$  and  $j$ ). Since the backward matrix is the inverse of the forward matrix, we have  $[t^k, \tau_i] = \delta_{ik}$ , which is the definition of a dual basis. This relation is depicted in Diagram H, for the transfer function between  $i$  and  $k$  is

$$\sum_j \tau_i^j t_j^k = \delta_{ik}.$$

Scalar multiples and linear combinations of dual vectors are defined analogously to their definitions for vectors in the original space, and are depicted in Diagrams I and J.

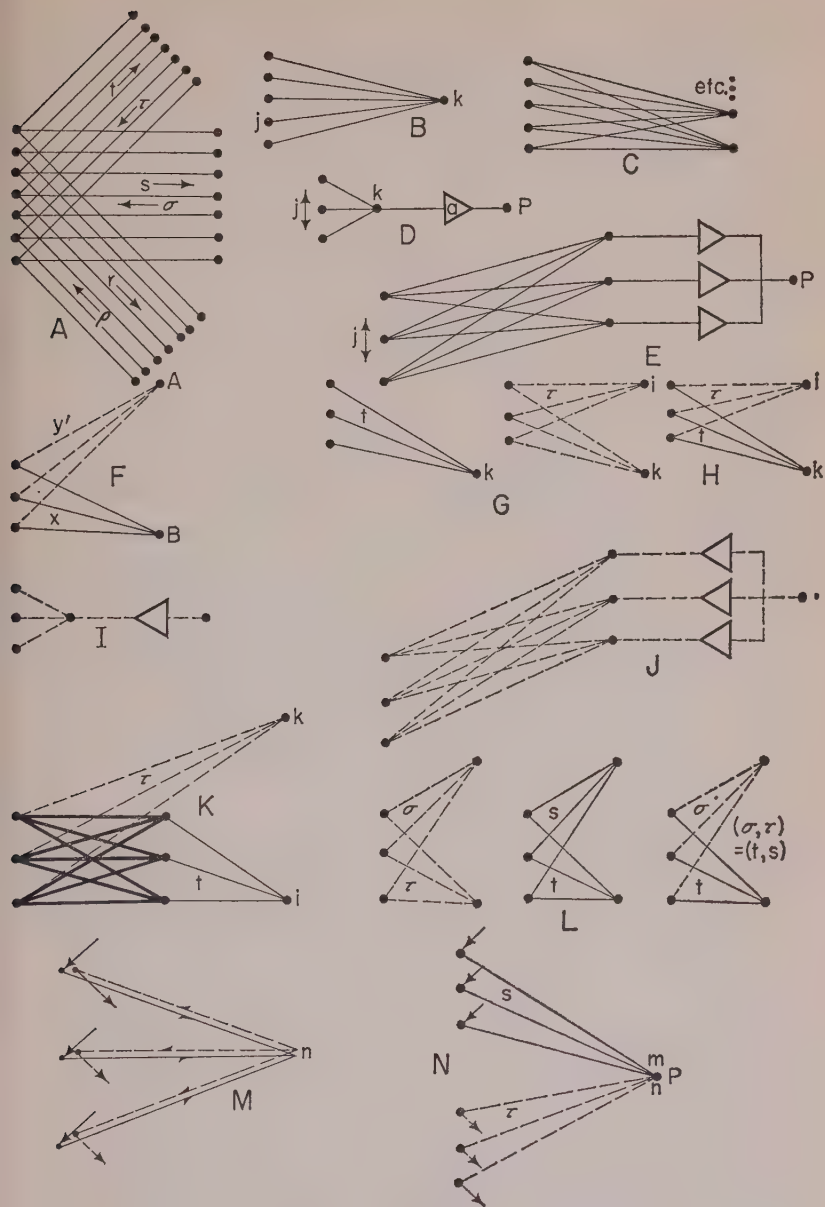


FIG. 1

6. *Inner product*: To decide how to define this, we observe that an orthonormal set must be self-dual. Therefore we should have  $(t^k, t^i) = \delta_{ik}$ . We might naturally try to define

$$(t^k, t^i) = \sum_j \bar{t}_j^i t_j^k,$$

and this equals

$$\sum_j \tau_j^i t_j^k,$$

which does equal  $\delta_{ik}$ , since in a unitary net  $\bar{t}_j^i = \tau_j^i$ . Therefore we interpret

$$(t^k, t^i) = \sum_j \tau_j^i t_j^k,$$

which is the transfer function from  $i$  to  $k$  in Diagram H. Thus the basis dual to the set  $\{t^k\}$  is the set  $\{\tau_i\}$ , but when we introduce the inner product, the set  $\{t^k\}$  becomes self-dual, as it should. We may replace either factor in the inner product by a linear combination of vectors in the natural way if we make the following provision. In order that  $(x, ay) = \bar{a}(x, y)$ , we must assume that the transfer function going backwards through a gain such as the triangular component in Diagram D is the complex conjugate of the transfer function going forward. This is the above-mentioned sole physical assumption made in the model, and is not a restriction at all if the transfer functions are real.

7. *Dirac notation*: In Diagram G,  $t^k$  would be the "ket"  $|k\rangle$  and  $\tau_i$  would be the "bra"  $\langle i|$  (dual to the "ket"  $|i\rangle$ ). Their inner product is obtained by combining them to form a bra(c)ket  $\langle i|k\rangle$ . This notation, which is very convenient when working with transformations involving a number of different bases (because the notation automatically carries out some of the mathematics and because the brackets leave room to indicate which bases the operators are being referred to), is also suggestive of the present diagrams. If we read the brackets from left to right, a ket has lines converging to a point and a bra has lines diverging from a point, and that is just how the diagrams look. Elements of matrices which represent operators look right too. For instance  $\langle k|A|i\rangle$  is shown in Diagram K.

8. *Operators and their adjoints*. The heavily drawn network in



Diagram K may be considered to represent an operator  $A$  operating on vector  $t^i$ . The transfer function from  $k$  to  $i$  represents either of the two equivalent expressions  $[At, \tau] = [t, A'\tau]$ . Thus we may regard the black part either as  $A$  or as  $A'$ , depending upon whether we multiply  $\tau_k$  by  $A'$  first (row vector times matrix on the right) and then take the scalar product with  $t^i$ , or else take the product of  $\tau_k$  with  $At^i$  (column vector times matrix on the left). The matrix of  $A'$  is the transpose of the matrix of  $A$ .

To interpret  $A^*$ , the hermitian conjugate of  $A$ , we must reverse our direction through the heavily drawn network. Just as the conjugate of a vector of forward paths converging to a point is a vector of backward paths diverging from a point, i.e. the same network traversed backwards, so the conjugate of the net representing  $At^i$  is the same network traversed backwards, and if we make the physical assumption that reversing direction in the operator network turns a gain into its complex conjugate, then the matrix of  $A^*$  is the hermitian conjugate of the matrix of  $A$ . (Remember that the indices become transposed according to our conventions because source and destination are interchanged.)

9. *Inner product of dual vectors*: In keeping with the previous definitions, we define  $(\sigma, \tau) = (t, s)$ , as depicted in Diagram L.

10. *Dyad*: This is just a combination of elements that have been previously defined. The dyad  $t^n\tau_n$ , or  $|t^n\rangle\langle t^n|$  is represented by Diagram M, from which we see that

$$\sum_n |t^n\rangle\langle t^n| = I.$$

A general dyad,  $s^m\tau_n$ , or  $|s^m\rangle\langle t^n|$ , is given by Diagram N, in which point  $P$  is simultaneously the  $m$ th terminal of  $s$  and the  $n$ th terminal of  $\tau$ .

We are now in possession of all the mathematical structures required by the formal model. As an example of the use of such networks, we shall conclude with a network diagram corresponding to the probabilistic selection of vector components. For reasons explained in the papers which presented the formal model, the utilization of the information represented by a vector required resolving the vector in some coordinate system and then picking any of the resulting components with a probability proportional to the squared magnitude of its expansion coefficient in that coordinate system. As explained in the exposition of the model,

this was accomplished by a process which followed ideas of N. Wiener involving random noise. In the notation which was used, the information bearing vector was the function  $\psi(t)$  in a function space, which might be expanded in terms of the basis  $\phi_1(t)$ ,  $\phi_2(t)$ , etc., so that

$$\psi(t) = \sum_n a_n \phi_n(t).$$

Then one computed the numbers

$$A_n = \int a_n \phi_n(t) dY_\alpha(t),$$

where  $dY_\alpha(t)$  is the instantaneous output of a complex-valued shot noise generator,  $\alpha$  being the parameter indexing the ensemble of possible random noises. Finally, one chose that  $\phi_n(t)$  for which  $|A_n|$  happened to be the largest of any of the  $A$ 's. This method, initially introduced solely because it works, turned out unexpectedly to lead to phenomena suggestive of psychological behavior. In the bracket notation, we may regard  $dY_\alpha(t)$  as the  $\alpha$ th basic vector of a random noise vector space and write

$$|\psi\rangle = \sum_n |\phi_n\rangle \langle \phi_n | \psi \rangle,$$

so  $a_n = \langle \phi_n | \psi \rangle$ , and  $A_n = \langle dY_\alpha | \phi_n \rangle \langle \phi_n | \psi \rangle$ . That is,  $A_n$  may be regarded as the  $\alpha$ th noise component of the  $n$ th  $\phi$  component of  $\psi$ . Now one has only to use the definitions which we have introduced to read off the structure of the net from the expression  $\langle dY_\alpha | \phi_n \rangle \langle \phi_n | \psi \rangle$ , making use of three unitary nets. Reading this symbol from left to right, we see as in Fig. 2 that one starts from a terminal  $P$  in a noise generating net (the gains of the diverging lines at some moment represent the  $\alpha$ th noise function), proceeds backwards through the noise net, forward to the  $n$ th terminal of the  $\phi$  net, then backwards through the  $\phi$  net, and finally forwards through a net representing  $\psi$  to point  $Q$ . The transfer function from  $P$  to  $Q$  along this path is the value of the desired expression. Such values are compared for all  $n$ , and the  $\phi_n$  giving the largest magnitude to the gain is selected.

Thus we have achieved the aim of this paper, which is to suggest something which, at least in principle, is a realization of a large part of the formal model for behavior that was presented in

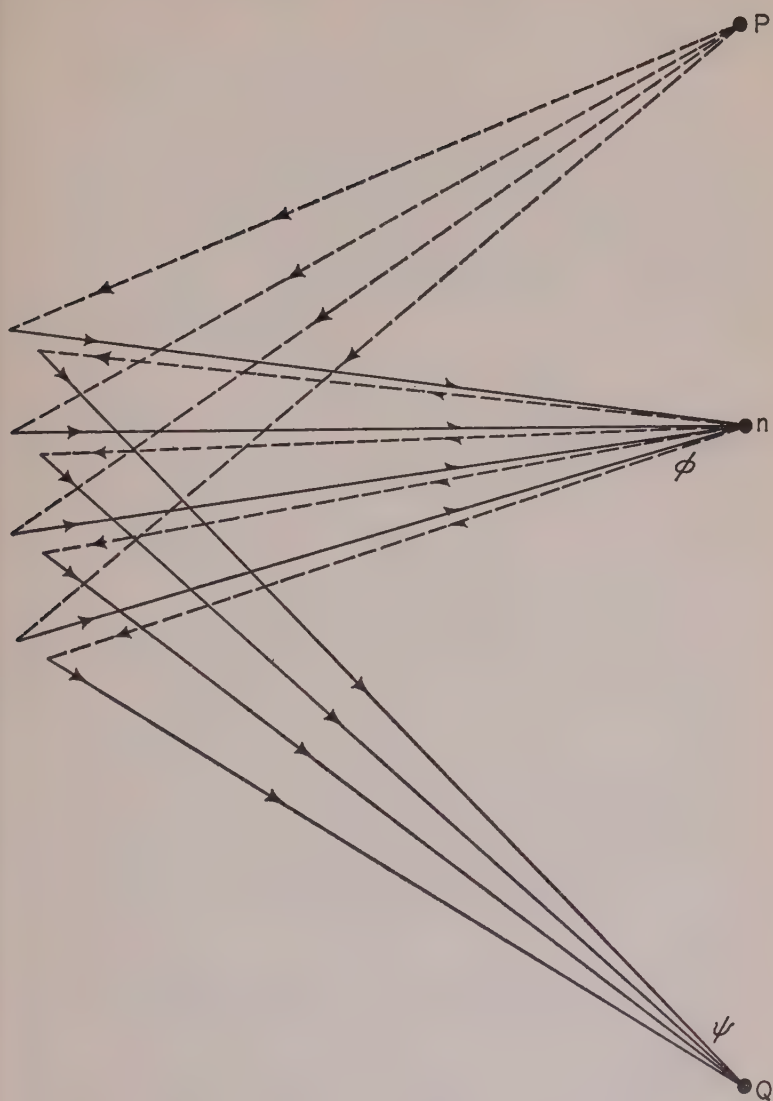


FIG. 2

previous publications. The immediate goal is not to produce a useful network, but rather, to produce anything at all which is simultaneously understandable in detail and describable in some of the language of perception.

