CYBERNETIC PROBLEMS IN BIONICS

Bionics Symposium 1966

CYBERNETIC PROBLEMS IN BIONICS

Edited by

HANS L. OESTREICHER
Aerospace Medical Research Laboratories
Wright-Patterson Air Force Base, Ohio

and

DARRELL R. MOORE
Air Force Avionics Laboratory
Wright-Patterson Air Force Base, Ohio

GB
GORDON AND BREACH SCIENCE PUBLISHERS INC.
NEW YORK · LONDON · PARIS

Science Publishers, Inc.,
150 Fifth Avenue, New York, N.Y. 10011

Library of Congress