## REFERENCES

1. P. H. GREENE, An approach to computers that perceive, learn, and reason, *Proc. Western Joint Computer Conference* pp. 181-6 (1959).
2. P. H. GREENE, A suggested model for information representation in a computer that perceives, learns, and reasons, *Ibid.* pp. 151-64 (1960).
3. S. FREUD, *The Interpretation of Dreams* Basic Books Publishing Co., New York, trans. by J. STRACHEY (1954).
4. E. MAYR, The emergence of evolutionary novelties, *Evolution After Darwin Vol. I. The Evolution of Life*, SOL TAX, ed., Univ. of Chicago Press, Chicago (1960).
5. C. F. A. PANTIN, The elementary nervous system, *Proc. Roy. Soc. Lond. (B)* **140**, 147-68 (1952).
6. W. H. THORPE, *Learning and Instinct in Animals*, Methuen & Co. Ltd., London, Eng. (1956).
7. J. PIAGET and B. INHELDER, *The Child's Conception of Space*, trans. by F. J. LANGDON and J. L. LUNZER, The Humanities Press, New York (1956).
8. H. WERNER, *Comparative Psychology of Mental Development*, rev. ed., Follett Publishing Co., Chicago (1948).
9. L. BENDER, A visual motor Gestalt test and its clinical use, *Res. Monogr. No. 3, Amer. Orthopsychiatric Assn.* (1938).
10. K. LORENZ, Comparative study of behavior (1939), in *Instinctive Behavior: The Development of a Modern Concept*, trans. and ed. by C. H. SCHILLER, International Universities Press Inc., N.Y. (1957).
11. J. Y. LETTVIN, H. R. MATURANA, W. S. McCULLOCH and W. H. PITTS, What the frog's eye tells the frog's brain, *Proc. I.R.E.* pp. 1940-51 (Nov., 1959).
12. R. A. HINDE, The conflict between drives in the courtship and copulation of the chaffinch, *Behaviour* **5**, 1-31 (1953).
13. R. A. HINDE, Appetitive behaviour, consummatory act, and the hierarchical organisation of behavior—with special reference to the Great Tit (*Parus Major*), *Behaviour* **5**, 188-224 (1953).
14. R. A. HINDE, Some recent trends in ethology, *Psychology: A Study of a Science*, ed. S. KOCH, Study I, Vol. **2**, McGraw-Hill Book Comp., N.Y. (1959).
15. K. LORENZ, The past twelve years in the comparative study of behavior (1952), in *Instinctive Behavior: The Development of a Modern Concept*, trans. and ed. by C. H. SCHILLER, International Universities Press Inc., N.Y. (1957).
16. W. SCHENKEL, Ausdrucks-Studien an Wölfen, *Behaviour* **1**, 81-130, 173-95 (1947).
17. T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc. New York (1958).

18. H. P. KRAMER and M. V. MATHEWS, A linear coding for transmitting a set of correlated signals, *I.R.E. Trans. on Info. Theory* IT-2, pp. 41-46 (1956). Reprinted as Bell Tel. Syst. Tech. Publ. Monogr. 2757.
19. M. C. GOODALL, Performance of a stochastic net, *Nature* **185**, 557-8 (1960).
20. M. C. GOODALL, Analytic net model of cognition, *Fourth Annual Meeting Biophysical Society* (February 24-26, 1960).
21. M. C. GOODALL, A basic model of cognition, *Unpublished report, Cornell Computing Center, Cornell Univ., Ithica, N.Y.* (1960).
22. B. KNIGHT, Some informal remarks on analogical organization schemes, *Tech. Memo. No. 9 (Project PARA)* Cornell Aeronautical Laboratories, Inc., Buffalo 21, N.Y. (1959).
23. P. M. MILNER, The cell assembly: Mark II, *Psychol. Rev.* **64**, 242-52 (1957).
24. P. R. HALMOS, *Finite Dimensional Vector Spaces*, *Ann. of Math. Studies No. 7*, Princeton Univ. Press, Princeton (1948).
25. P. A. M. DIRAC, *The Principles of Quantum Mechanics*, 3rd ed., Clarendon Press, Oxford, Eng. (1947).





**JOHN R. TOOLEY**

*Texas Instruments, Inc.*

## THRESHOLDING AND MICROMINIATURIZATION WITH SEMICONDUCTORS

Ultimately, in any discussion of self-organizing systems, the question of hardware must be discussed if synthesis is to be attempted. At the risk of being premature, I would like to describe for you two recent advances in semiconductor device technology. The first is an electronic device called the Esaki or tunnel diode. The second is a technique for fabricating electronic networks called the Solid Circuit.\*

In June of 1958 a letter appeared in the *Physical Review*<sup>(1)</sup> from Leo Esaki wherein he described an anomalous  $p-n$  junction diode characteristic. He observed a region of voltage and current where the current through the diode decreased as the voltage across it increased; it was a region of *negative* resistance. Figure 1 shows a typical  $V-I$  characteristic.

Negative resistance devices are not new to the electronics field, but have been proposed, discussed and used for many years. However in the past the negative resistance device has had associated with it serious disadvantages with respect to power, size, speed and fabrication. In the case of the tunnel diode this is not the case, as will become evident in the remainder of what I have to say.

For the physics of the tunnel diode and a theory of the physical origin of the negative resistance let me refer you to Esaki's letter or any of the more recent statements in the literature.<sup>(2,3)</sup> Suffice it to say that when a  $P-N$  junction diode is formed from very heavily doped semiconductor material (e.g. GaAs) the energy

---

\* Reg. trade mark.

band structure becomes degenerate. The result is a situation where classically a conduction electron would be unable to traverse the energy barrier at the junction but quantum mechanically can penetrate ("tunnel" through it) with a sufficiently high probability such that these electrons can make a significant contribution to the current flow through the diode.

The tunneling probability initially increases with increasing forward voltage but then decreases. Further increases in forward

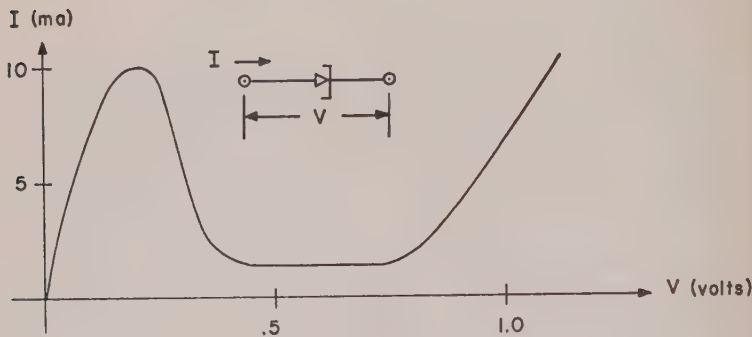


FIG. 1

voltage cause normal conduction processes to take over and tunneling processes to become negligible.

It is of interest as a logical device because of its very small size (active volume  $\leq 10^{-6}$  cu. in.), low power, and high speed. A maximum power dissipation of less than 0.1 mW can be achieved quite easily. The device operates in the kMc range with switching times on the order of  $10^{-10}$  sec. These figures correspond to the best results achieved to date but do not appear to represent ultimate limits.

Let us consider the simple series circuit shown in Fig. 2a, of a voltage source, load resistor and tunnel diode. We can portray the operation of the circuit graphically by drawing a load line corresponding to the equation:

$$V_D = E - IR$$

on the  $V$ - $I$  characteristic of the tunnel diode as shown in 2b.

The two points  $(V_1, I_1)$  and  $(V_2, I_2)$  correspond to the two stable simultaneous solutions of the tunnel diode characters and load equations. (A third point  $(V_3, I_3)$  can be shown to be unstable under small perturbations in  $V$  or  $I$ .) If  $E$  is allowed to vary, the result for increasing  $E$  is a shift of the load line parallel and upwards to itself. One notes that for  $E > E'$  where  $E' \simeq V_p + I_p R$ , there no longer are two stable solutions. In fact, if our initial operating point had been one where  $V_D < V_P$  and therefore  $I < I_p$  then as  $E$  increased to  $E' + \Delta E$ , we would have observed

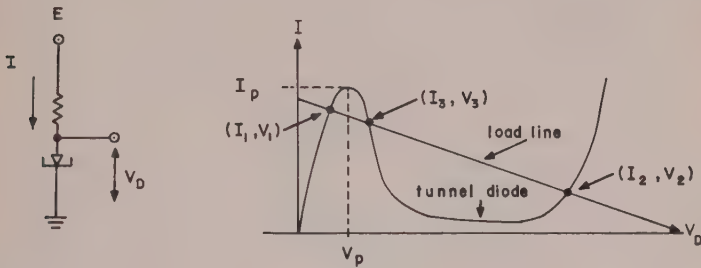


FIG. 2

a rapid switching of the diode voltage and current to the new values corresponding to the single high voltage solution. One thus sees that the tunnel diode is a device which provides us with current threshold,  $I_p$ , which if exceeded results in a discontinuous change in the voltage across it, from a low to a high value.

The application of the tunnel diode as a threshold logic generator is obvious. If each of several inputs supply some increment of current,  $\Delta i_j$ , all of which are summed and passed through a tunnel diode the voltage across the diode will switch only when

$$\sum_j \Delta i_j > I_p.$$

Figs. 3b and 3c show two possible gate circuits. The series diode in the inputs are decoupling diodes which serve to prevent a change of the inputs voltage due to a change in the output voltage. In case the tunnel diode is GaAs the decoupling diode may be a Ge switching diode.

Since my goal here is not the design of tunnel diode circuitry but rather a qualitative sketch of their utility in binary logic, I hasten to point out the oversimplification made in Fig. 3. The constraining inequalities assume the tunnel diode is always operating at one of the two points,  $(I_0, V_0)$  or  $(I_1, V_1)$ . Clearly this is not true. Consider the operating point after one of three inputs has

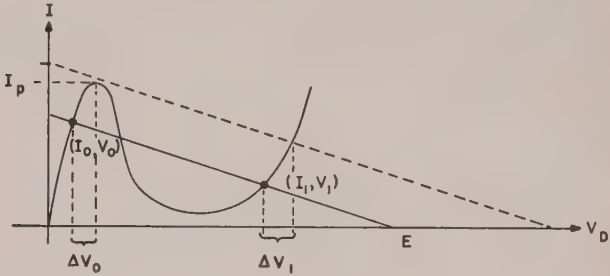


Figure 3(a)

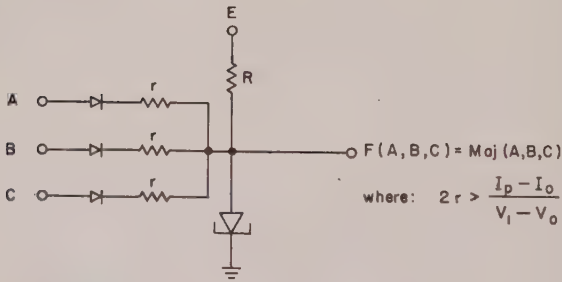


Figure 3(b)

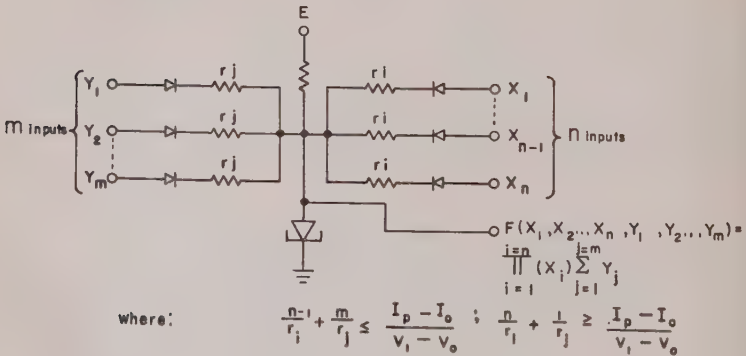


Figure 3(c)

FIG. 3



$V_1$  applied to it; the operating point is now  $(I_0 + \Delta I, V_0 + \Delta V)$  where  $|\Delta V| < |\Delta V_0|$  and  $|\Delta I| < |I_p - I_0|$ . Similarly,  $(I_1, V_1)$  is a function of the inputs. However, practical design techniques can be worked out which take these variations into account, as second-order effects when  $\Delta V_0$  and  $\Delta V_1$  are small compared to  $|V_1 - V_0|$ .

A more serious obstacle in the way of realization of large nets of tunnel diode threshold logic is essentially that of the practical uncertainty always present in the threshold of a tunnel diode gate. To see that all the static circuit parameter variations manifest themselves as threshold variations we note that the threshold integer,  $\theta$ , of a tunnel diode gate may be written approximately as:

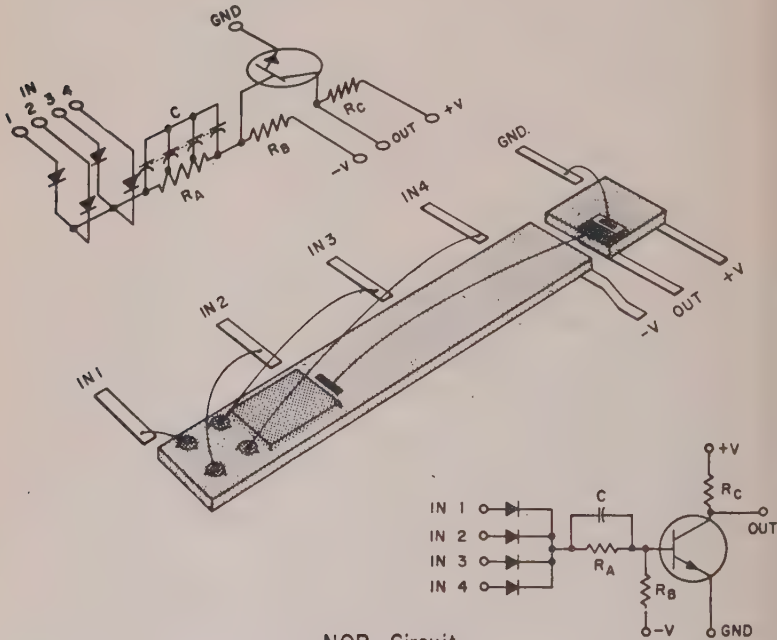
$$\theta \approx \frac{I_p - I_0}{(\overline{\Delta i})}$$

where  $I_p$  and  $I_0$  are defined as before and  $(\overline{\Delta i})$  is the mean current increment supplied by each input. In the case of unequal input weights  $(\overline{\Delta i})$  must be taken as the greatest common divisor of the different input current increments. Thus threshold variations,  $\delta\theta$ , will arise whenever  $I_p$ ,  $I_0$  or  $\overline{\Delta i}$  vary. Uncertainties in these three parameters will always exist in large networks because of manufacturing tolerances on the resistance values and diode characteristics. Even if these uncertainties are removed by careful selection of components there will be uncertainty in  $I_0$  because of power supply noise and changes in  $I_p$  and  $\Delta i$  through thermal drift. One might conclude that logical designs dependent on a fixed threshold are doomed to failure. However, a standard circuit design technique known as "worst-case" design can eliminate the problem for relatively small variations in the parameters by restraining these parameters' variations to such ranges as to cause variations in  $\theta$  to be less than  $\pm \frac{1}{2}$ . Perhaps the functional redundancy being worked on by W. S. McCulloch and M. Blum will eliminate the problem for the cases of larger variations in  $\theta$ .

The second topic I would like to discuss with you is the Solid Circuit semiconductor network.<sup>(4)</sup> I think you will find this of interest because it represents an electronic circuit fabrication technique permitting a reduction ratio of 100 : 1 in size and weight over conventional circuits fabricated from semiconductor

components, as well as a potential improvement in reliability. Since the systems we have been considering here may well require billions of elements in their electronic realization both of these factors will be of crucial importance.

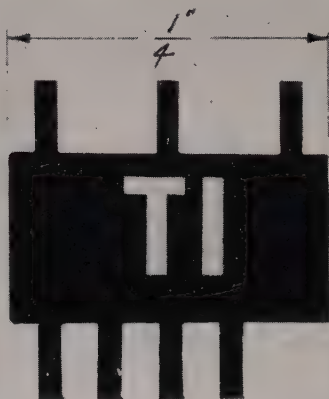
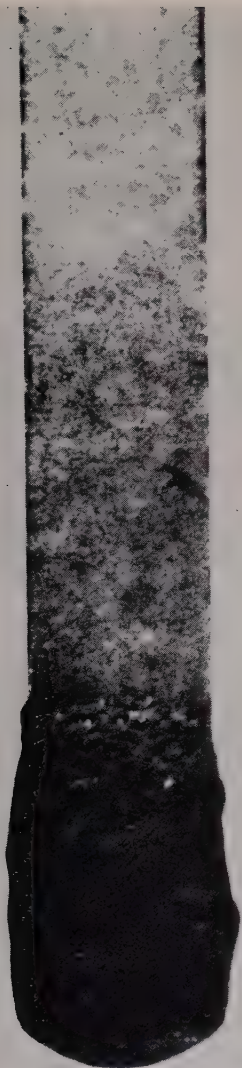
A Solid Circuit semiconductor network is a complete electronic circuit fabricated within a semiconductor material. By selection



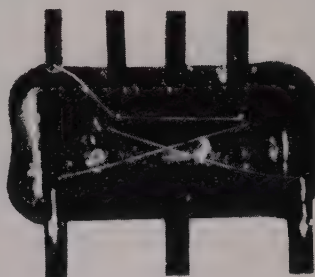
NOR Circuit  
SOLID CIRCUIT semiconductor network

FIG. 4

and shaping of conductance paths on and through the semiconductor material, it is possible to obtain such functions as switching, counting, oscillation and amplification. Networks formed in this manner are truly integrated; one network element cannot always be distinguished from another. Nevertheless, the properties of these networks are considered in terms of conventional elements such as resistors, capacitors, transistors and diodes in order to effect an orderly understandable design approach. Conventional



HERMETICALLY PACKAGED UNIT

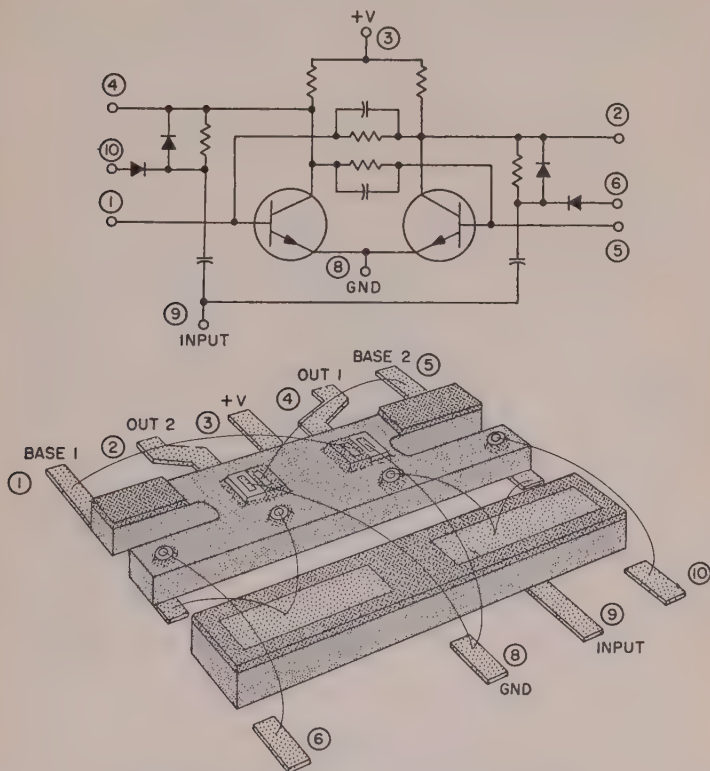


SIZE COMPARED TO PAPER MATCH

FIG. 6



circuit analysis can be used as the point of departure in device design. The resultant semiconductor network will be electrically equivalent to circuits designed with conventional components.



Layout Of Bistable Multivibrator (Type 502)  
SOLID CIRCUIT semiconductor network

FIG. 5

Figures 4 and 5 show two examples of semiconductor networks, a NOR circuit and a multivibrator, respectively.

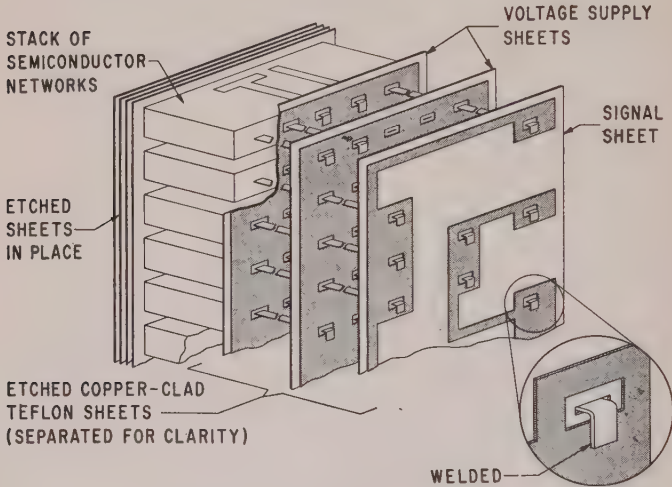
The size and weight reduction offered by semiconductor networks is unequalled by any other microelectronic product. Figure 6 shows size comparisons of a multivibrator network and a paper match.

In interconnecting devices of this type, a considerable amount



of space is required for wiring and connectors. Interconnection schemes which lay the units out flat permit in-circuit testing and replacement but waste a considerable amount of space. Placing units in a stack requires considerably less space but is somewhat more difficult to assemble and service. A method of stacking these very small packages which has been used in the past is shown in

## PACKAGE INTERCONNECTION



**SOLID CIRCUIT** semiconductor networks

FIG. 7

Fig. 7. Here the packages are stacked, but because of the very close spacing between the leads, it is not possible to provide all of the wiring in a single plane. Thin sheets of Teflon clad with copper are used to provide multiple planes for wiring. One sheet may be used for each supply voltage. Each sheet is pierced with holes which provide electrical and mechanical clearance for the leads. Any lead may pass straight through a hole and be insulated from the sheet or may be bent over and connected to it. A similar sheet or sheets may be used with a wiring pattern etched on it to provide the signal paths.

The fabrication of Solid Circuit semiconductor networks has been closely related and allied to the diffusion processes and techniques that have produced reliable diffused transistors and diodes since 1957. The diffusion process lends itself to high volume production where extremely close dimensional tolerances are held. Photomasks with tolerances less than 0.0001 in. are used to define the areas which are to be diffused or shaped. In essence, these photomasks are the only variable tooling required. A semiconductor network manufacturing line can produce a variety of circuits by merely altering the diffusion and photomasks. Also many semiconductor wafers are simultaneously diffused, yielding devices which are essentially images of one another.

In addition, semiconductor networks offer the potential of improved reliability because of the following factors. (1) Only high-purity single crystal material is used. (2) By using a single material, up to 80 per cent of connections required for a conventional circuit are eliminated. (3) The small number of process steps (approximately 15-20) required, compared to the process steps (approximately 200) needed for an equivalent conventional circuit, permit stringent process controls to be economically applied. (4) The entire circuit is hermetically sealed, and both active and passive elements are protected against environmental conditions. (5) The very small mass makes a semiconductor network difficult to damage by shock and vibration.

#### REFERENCES

1. L. ESAKI, *Phys. Rev.* **109**, p. 603 (June 15, 1958).
2. W. W. GARTNER, Esaki or Tunnel Diodes, Part I, *Semiconductor Products*, p. 31 (May, 1960).
3. W. W. GARTNER, Esaki or Tunnel Diodes, Part II, *Semiconductor Products*, p. 36 (June, 1960).
4. J. W. LATHROP, R. E. LEE and C. H. PHIPPS, Semiconductor networks for microelectronics, *Electronics*, p. 69 (May 13, 1960).

#### DISCUSSION

ZOPF: I have something that is not a question, but rather a comment I think appropriate. Two years ago I proposed forming a company called Magnafake, to manufacture gigantic paper clips and enormous matches, so that we would no longer need microminiaturization.

ASHBY: Can I ask: are the connections to the diode very critical in the way they have to be placed? To put it crudely, suppose you took a bucket of them. Would proper connections be excessively rare in such a mixture?

TOOLEY: I do not understand.

ASHBY: Suppose you took a bucketful of these diodes. Do the contacts and the way they come together, are the connections so critical, that the great bulk of contacts would be merely useless electrically?

TOOLEY: Are you thinking of the tunnel diode?

ASHBY: Yes.

TOOLEY: Yes, I think that would be the case because you need to supply series resistance with these. Possibly if you connected some resistors on these, you would have something that would be useful, but I do not think that you would have anything useful if you did just this.

ASHBY: I am thinking that in forming a random network, the point would be whether you would just finish up with nothing at all, where no current could get through anywhere.

TOOLEY: I think if you first specified that you had all of these connected to a power supply with a series resistance, and then made random interconnections from that point, since you would still have only one node, that then indeed you would have something interesting. As you vary the bias on these, you would determine what the function is it is generating, whether it would be "and" or "or" or some combination of these, or a majority type organ; you would obviously have random connections to random functions. Possibly it would be interesting and useful.

PLATT: If you use a principle which would tend to line them up with a certain polarity, the number that would be useful would be very much larger.

TOOLEY: Yes.

VON FOERSTER: I think the useful unit is a diode in every respect, is it not?

TOOLEY: Correct.

VON FOERSTER: I mean, you have to form a unit a little bit larger than the diode. Again you are developing Solid Circuits. This is only half an effect of the tunnel diode. May I ask another question? I have asked so many dollar questions already. How much would a cubic foot of those things cost, would you say?

TOOLEY: A cubic foot of tunnel diodes?

VON FOERSTER: No, no. I mean a cubic foot of solid circuits.

TOOLEY: Let me tell you what the cost of one of them is right now.

VON FOERSTER: No, I asked about a million of them. One, I know, is very expensive. A million is about the same?

TOOLEY: I would not know how to answer that, Dr. von Foerster.

VON FOERSTER: What would a single one cost?

TOOLEY: A single one would cost you about \$200-400.

VON FOERSTER: Oh, that is peanuts.

TOOLEY: A million at that rate would be several million.

VON FOERSTER: But why do you ask so much? The circuit is so little.

TOOLEY: This is because we have made only a few.

**ALFONSO SHIMBEL***Illinois State Psychiatric Institute*

# A LOGICAL PROGRAM FOR THE SIMULATION OF VISUAL PATTERN RECOGNITION

## INTRODUCTION

In what follows there will be presented a set of operations that, with varying degrees of success, yield certain properties of two-dimensional patterns that are invariant with respect to any combination of linear transformations of those patterns. A visual pattern is herein described as any bounded function,  $I(x, y)$  defined for some finite region in the first quadrant of a cartesian coordinate system. No other restrictions are placed on  $I(x, y)$  and it may be thought of as a distribution of intensities of light on a black background. Obviously, such patterns may consist of lines, dots, continuous cloud-like masses, or any combination of these.

The program may be thought of as if being performed with a piece of paper, pencil, protractor and similar devices, but the reader will readily infer the applicability of scanning and computational methods that are so familiar to the simulation engineer.

*The Program*

1. Assume the pattern to be located somewhere in the region  $0 \leq x \leq X$ ,  $0 \leq y \leq Y$ , and compute a series of expressions

$$\int_0^x I(x, y_i) y_i dx \quad (i = 1, 2, 3, \dots, K)$$

---

\* *Ed. note.*—Although the following paper was not presented to this symposium, several participants called our attention to its pertinence.

where the numbers  $y_i$  denote the levels of a sequence of straight lines like those that form the raster of a television screen. Compute the sum of those integrals.

2. Compute

$$\sum_{i=1}^K \int_0^X I(x, y_i) dx.$$

3. Using the results of steps 1 and 2 compute  $\bar{y}$ , the  $y$ -coordinate of the centroid of the pattern. Strictly speaking, of course, only an approximation to  $\bar{y}$  will be thus obtained, an approximation whose accuracy increases with increasing  $K$ .

4. By analogous means find  $\bar{x}$ .

Steps 1 through 4 serve the purpose of finding a reference point in the pattern. Points other than the centroid might in future analyses prove more useful. However, in this presentation, the primary reference point will be the centroid.

5. Map the pattern into a polar coordinate system with its centroid at the pole.

6. Let  $\mathcal{I}(\rho, \theta)$  be the polar transformation of  $I(x, y)$  and compute a series of expressions

$$\int_0^R I(\rho, \theta_j) \rho d\rho \quad (j = 1, 2, 3, \dots, N)$$

for which  $R$  is the radius of a circle with center at the pole large enough to contain the entire pattern. The angles  $\theta_j$  are to be chosen so as to conform to the orientation of a series of equally spaced radial lines like the spokes of a wheel.

7. Compute a series of expressions

$$\int_0^R \mathcal{I}(\rho, \theta_j) d\rho.$$

8. Using the results of steps 6 and 7 compute the set of numbers  $\{\bar{\rho}_j\}$  that represent the radial components of the centroids of the pattern when viewed along each of the radial lines.

9. Compute the mean of  $\{\bar{\rho}_j\}$  (call it  $\rho^*$ ) and using  $\rho^*$  as a unit of distance normalize the set  $\{\bar{\rho}_j\}$ . Label the normalized set  $\{\bar{\rho}'_j\}$ .



10. Plot the numbers  $\{\rho_j^*\}$  (using the appropriate angles  $\theta_j$ ) in a new polar coordinate system and connect them by means of a smooth curve.

The closed curve thus obtained can be thought of as a normalized panoramic display of the first moment of the light distribution in the pattern as seen from its centroid. By using  $\rho^*$  (of step 10) to normalize the size of the original pattern it then becomes possible, by completely analogous steps, to compute the second, the third, and as many higher moment abstraction contours as might be desired.

Now let us consider the class of all patterns that could be adequately displayed on a television screen having a radial (spoke-like) raster with, let us say,  $M$  discrete all-or-none photosensitive elements along each radial line. By "adequately displayed" one should infer that rotation of the raster will not produce changes in the picture.

It is mathematically obvious but important to note that because of the discreteness of the picture all of the information contained in any display on such a raster will also be contained in the first  $M$  abstraction contours of that display. In other words (except for rotation, which will be dealt with later), the *class* of all linear transformations on a pattern,  $I(x, y)$ , maps uniquely into the first  $M$  abstraction contours of that pattern and vice versa. Note, also, that inasmuch as the intensity of the pattern always enters the computations of the program in a linear fashion and all of the integrals containing the intensity are in every case eventually divided by another integral that contains that same intensity, the first  $M$  abstraction contours are also invariant with respect to any uniform change in the overall intensity of the pattern.

The question of whether two different sets of abstraction contours were derived from patterns that differ only by some combination of linear transformations can now be answered. This is done by the pair-wise superimposition of the two sets of abstraction contours and the measurement of the total area enclosed by the superimposed pairs minus the total area common to them. Let us call this number  $\Delta A$  and continue to measure it as all of the contours of one set are simultaneously turned through  $360^\circ$  with respect to those of the other set. The minimum of  $\Delta A$  obtained by this process is a measure of the "similarity" of the two sets and,

therefore, is also a measure of the similarity of the original patterns. Note that the superimposition of the corresponding pairs of contours must be done so as to make the poles of their respective coordinate systems coincide. Also, the angular orientation of the abstraction contours with respect to one another, within any given set, must be maintained. It must be maintained, in fact, throughout the process that measures  $\Delta A$ .

#### IMPLEMENTATION OF THE PROGRAM

Although the program, as so far outlined, does indeed give a means for the recognition of invariance under linear transformation (including intensity) there are certain difficulties associated with its implementation and use. One of these is the fact that for any adequate raster,  $M$  would need to be relatively large and, therefore, a complete set of abstraction contours would require a great deal of computation and storage space. The case, however, is not as bad as it may seem at first glance. In fact, most "pictures" of our practical world are highly redundant in the information theoretical sense. Therefore, the computation of a complete set of abstraction contours would also, in most practical cases, be redundant.

For example, consider the class of all simply closed curves such that every point on the curve can be reached from the centroid by a straight line that does not touch any other part of the curve. *All of the information contained in such a curve is also contained in its first moment abstraction contour!* The remaining  $M-1$  abstraction contours are completely superfluous.

For any relatively simple practical application of the program, for example the recognition of a naval vessel by its silhouette, it would seem highly improbable, speaking intuitively, that even several hundred different shapes would ever require more than the first moment abstraction contour for reliable performance. For the recognition of a small set of symbols like our printed alphabet this is almost certainly true.

#### GENERAL CONSIDERATIONS

It must be remembered that invariance under linear transformation is not equivalent to the recognition of pattern "similarity" as performed by humans and other mammals. For example, the

patterns in Fig. 1 are all readily recognizable as variations on the twentieth letter of the English alphabet. They are *not* linear transformations of each other. Nevertheless, we still regard them as somehow similar.

It is interesting to note in this connection that the first moment abstraction contours of all these patterns would be remarkably similar. This is because the linear character of the first moment combined with the normalizing process described in step 9 makes this contour highly insensitive to slight changes in the general aspect of the pattern. Minor perturbations, so to speak, are



FIG. 1

ignored. This property of the first moment abstraction contour could undoubtedly be enhanced by averaging adjacent  $\bar{\rho}_j$  before making the final plot. The higher moment abstraction contours are increasingly sensitive to details in the pattern owing to their increasing non-linearity with respect to distance.

In the final analysis it will be possible to judge how well and how economically this program or modified segments of it can imitate "shape" perception only by trying it on a large variety of shapes or by supplying more appropriate mathematical theorems on "shape" perception. The program itself may serve as a basis for defining shape somewhat more generally than the restricted notion of invariance under linear transformation. Other definitions more closely resembling our subjective notion of shape may eventually suggest themselves.

#### PROGRAMMING AND STORAGE

In simulating the foregoing logical program it might be arranged for the machine to do its computations on the pattern and then to store the results, or to compare its results with abstraction contours already stored. Thus, programming the machine to recognize a given pattern, would consist of "showing" it an example.

It would be wasteful of time for the machine to "scrutinize" all of its storage before concluding that a newly analysed pattern has been previously seen. This can be obviated by indexing stored patterns according to some simple property of, for example, the first moment abstraction contour. The area that it encloses might be a useful index. If this proves in use to be too insensitive to shape one might try using the ratio of the perimeter to the area. Many possibilities for indexing can be easily invented and explored.

## NAME AND SUBJECT INDEX

- ABBOT, E., 237  
 Abelian group, 48  
 Ability, 307  
 Abstraction of contours, 523-4  
 Acrasin, 246-250  
 Active channels, 403ff  
 Active elements, 339, 403ff  
 Activity  
   cognitive, 472  
   electroencephalographic, 312  
   sequential, 283  
   varieties of, in Beurle model, 293-298  
 Activity function, in cybernetic factory, 68  
 Activity sets, in Beer model, 55  
 Actuators, 184  
 Acuity, vernier, 321-322  
 Adaptation, 27, 272, 275, 327, 337, 341, 394  
 ADDISON, A., 250  
 Additivity, 354  
 Addresses, 320-321, 323  
 Adjoint operators, 504  
 Afferents, of formal neurons, 93  
   interaction of, 94, 109, 122, 132-133  
 AIKEN, H., 382  
 Algedonic control, 64, 72, 76, 89  
 ALLANSON, J., 165, 178  
 All-to-all principle, 128-131, 167-170  
 Alpha system, in perceptron, 391, 393, 395  
 AMAREL, S., 131-132, 204, 227, 382, 421, 443-483  
 Ambiguity (*see* Indeterminacy)  
 Amnesia, retrograde, 306  
 Amoebae, 248-249  
 Amphecks, 114  
 Amplification, as an encoding, 162  
   of intelligence, 30  
 Analog, brain, 25ff, 291ff  
 Analog junctions, in neuristor, 415  
 Analog of fabric (*see* Fabric)  
 Analysis by synthesis, 457-458  
 Analysis of variance, 256  
 Analytic *a priori*, 34  
 Analytic models, 69  
 Analyticity, 330-333  
 ANDERSEN, E. E., 321  
 ANDERSEN, H. C., 346  
 ANDERSON, T. W., 508  
 ANDRE-BALISAUX, G., 289  
 Anisotropy (*see also* Assymetry), 294, 305  
 Annihilation, of pulses, 406  
 Anti-coalition structures, 251  
 Aperiodicity, of reinforcement, 313  
 Approbation, 56  
*A priori*, function, 386  
*A prioris*, in evolved organisms, 338  
 Aptitudes, 307  
 Arcs, uniformly curved, 318  
 Arithmetic unit, 85  
 ARORA, H. L., 289  
 Arousal mechanism, 64, 71  
 Array, of magnets, 418  
   random, 317  
   time-dependent, 319  
 Arrival function, 68  
 Artefact, brain, 25ff  
 ASHBY, W. R., 31, 55, 58, 60, 76, 80-82, 86, 89, 92, 229, 238, 255-278, 328, 330, 341, 346, 520  
 ASHENHURST, R. L., 377, 382  
 Aspirations, 307  
 Assemblies, cell, 490, 495  
 Asseverated functions, 68-69  
 Asseveration, 66-70  
 Association, strength of, 455, 466-467  
   mental, 283  
 Association area, of cortex, 398  
 Association state, 393  
 Association units, 310, 312, 388ff  
 Assymetry, detection of, 288  
 Ataxia, 26  
 ATTARDI, D., 289  
 Attenuation, 294-296, 301, 313, 403  
 Attitude, 325-346  
 Automata, 25ff, 229ff, 326, 261, 369  
   reliability of, 135ff  
 Automation, 26-27, 29  
 Autonomic response, 27  
 AVERILL, H. L., 321



- Avogadro's number, 372  
 Axonal error, 122-123  
 Axons, 40, 305
- Background, irrelevant, 297, 301  
 Baconian cipher, 344  
 Barium titanate, 420  
 BARTKY, S., 228  
 Baye's theorem, 302, 303, 310  
 BEER, S., 25-89, 231, 237, 328, 330, 336, 350, 387  
 BEES, 258-259  
 Believing, 333  
 BENDER, L., 508  
 BERENDSEN, H., 116  
 Bertrand paradox, 350, 356  
 BEURLE, R. L., 236, 248, 276, 278, 291-314, 399-400  
 Birds, 323, 489  
 BIRKHOFF, G. S., 179, 482  
 Bits, natural, 5  
 Black box models, 69, 75, 229-230, 335  
 Blind spot, 279, 319  
 Block code, 139  
 Blood-sucking reflex, 273  
 BLUM, M., 95-119, 121-123, 131-133, 145, 165, 177-178, 515  
 Bone, 339  
 Boolean logic, 81, 167, 172  
 Boolean functions, 200  
 Boolean lattice, 175  
 Boole-Schroder functions, 154-155, 161ff  
 Boundary positions, 319  
 BOURBAKI, N., 55, 80  
 Bourbaki set product, 57  
 BOWMAN, J. R., xiv-vi, 62, 80, 83, 85-86, 131, 250, 417, 424  
 Brain, 264, 279-290  
   long term behavior (*see* Memory)  
 BRAIN, A. E., 426  
 Brain models, 25ff, 291ff, 385ff  
 Brain stem, 62  
 Brain waves, 312  
 BREMER, F., 289  
 BRIHAYE, J., 289  
 BRODAL, A., 63, 80  
 Buffon needle problem, 348-351, 364, 367  
 Bundling, von Neumann, 128-131, 170, 193, 196-197, 205
- BURKS, A. W., 119  
 BURNS, B. D., 325, 346
- Calculus of propositions (*see also* Propositional calculus), 114  
 CAMERON, S., 178  
 CANNON, W. B., 263  
 Canonical form, 190  
 Capacitance, generalized, 404  
 Capacity theorem, 203  
 CARNAP, R., 482  
 Carrier wave, 417  
 CARROLL, L., 253  
 CARATHEODORY, C., 345  
 Cardiovascular system, 25  
 Cat, 87, 285, 324  
 Cataracts, 317, 323  
 Categorical associations, 37  
 Categorization, 334, 344, 390, 399, 400, 450, 456  
 CAUCHY, A. L., 263, 367  
 Causality, 186  
 Cell assemblies, 490, 495  
 Cell body, 94  
 Cell density, 40  
 Cell masses, 291ff  
 Cell properties, 291-293  
 Centroid, 523  
 Cerebral control, 283  
 Cerebral cortex (*see* Cortex)  
 Cerebellum, 27, 282, 285  
 Cerebrum, 27, 28, 64  
 Chains (math.), 176  
 Chains of neurons, 294, 304-305  
 Channel capacity, 61-62, 137-141, 203, 273  
 Channel matching, 146  
 Characteristic function form, 251  
 Characters, 348  
 Cheese game, 76  
 Chemistry, 84, 257, 422ff  
 Chemotaxis, 248  
 CHERRY, E. C., 178  
 Chess, 251-253, 333  
 Chess playing machines, 333  
 Chiastic symbols, 91-95  
 Choice, 298-299, 343  
 CHOMSKY, N., 482  
 Chords, random, 351ff  
 Chromosomes, 83, 292  
 Cipher, baconian, 344  
 C-nets, 183-184, 191-192, 196, 226

- Coalitions, 234-236, 245-246, 251-252
- Coding, (*see also* Decoding, Encoding) 135ff, 343-344, 425ff
- Coenetic variable, 266
- Cognition, 285, 315ff, 343-345, 385ff, 443ff
- Cognitive functions, 493
- Collaterals, 40, 63
- Colliculus, superior, 323
- Collision of pulses, in neuristor, 408-411
- Colloid computation, 74, 76, 86
- Collusion, 14, 17, 20-24
- Color discrimination, 282
- Combinatorial analysis, 428
- Commissures, cerebral, 285, 289
- Commissurotomy, 284
- Communication and conditionality, 257
- Community set, 99, 112-113
- Compasses, magnetic, 417-421
- Compatibility, of logics, 162
- Compensating devises, 280
- Competition, 17-20, 242, 344
- Complementarity, 288
- Complementation property, 116
- Complements, 111, 113, 115
- Complexity, 272-273, 454, 464
- Component replacement, 184, 199
- Componentry, 73-75, 403ff, 511-520
- Composition rules, 239
- Computability, 58
- Computable values, 42-43, 45
- Computation capacity, 137-141, 203
- Computation, 135, 137-139, 160, 170, 199-200, 203, 205, 213, 398
- Computers, 75, 83, 230, 326, 403
- Conditionality, 255-258, 267
- Conditional probability machine, 397-398, 400
- Conditional reflex, 291-314, 327
- Conditioning mechanism, 63-64
- Confidence numbers, 455
- Confidence in beliefs, 468
- Configurations, 34-35, 41, 53-60, 67, 69, 97, 112, 146
- Congruence, 319, 354, 366
- Conjugate, hermitian, 494, 505
- Connectivity, 37, 130, 234-235, 240, 267, 281, 284, 286, 288, 293, 297, 305, 392, 426, 495
- Conservable quantities, 232, 250-251
- Constancies, 338-389
- Constraint, 73-74, 81-83, 88, 257, 276, 310-311, 335-338, 341, 445
- Construction objective, 343
- Consumption, 65, 326
- Contact nets (*see also* C-nets), 191
- Contacts, relay, 379
- Content, 45
- Context, 325, 332-333, 341-344
- Continuity, 277, 394
- Contours, abstraction of, 523-524
- Contradiction, 97, 132
- Contrast effects, 323
- Control, 28-29, 62, 72, 284
- Controllable program generator, 481
- Convergence, 263, 277
- Conversation, 230
- Convex curves, 359
- Convex hulls, 361
- Convulsions, 133
- CONWAY, T. P., 77
- Cooperation, 16ff, 234, 242, 299, 424
- Cooperative processes, 250-251, 424
- Coordination, 263
- Coordinatization, 354-355
- Corporate behavior, 307
- Corpus callosum, 284-289
- Corrective devises, 280
- Correlation, 38, 257, 302, 490
- Corresponding cells, 315
- Cortex, 32, 36-40, 52, 54-56, 59, 63, 66, 279-284, 288, 292, 314, 323
- Cortical layers, 36-41, 282, 312
- Cortico-cortical fiber system, 285
- Cosmic rays, 315, 319
- Coupling, 65, 404, 407
- COWAN, J., 81, 103, 121-122, 130-132, 135-179, 228, 231, 396, 398-400, 483
- CRAGG, B. G., 421, 424
- CRANE, H. D., 403-415
- Critical values, 236
- Cross talk, 38
- CROWLEY, T., 227
- Cryotron, 191
- Crystallization, 422-424
- Crystals, single, 73
- CUNNINGHAME-GREEN, R. A., 75
- Curiosity, 264
- Curvature, 318, 360, 366
- Cybernetic, factory, 25-89

- Cybernetician, 74, 325, 343  
 Cybernetic research, 75  
 Cybernetics, 25, 28, 325, 342  
 Cyclic connectivity, 234
- Daphnia*, 76  
 Death, 77, 122, 238  
 Decision, 65, 229, 234, 238, 330, 344, 349  
 Decomposition, of functions, 376-380  
 Decoding, 139, 169-170, 203, 343  
 Defection, 12-13, 18  
 Definition, exhaustive, 341  
 Delay lines, acoustic, 425  
 Demand function, 69  
 Dendrites, 38, 40, 94, 282, 294  
 Departures function, 68  
 Depletion, 249-252  
 Discribability, 330  
 Detection, of straight lines, 315  
 Deterministic behavior, 182-185, 195, 226  
 Deterministic events, 182-183  
 Deterministic systems, 182-186, 189-190, 226, 230, 329  
 DEUTSCH, M., 17  
 Development, 56, 315, 450  
 Diagonalization, 492  
 Dichotomies of function, 55  
 Dichotomy, 55, 313, 392, 397  
 Dielectric properties, 420  
 Differential geometry, 399  
 Diffuse damage, 279  
 Diffuse light (*see* Vision)  
 Diffusion, 74, 232, 519  
 Diffusion network, 232  
 Digital access, 86-87  
 Digital computer, 185, 328  
 DI LAMPEDUSA, G., 346  
 Dilations, 355  
 Dimentional analysis, 420  
 Dinosaur, 28, 86  
 Diodes, 75, 196, 511-513  
 Dipoles, 419  
 DIRAC, P. A. M., 509  
 Dirac notation, 500-501, 504  
 Disapprobation, 56  
 Discrimination, 285, 307, 310-314, 315-323, 347, 350, 385-402, 485, 521-526  
 Disordered structure, 279  
 Displacement, 318-321  
 Distance measures, 140  
 Distributions, 70, 201  
 Disturbance, 94, 266, 294  
 DITCHBURN, R. W., 316, 321  
 Dog, 389, 489  
 Domain, 75, 183, 245, 250, 369  
 Dominance, 27, 285, 289  
 "Don't care" conditioning, 146  
 Dormancy, 259  
 Doughnut, 87  
 Drawing by children, 487  
 Dream images, 489  
 Dreaming, 333  
 Dual basis, 502  
 Duality, 111, 116  
 Duplication, right-left, 284  
 Duplication rule, 248  
 Dyad, 505  
 Dynamic and static methods, 317  
 Dynamics, 258, 259, 342  
 Dynamic system, 260, 277
- ECCLES, J. C., 38, 40, 80, 311  
 Ecology, industrial, 25  
 Ectoplasm, 261  
 Edema, post-operative, 280  
 EDEN, M., 137, 178  
 Edge effects, 323  
 Efficiency, 36  
 Efficiency index, 7  
 Eigenvalues, 493  
 Eigenvectors, 492  
 Elastoviscosity, 421  
 Elective devises, 344  
 Electronics, molecular, 73  
 Elephant, 337, 340  
 ELIAS, P., 135, 137, 139, 141, 145, 177-178, 227-228  
 Embryo, 401  
 Embryology, 56  
 Encephalogram, 70  
 Encoding, 139, 203, 443  
 Endocrine system, 25  
 Energetics, 232  
 Energy-band structure, 512  
 Energy storage devices, 404  
 Energy variable, 406  
 Entropy, 63, 327, 423  
 Environmental fields, 237, 242  
 Epidemics, 275

- Epigenetic landscape, 56, 83, 85  
 Epilepsy, 122, 284, 289  
 Equidistance, perception of, 318  
 Equifinality, 52  
 Equilibration, 341  
 Equilibrium (*see also* Adaptation),  
   84, 263, 270-272, 263, 338, 341,  
   419, 423  
 Equiprobability, 337  
 Equivalence, 72, 185, 187, 390  
 Equivalence classes, 187  
 Ergodicity, 336, 339  
 Error, 91-94, 95-119, 121-133, 135-  
   179, 181-228, 335-336  
 Error curve, cumulated, 3, 5-7  
 Error-indication functions, 212  
 Error location, 184, 212, 216  
 Error probability, 123-125, 129  
 Error sources, in abstract neurons,  
   122  
 Error types, 181ff  
 ESAKI, L., 511, 519  
 Ethical codes, 310  
 Euler circles, 91  
 Eutectic curve, 422  
 Even-weight functions, 141  
 Evolution, 27, 229-253, 265, 269, 271,  
   369, 371, 375, 387, 393  
 Evolutionary models, 229-253  
 Exchangability, 334  
 Excitation, 144, 293-303  
 Execution time, 188, 190  
 Expected gains, 17-19  
 Expenses function, 69  
 Experimental method, 333  
 Explanation, 244, 331  
 Extensional tests, 334  
 Externalization, 327  
 Extinction, 275  
 Extracellular influences, 339  
 Eye, 315-323  
 EYRING, H., 331  
  
 Fabric, 73-76, 82, 86-88, 240  
 Facial expression, 489  
 Facilitation, 46, 63, 86  
 Factory, cybernetic, 25ff  
 Fallibility (*see* Error)  
 Fan-in mechanism, 409  
 Fan-out mechanism, 407  
 F-deterministic system, 186  
  
 Feedback, 27, 66, 69, 283, 400, 495  
 Feeling, 72  
 FENDER, D. H., 321  
 Ferro-electrics, 250, 420  
 Ferromagnetics, 250  
 Fibers, nerve, 86, 286, 289  
 Fiduciary level, 142  
 Fighting behavior, 489  
 Figural analysis, 322  
 Filters, 70, 298, 327, 347  
 Finite integrals, 354  
 FIRESTONE, F., 228  
 Fish, 282, 314, 323, 489  
 Fixation, one-shot, 313  
 Fixation point, 316  
 Flick, in eye movement, 316  
 Flow, streaming, 420-421  
 Fluctuation, 70, 275  
 Flux, 282  
 Focal conditions, 89, 263-264, 266  
 Food distribution, 232, 234, 237, 240,  
   243, 250  
 Forced teaching, 426  
 Forebrain ablation, 282  
 Formation of synapses, 282  
 Fovea, 316  
 FREGE, G., 342  
 FREUD, S., 508  
 Frog, 89, 316  
 Frontal lobe, 279  
 Function of state, 327  
 Function, preservation of, 279  
 Functional completeness, of logics,  
   172  
 Functional, linear, 502  
 Functional geometry, 315-323  
 Fuse, chemical, 404  
  
 Gabor uncertainty (*see also* Un-  
   certainty), 81  
 Gabor-McKay theory (*see also* Struc-  
   tural information), 230  
 Gait, 283  
 Gallium arsenide, 512-513  
 Games, 11ff, 76, 84, 234, 251, 327, 330,  
   335  
 Gamma system, in perceptron, 395  
 GARNER, W. R., 256-257, 278  
 GARTNER, W. W., 519  
 GASTAUT, H., 63-64, 80  
 Gate elements and circuits, 132, 183-  
   184, 192-196, 194ff, 204, 226, 414

- GELERNTER, H., 482  
 Generalization, 162, 313, 391, 393  
 Generator, random-number, 75  
 Generator, threshold logic, 513  
 Genetic determination of pattern, 323  
     371, 487  
 Genetics, neo-Darwinian, 56  
 Geniculate nucleus, lateral, 323  
 Genotype and phenotype, 386  
 Geometry, 87, 315-323, 347-368, 399  
 GEORGE, F. H., 74  
 Gestalten, 33-35, 48, 56, 485-509  
 GIBSON, E. J., 323  
 Goal attainment, 474  
 GOODAL, M. C., 493-497, 509  
 GOEDEL, K., (*see also* Computability),  
     58  
 Gonads, 284  
 GOSWAMI, P., 483  
 Governments, 312  
 Graeco-Latin squares, 428  
 GRAFSTEIN, B., 290  
 Grain size, 319  
 Graph, 234, 259  
 Grau, 483  
 GRAY, J., 290  
 Greatest lower bound, 477  
 GREENE, P. H., 400, 485-509  
 Grid size, 348-349  
 Groups (math.), 48, 337, 355, 421  
 Group behavior, 1-24, 229-253  
 Grouping, 37  
 Gyri, central, 52-53  
  
 Habituation, 46, 63, 275  
 Haemorrhage, 271  
 HAGELBARGER, D., 227  
 Half-brain preparations, 284  
 HALLE, M., 457, 482  
 Hallucination, 333  
 HALMOS, P. R., 509  
 HAMMING, R., 178  
 Hardware, electronic, 339  
 HARTSHORNE, C., 119  
 HAYEK, F., 382, 383, 399  
 Heat flow, 191  
 HELMHOLTZ, L., 316  
 Hemispherectomy, 284  
 Heuristics, 329, 462, 469  
 Higher mental functions, 340  
 "Hill-climbing" devices, 81, 238  
  
 HINDE, R. A., 508  
 HIRST, H. J., 77  
 Homeo-projection, 289  
 Homeostasis, 31, 64, 71, 84, 263, 271,  
     339  
 Homeostat, 58, 71, 72, 268  
 Homing instinct, 281  
 Homogeneity, 43-44, 85, 415  
 Homomorphism, 31, 70, 82, 331  
 Homunculi, cortical, 56  
 HOPKINS, D. A., 77  
 HORRIDGE, G. A., 311  
 HOWES, S. R., 77  
 HOWLAND, B., 131  
 HUFFMAN, D., 228  
 Hull, convex, 361  
 Hypercubes, 146  
 Hypothalamus, 284  
 Hypothesis formation, 443-483  
  
 Ideal components, 193  
 Identification, 344, 390  
 Identity, 342  
 Ideographs, 91  
 Ignorance, 327-329  
 Images, visual, 316, 323  
 Implants, in cortex, 280  
 Implicant number, 213  
 Impulses, 56, 63, 94  
 Independence, 123, 132, 257, 332  
 Indeterminacy, 43, 55  
 Indeterministic, systems, 185, 189-  
     190  
 Indexing, 526  
 Indication, of error location, 218  
 Individualism, in games, 17  
 Induction, 115, 453, 456  
 Inductive logic, 456  
 Industry, as cybernetic system, 25-26  
 Inertia, 417  
 Infimum, 477  
 Information (*see* Metrical informa-  
     tion, Structural information)  
 Information measure, 135  
 Information storage (*see also* Mem-  
     ory), 8-9, 400, 425  
 INHELDER, B., 508  
 Inhibition, 93-94, 122, 133, 283, 289,  
     293-294, 299, 303  
 Input, 37, 74, 166, 183, 395, 328, 400  
 Instability, 293



- Instinct, homing, 281  
 Integral, geometry, 347-368  
 Integratable function, 355  
 Integration, 55, 263, 277, 286  
 Intelligence, 30, 76, 270-277, 327-328, 331  
 Interaction, 181, 230, 329, 338, 409  
 Interfaces, 74  
 Internal state, 261, 328, 343  
 Interneurons, 38, 40  
 Intersection measure, 393, 397  
 Intuition, 299, 307  
 Intuitive behavior, 292  
 "Intuitive inspiration", 309, 310  
 Invariance, 71, 317, 320, 352, 354, 357  
 I.Q., 328  
 Irregularities, developmental, 316  
 Isolates, 326  
 Isolation, 181-182, 272  
 Isomorphism, 31, 164, 260  
 Iterative line, 404
- JEFFREY, R. C., 261, 278  
 JENNINGS, H. S., 275, 278  
 JENNINGS' law, 275  
 JORDAN, E. C., xi  
 JOSEPH, R. D., 396  
 Judgement, 34-35, 62, 72  
 Junctions, neurister, 407-413
- KAPLANSKY, I. M., 428  
 KAUTZ, W. H., 428  
 Kidneys, 284  
 Kinetics, 423  
 KISA, S. I., 178  
 KLÜVER, H., 321  
 Knife cuts, in cortex, 280  
 KNIGHT, B., 509  
 Knobs, synaptic, 38  
 Knowledge, 331  
 KOCHEN, M., 178  
 KRAMER, H. P., 509
- Labeling, 344  
 Labour function, 69  
 LANGER, S., 330, 346  
 LANGFORD, C. E., 178  
 Language, 34-35, 45, 72, 83, 87, 253, 320, 331, 444-445, 447, 457
- LASHLEV, K. W., 284  
 LATHROP, J. W., 519  
 Lattices, 131-132, 147, 174-175, 342-343, 479  
 Learned elements, 323  
 Learning, 1-11, 63, 71, 285, 291-314, 315-323, 327-329, 385-402, 443, 486  
 Learning capacity, 392  
 Learning curves, 392  
 Learning experiments, 1-24, 398  
 Least upper bound, 478  
 Lebesgue measure, 343, 353  
 Lecithin, 75  
 LEE, R. E., 519  
 Leech, 273  
 Length measures, 364  
 Lesions, 63, 279, 284  
 Letter game, 76  
 LETTVIN, J. T., 131, 231, 316, 321, 401, 489, 508  
 LEWIS, C. I., 135, 172, 178  
 Lewis logic, 154-155, 161, 164, 172-175  
 Lie algebra, 132  
 Life, 269, 271-272  
 Lifetime, 184, 198, 225, 226  
 Line segment, oriented, 348, 364  
 Line-splitting, 169  
 Linear combination, 502  
 Linear equations, 76  
 Linear programming, 28  
 Linear transformations, 244, 523  
 Linguistic analysis, 457  
 Linguistics (*see* Language)  
 Links, associative, 466  
 Lipids, 74-75  
 Liquids, aqueous, 74  
 LISSMAN, H. W., 290  
 Literals, irredundant, 202  
 Living components, 76  
 LLOYD, D. P. C., 133  
 Localization, cortical, 56  
 Locomotor gait, 283  
 LÖFGREN, L., 149, 178, 181-228, 399  
 Logic (*see also* Boolean, Digital Inductive, Many-valued, non-Aristotelian, Non-planar, and Probabilistic logics), 292, 338  
 Logic of probable arguments, 123  
 Logic functions, 91-94, 95-119, 121-133, 135-179, 369-383, 443-483

- Logical depth, 149, 151, 170  
 Logical elements, universal, 130  
 Logical information (*see* Structural information)  
 Logical notation, McCulloch-Pierce, 91-94  
 Logical stability, 204, 248  
 Logons, 231  
 Longimeter, 367  
 Loop-matrices, 201  
 Loops, homeostatic, 283  
 LORENTE DE NO, R., 38, 80  
 LORENZ, K., 487, 508  
 LOTKA, A. J., 261, 278  
 LOWENSCHUSS, O., 178  
 LUCE, R. D., 251  
 Luce psi function, 235, 252  
 LUKASIEWICZ, J., 135, 171, 172, 178  
 Lukasiewicz logic, 172, 174  
 Lungs, 284  
  
 Machine, 69, 181, 260-262, 277, 333-336, 369  
 MACKAY, D. M., 135, 178  
 Magnets, 422  
 Magnetic cores, 426  
 Magnetic drum store, 85  
 MAGOUN, H. W., 63, 80  
 Maintenance cost, 235  
 Maintenance, dynamic, 338  
 Majority function, 202  
 Majority organ, 124, 125, 127  
 Management, 72  
 Man from Mars, 251, 252, 253  
 Manifold, 355, 357, 362, 363  
 Many-valued logic, 128, 135-179  
 Mapping, 45, 48-49, 59, 63, 76, 183, 192, 262, 343-344, 398  
 MARCH, J. G., 255, 278  
 Markov chains, 61, 81  
 MARSHALL, W. H., 319, 321  
 MARVEL, C. S., 420  
 Material, high-variety, (*see* Fabric)  
 Materiality, 260  
 Mathematics, 331  
 MATHEWS, M. V., 509  
 Mating behavior, 489  
 Matrices, 83, 86, 330, 491-498  
 MATTHEWS, B. H. C., 133  
 MATURANA, H. R., 321, 508  
 Mautner, 483  
 MAYR, E., 508  
  
 Mean free path, 198  
 Means-end readiness, 468  
 Measure, 82, 185, 352-354, 358, 362-367  
 Measure function, 185  
 Mechanism, analog of, 82  
 Mechanistic behavior, 260-262  
 MEDAWAR, P. B., 344, 346  
 Medulla, 31  
 Membranes, semi-permeable, 74  
 MENGER, K., 131  
 Memory (*see also* Storage), 85-86, 264, 276, 295, 298-299, 305, 313, 327-328, 371, 375, 386, 390, 425  
     associative, 458  
     chronological, 71  
     computer, 328, 425ff  
     distributed, 425-442  
     gestalt, 71  
     long term, 291, 296, 304-309  
     mechanisms of, 395  
     physical basis of, 276-277  
     physiological, 86  
     reactivation of, 282  
     short term, 291, 304-309  
 Memory load, 1, 2, 4, 9  
 Memory sequence, 304  
 Memory trace, 299, 301, 303  
 Mental activity, communal, 310  
 Mental functions, 333  
 Mental models, 443  
 Mesencephalon, 38  
 Message, to-whom-it-may-concern, 313  
 Metabolism, 64  
 Meta-classes, 399  
 Metalanguage, 72, 83, 253  
 Metastability, 422  
 Meteorites, 265, 279  
 Metrical information (*see also* Structural information), 48, 135, 142-143, 152, 231, 270, 330  
 Metrons, 231  
 Microlevel, redundancy, 149  
 Microminiaturization, 511  
 Micromodules, 73, 511-520  
 Microprogramming, 85, 86  
 Midbrain, 281  
 Midbrain optic lobe, 314  
 MILLER, G. A., 482  
 Mill, steel, 65  
 MILNER, P. M., 495, 509

- Miniaturization, electronic, 403, 511-520  
 Minimal complexity, of nets, 202  
 Minimum energy, 250-251  
 MINSKY, M., 93  
 Minsky-Selfridge diagram, 98  
 Mirror images, 287  
 Missile gaps, 326  
 Missile systems, 326  
 Models, 25-89, 229-253, 385-386, 443, 483-509  
 Modes, 488ff  
 Molecular electronics, 74  
 Molecular level, 82  
 Molecules, domains of, 73  
 Moment, dipole, 420  
 Momentum, conserving angular, 87  
 Money function, 69, 70  
 Monitoring, by U-machine, 56  
 Monkey, 264, 285, 323  
 Monte Carlo methods, 230, 360  
 Mont St. Michel, 343  
 Monostable circuits, 404  
 MOORE, E. F., 119, 165, 227, 228, 370, 382  
 Mosaic, retinal, 316-319, 320  
 Motions, eye, 316-317, 322  
 MOTT, 382  
 Move neighborhood, 239  
 MULLIN, A., 83, 132  
 MULLER, D., 228  
 Multilevel circuits, 379  
 Multiplier, 85  
 Multivariate analysis, 256, 492  
 Multivibrator, 517  
 Muscles, 248, 339  
 Mutation, random, 56  
 MYERS, R. E., 290  
 Mystique, 329  
 MCCARTHY, J., 228  
 MCCULLOCH, W. S., 38, 52, 63, 80, 85, 87-88, 91-94, 95, 103, 119, 121-123, 131-132, 135-137, 165, 177-178, 204, 228, 231, 249, 281, 284, 289, 316, 321, 323, 369, 381, 386, 450, 482-483, 508, 515  
 McCulloch nets, 91-179  
 MCGILL, W. J., 256-257, 278  
 MCSHANE, E. J., 178  
  
 Naming, 344  
 N-ary operation, 132  
 Natural selection, 275  
 Nature, 335, 337, 338  
 Navigation, 281, 323  
 Needle, Buffon, 348, 350, 364, 365  
 Neighborhoods, food, 232  
 Neocortex, 284  
 Nervous system, 266  
 Networks  
   analysis and synthesis of, 121  
   classification of, 190-191  
   food-distribution, 231-232  
   gate, 183  
   logically stable, 95-119, 121-133, 135-179, 202  
   neuronal, 95-179, 279-290, 291-314, 385-402, 403-415  
   randomly connected, 292, 298, 304  
   redundancy of, 124, 200-203  
   self-repairing, 183  
   Tee, 413  
   of relaxation oscillators, 499  
 Neuraxis, 64  
 Neuristor, 403-415  
 Neurodynamics, 132  
 Neuron, 91-94, 95-119, 121-133, 282, 305, 339, 386-387, 402-404  
 Nicotinic acid, 420  
 "Nits" (*see* Bits, natural)  
 Node-element, 460  
 Noise, 70, 94, 136, 165, 297-298, 490-491, 506  
 Non-Aristotelian logic, 171  
 Non-congruence, 319  
 Non-conservative system, 391-398  
 Non-linear functions, 82  
 Non-linear media, 74, 299  
 Non-metric variables, 256  
 Non-planar logic, in neuristors, 411  
 NOORDENBOS, W., 64, 80  
 "Nor" circuits, 517  
 Normal coordinates, 259  
 Notation, Dirac, 500, 501, 504  
   McCulloch-Pierce, 91-94, 448, 460  
 Novelty, of environment, 298  
 NOVIKOFF, A. B. J., 83, 132, 249, 347-368, 382, 399, 401  
 Nucleation, 422  
 Numbers, computable, 58  
   Goedel, 58, 61  
   quasi-pseudo-random, 82  
   rational, 83  
 Objectification, 327, 334

- Observer, uncertainty of, 258  
 Observer-system interaction, 181, 230, 329, 338  
 Odd-weight functions, 141  
 OLIVER, B. M., 163, 178  
 Open systems, 250  
 Operant, 86  
 Operational research, 28, 66, 69, 77  
 Operators, 86, 498, 504  
 Optic lobes, 281-282, 323  
 Optic nerves, 281  
 Optic tract, 281  
 Order (*see also* Organization), 327, 330, 342, 423  
 Order-continuous functions, 157  
 Order-preserving functions, 157  
 Organic materials, 74  
 Organization, 255-278, 342  
 Orientation, 349-350  
 Orthogonal bases, 489  
 Orthonormal functions, 490, 496  
 Oscillation, patterns of, 490  
 Osteoblast, 339  
 Osteoclast, 339  
 Overdetermination, of mathematics, 331  
 Overhead, 69  
 Overlap, of patterns, 347, 366
- Pain, 64, 76  
 PANTIN, C. F. A., 487, 508  
 Paradox, Bertrand, 350  
 Paralleling of channels (*see also* Bundling), 124  
 Paratheory, 346  
 Parity, of jots, 140  
 Part and whole, 258-260, 340  
 Part functions, 275  
 Particle physics, 447  
 Partitioning, 343  
 PASK, G., 31, 74-75, 80, 82, 95, 229-253, 266, 328-330, 387, 400  
 Passive elements, 339  
 Pathetic fallacy, 331  
 Pattern perception (*see* Recognition)  
 Payoff functions, 11, 234, 244-247, 251-252  
 Pendulum, 259, 270  
 PENFIELD, W., 56  
 Perception, 285, 319, 322-323, 342-334, 525  
 Perceptron, 321, 347-348, 365, 385-402  
 Perceptual elements, 485  
 Performance criterion, 84  
 Permissible transition, 252  
 Permutations, 111, 337  
 Personalism, 520  
 PETERSEN, W., 137, 178  
 Petrine method, 335  
 Phase-space, 35, 71, 82, 86  
 Phenotypes, 386  
 PHIPPS, C. H., 519  
 Photocells, 397  
 Physical world, 257  
 Physics, 257, 258, 332  
 PIAGET, J., 276, 450, 482, 508  
 Piano playing, 283  
 PIERCE, C. S., 114, 119  
 Pierce ampheck, 141, 154, 157  
 PIERCE, J. R., 163, 178  
 Pigeons, superstitious, 401-402  
 Pilot, automatic, 271  
 PITTS, W. H., 131, 321, 369, 381, 386-387, 508  
 Plane waves, 299  
 Plant function, 69  
 Plasmodium, 250  
 Plasticity, 302, 329  
 PLATT, J. R., 311, 315-323, 397, 399, 401-402, 520  
 Pleasure, 76  
 POINCARÉ, H., 261, 364  
 Point of inflection, 86  
 Poisson's series, 299  
 Polarization, resting, 339  
 Polyhedra, n-dimensional, 71  
 Polymer, 420-421  
 Polypecks, 114, 116, 118, 130  
 Polyps, coral, 295, 311  
 POPPER, K. R., 327, 346  
 Possibilities, 257  
 Post-development mechanisms, 316  
 POST, E. L., 135, 178  
 Post logics, 135, 164-165, 171-175, 289  
 Post-Lukasiewicz logic, 162-165, 167, 173, 175-176  
 Postulation, 444  
 Potential, resting, 339  
 Potentialities, representation of, 486  
 Power supply, 339  
 PRANGE, G., 103

- Pre-baiting, 265  
 Predicates, independence of, 337  
 Prediction, 69, 456  
 Predictive mode, 458  
 Prejudice, 325-326, 330, 343  
 Pre-programming, 336, 399  
 Pre-structuralization, 399  
 PRIBRAM, K. H., 264, 278, 482  
 Prisoner's dilemma game, 12, 14  
 Privacy, of perception, 320  
 Probabilities logic, 95, 123, 136  
 Probability  
   conditional, 46, 74, 304  
   cumulative, 124  
   of error, 95-119, 121-133  
   geometric, 348, 350  
   tunnelling, 512  
 Problem language, 76  
 Problem solving, 328  
 Procedure space, 466  
 Procedures, in theory formation, 465-466  
 Processing power, 453  
 Procrustes, 340  
 Product, Bourbakian, 61  
 Product, inner, 504  
 Product logic, 172  
 Product space, 257  
 Production systems, 65  
 Profit, 83, 84  
 Program, simulation, 394, 397  
 Programming, 336, 399, 525  
 Projection, 368, 396  
 Projective geometry, 87  
 Proof, 331, 333  
 Propagation, all-or-none, 404  
 Properties, 258, 327, 341  
 Propositional Calculus (*see also* Calculus of Propositions), 152ff, 450, 479  
 Propositional functions, 447  
 Protein, 76, 270  
 Psychology, 256, 333, 389  
 Pulse interaction, in neuristor, 409ff  
 Pyramidal cells, 38, 288  
  
 Quadrant, ventral visual, 281  
 Quadrigeminal plate, 285  
 Quantization violation, 192  
 Quantum of action, 185  
 Quicksand, 343  
  
 QUINE, W. V. O., 343, 382  
  
 Radiation, 74, 279, 315, 319  
 Railroad crossing problem, 412  
 Random networks (*see* Networks)  
 Random variables, 60, 144, 349  
 Randomness, 144, 350-353  
 Randomizing operations, 75, 117, 238  
 Range, 183  
 RAPOPORT, A., 1-24, 249, 251-252, 313  
 RASHEVSKY, N., 238  
 Rat, 265, 389  
 Rate-dependent properties, 421  
 Ratio, signal-to-noise, 319  
 READ, H., 331  
 Reading, 112  
 Realizability, 453  
 Receptors, sensory, 64, 309, 317, 389  
 Reciprocal innervation, 289  
 Recognition, 37, 62, 315-323, 342-345, 348, 367, 385-402, 444, 457-458, 521-526  
 Reciprocity, 338  
 Reciprocity of connections, 288  
 Rectifiable curve, 350, 364  
 Reducibility, 256  
 Redundancy, 9, 37, 52, 63, 81, 136, 138-139, 149, 196-197, 200, 248, 283, 493, 524  
 Reference frame, 331-332  
 Reflex, blood-sucking, 273  
   conditional, 291-314, 327  
   stinging, 273  
 Refractory period, 245, 293, 404, 411  
 Regeneration, 281-282  
 Regularity of behavior, 261  
 Regulation, 275  
 Reinforcement, 282-283, 296, 302, 313, 391, 393-395, 398, 402  
 Reinforcement control system, 386, 389, 391-392, 398  
 Relabeling, 337  
 Relational properties, 258, 341  
 Relations, 35, 276, 303, 327, 329  
 Relative position, 191  
 Relativity, 266  
 Relaxation oscillations, 499  
 Relay, 191, 213, 379  
 Releasers, 487  
 Reliability, 91-94, 121-133, 135-179, 181-228, 370



- Religion, 84-85  
 Repair  
   biological, 280  
   machine, 181-228  
 Repartitioning, 332  
 Repetition, 313  
 Replacement action, 220  
 Representation, 81, 288, 303, 310, 331  
   336, 342, 452, 485-509  
 Reproduction, 25, 30, 75  
 Requisite variety, 31, 81, 277  
 Research methodology, 444  
 Reservoir components, 218  
 Resistance, negative, 511  
 Resolution, of physiological states,  
   275  
 Resonance, 252, 419  
 Response units, perceptron, 386, 389,  
   402  
 Responses, 38, 40, 298-310, 398, 402,  
   489  
 Resting energy, 404  
 Restoring force, 417  
 Restoring organs, 142, 169  
 Restriction, 257  
 Reticular formation, 62, 64  
 Retina (*see also* Mosaic), 281-282,  
   315-316, 323, 348, 350, 353, 393,  
   521ff  
 Retinal elements, 315, 397  
 Retrograde amnesia, 306  
 Retrospection, 303  
 Reverberations, 413  
 Reverberatory chains, 304-305  
 Reward function, 56, 64, 72  
 Rhythmic activity, 38  
 RIGGS, L. A., 316, 321, 323  
 Rigor, 343  
 Rings, of neuristors, 412  
 R-machine, 63, 64, 71, 72  
 Roots, dorsal, 283  
 ROSE, A., 179  
 ROSEN, C. A., 88-89, 250-251, 382,  
   425-442  
 ROSENBLATT, F., 249, 312, 321, 385-  
   402, 425, 427, 442  
 Rotation, 323, 352-355  
 Roulette wheel, 230  
  
 SADLER, H., 77  
 Scalar combination, 502  
  
 Scale transforms, 70  
 Scanning, 37, 40, 81, 317, 319, 323,  
   471  
 SCHELLING, T. C., 13  
 Schemes, compound, 464  
 SCHENKEL, W., 508  
 SCHUTZENBERGER, 132, 483  
 Science  
   controlled *vs.* descriptive, 82  
   physical, 332  
 Scientists, 340  
 SCOTT, M. D., 77  
 Seeing, 333  
 Segment, oriented, 350  
 Selection, 274, 277, 344  
 Selection, in synopsis, 282  
 Selective information (*see* Metrica  
   information)  
 Selectivity, 298, 422, 423  
 "Self", 269  
 Self-collision, of pulse trains, 410  
 Self-congruence, 317, 318, 322  
 "Self-connecting", 267  
 Self-constructed components, 74  
 Self-designing processes, 74  
 Self-healing, 404  
 Self-organizing potential, 2, 19, 23  
 Self-repair, 181-228  
 SELFRIDGE, O., 93, 343, 346  
 Self-reproduction, 244  
 Semiconductors, 74, 75, 511-520  
 Senility, 279  
 Sensation, 66, 72  
 Sensibility, 32ff, 52  
 Sensitivity, variation of, 319  
 Sensory configurations, 33ff  
 Sensory elements, 312  
 Sensory-sensory interaction, 286  
 Separability (*see also* Independence),  
   83, 256, 330, 332, 340  
 Sets, 32, 55, 60-62, 71, 81, 83, 99, 112-  
   113, 331, 333, 490  
 Sequence of causes, 186  
 Sequences, 283, 309  
 Sequential circuits, 183  
 Shannon capacity theorem, 203  
 SHANNON, C. E., 119, 163, 165, 178,  
   228, 278, 330, 370, 376-377, 379,  
   382  
 Shannon-Moore nets, 103  
 Shape detection, 323  
 Shear, 420

- Sheffer functions, 97, 114, 116, 141–142, 153–154, 157, 161–164
- SHERWOOD, S., 84, 87, 250, 289, 342, 400
- SHIMBEL, A., 313, 521
- SHOLL, D. A., 53, 80, 293, 311
- Side groups, chemical, 421
- Signalling activity, 232, 234
- Signal processing, 312
- Similarity, 311, 343, 392, 394–395, 523–524
- SIMON, H. A., 255, 278
- Simultaneity, 302
- SKINNER, B. F., 401, 402
- Skinner procedure, 399
- Sleep rhythm, 63
- Slime mold, 246, 248, 250, 265, 326
- Sludges, 326, 330, 401
- Social amoebae, 246
- Social inertia, 252
- Sociological situation, 252
- Sociology, 236
- Solid Circuit, 511–515
- SOMMERHOFF, G., 89, 263, 264, 266, 278, 343, 346
- Source matching, 146
- Specialization, 28, 46, 273
- Speculation, 303
- SPERRY, R. W., 279–290, 313, 323
- Spinal chord, cat, 133
- Spinal dog, 26, 28
- Spores, 248, 250
- Springs, 259, 417
- Stability, 187, 293, 299, 303
- Stabilized retinal images, 316, 323
- Stabilizing mechanisms, 87, 291, 299, 301
- Stationary errors (*see* Error types)
- Statistical analysis, 253, 256, 492
- Statistical homogeneity, 70
- Statistical variables, 492
- Statistician, 253
- Steel mill, 89
- Steiner's triples, 428
- STEINHAUS, H., 367
- Step functions, 344
- Steroids, 248
- STEVENS, K. N., 482
- Stimulation, physiological, 312
- Stimuli, 64, 135, 297–299, 301–302, 487
- Stinging reflex, 273
- Stochastic processes, 66, 275
- Stock utility functions, 68
- Storage of information (*see also* Memory), 85, 306, 328, 425–426, 432, 525
- Straightness, 320
- Strategies, 1–24, 233–234, 462
- Strikes, 83
- Structural information, 48, 231 (*see also* Metrical information)
- Structure d'ensemble, 480
- Strychnine, 94, 133
- Subcellular mechanisms, 339
- Submanifold, 357
- Subset, constrained, 257
- Substitutions, extensional, 342
- Supercooling, 422
- Super iteration, 226
- Superposition, 391, 406, 489
- Superstition, in pigeons, 401–402
- Supplemental complementarity, 287–288
- Supply function, 68
- Suppression, 297, 302
- Supremum, 478
- Surround, inhibitory, 283
- Survival, 28, 56, 67, 84–85, 271, 292, 337, 340
- Svoboda, 401
- Swarming, 259
- Switching diodes, 513
- Switching nets, characteristic behavior, 190
- Switching function, 376
- Switching theory, 369, 383
- Symbol, 87
- Symbolism, 332
- Symmetric function, 379
- Symmetry, 172, 176, 286, 423
- Synapse, 38, 94, 266, 282, 495
- Synaptic selection, 282
- Syncategorematic, 342
- Synchronization interval, 187
- Synchrony, of states, 122
- Synthetic *a priori*, 35, 243
- Synthetic *a posteriori*, 34–35
- Systems theory, general, 260
- Szent-Gyorgi, 340
- Tachistoscopic vision, 322, 344
- TALBOT, S. A., 319, 321
- Taurus, 87

- Tautology, 97, 132  
 Taxis, 76  
 TAYLOR, W. K., 425, 442  
 Teaching, iterative, 431  
 Techniques, multiple multiplexing, 66  
 Tectum, 202, 282, 314  
 Teleology, 84, 341  
 Telephone switching systems, 375  
 TEMPERLEY, H. N. V., 421, 424  
 Temporal lobe, 279  
 Temporal structure (of computable values), 47  
 Temporary errors (*see also* Error types), 121-133  
 Temptation to defect, 12, 23  
 Ternary operations, 132  
 Thalamic integration, 55  
 Thalamus, 55, 63  
 Thermionic diodes, 75  
 Thermister, 404, 406  
 Thermodynamics, 87, 250, 424  
 Thinking, 303-7, 332, 333  
 Thinking machine, 333  
 Thinking, sequential, 310  
 THORPE, W. H., 487, 508  
 Thought processes, contemplative, 292  
 Thread structures, 31, 75, 248, 266  
 Threshold, 94, 96, 121, 123, 129-30, 276, 293, 296, 301, 313, 396, 404  
 Threshold functions, 383  
 Threshold logic generator, 513  
 Threshold switching cell, 458  
 Time, 372, 375  
 Time delay, 96, 122  
 Time-ordering, 183  
 Time series, 44  
 T-machine, 51-72  
 TOCHER, K. D., 75  
 TOLMAN, E. C., 468, 482  
 Tonic interplay, 339  
 TOOLEY, J. R., 511, 520  
 Topology, 74, 241, 243, 392  
 Torus, 87, 231, 242  
 Trajectories, 51-52, 259  
 Transducers, 312, 386  
 Transfer components, 226  
 Transfer errors, 184, 226  
 Transfer functions, 83, 121-2, 491, 501  
 Transformations, 40, 44, 47, 49, 60, 70, 110, 259, 394, 492-3  
 Transformation group, 355  
 Transition interval, 182, 185, 187  
 Translating machines, 333  
 Translation, of patterns, 87, 349, 352-5, 447  
 Transmission, signal, 403  
 Transmission line, 417, 421  
 Transmitter, radio, 267  
 Tree, 469  
 Tree-growth, process, 471  
 Tremor, 322, 316  
 TRENT, H., 228  
 Triangle, equilateral, 351  
 Trigger variable, 404-6  
 Tropism, 76  
 Truth, 182, 332  
 Truth tables, 91  
 Tunnel diode, 511-20  
 TURING, A. M., 86, 228, 369, 381, 386  
 T-U-V system, 63  
 Twins, 315  
 Two-valued calculi, 152  
  
 Ultrastability, 31, 55  
 U-machine, 56, 76  
 Unanalyzed states, 258, 259  
 Uncertainty, 4-7, 82, 256-7  
 Undecidable situations, 238  
 Uniform switching nets, 183  
 Unity, functional, 344  
 Universal elements, 141  
 Universe of functions, 81  
 Unscrambling, 493  
 Utility, 341  
 UTTLEY, A. M., 397, 399, 425, 442  
  
 Value judgment, 84  
 Variance, 256, 493  
 Variation, independent, 332  
 Variety, 31, 56, 61-2, 73, 81-2, 273, 330  
 Vectors, 498-9  
 Vector-space model, 486, 491  
 Venn diagrams, 91-4, 96-7, 101, 130  
 VERBEEK, L., 103, 121-33, 165, 167-8, 177-8  
 Verisimilitude, 335  
 Visceral function, 62  
 Vision, 280, 323, 489  
 Visual cliff, 323

- Visual field, 279, 320  
 V-machine, 54-72  
 Volley, afferent, 40  
 VON BERTALANFFY, L., 132, 261, 278  
 VON FOERSTER, H. M., xii-xvii, 81, 85,  
   87-9, 131, 227, 248-9, 398, 401,  
   520  
 VON NEUMANN, J., 95, 103, 119, 124,  
   127, 130-1, 135-46, 154, 158-9,  
   165, 178, 205-6, 228  
 von Neumann model, 252  
  
 WADDINGTON, C. H., 56, 80, 83  
 WALL, P. D., 131  
 Wasp, 273  
 WATANABY, S., 482  
 Watch, 270  
 Weather prediction, 447  
 WEAVER, W., 178, 278  
 Weight, of binary sequences, 140  
 Weighting, unanalyzable error-cor-  
   recting, 69  
  
 Weights, variable, 393  
 WEISS, P., 119  
 WERNER, H., 508  
 WESTON, P., 236, 399  
 WEYMOUTH, F. W., 318, 321  
 Wheel of chance, 238  
 Whirligigs, 401  
 WIENER, N., 38, 80, 506  
 WILLIS, D. G., 80-1, 336, 369-83,  
   386, 425, 442  
 WILSON, W., 424  
 Wires, 191, 196, 403  
 WITTGENSTEIN, L., 91, 332-3, 342, 346  
 Work study, 28  
  
 Yield function, 68  
 YOVITS, M., 178, 288, 249-51  
  
 Zollner's illusion, 322  
 ZOPF, G. W., 87, 227, 325-46, 420-2,  
   519

















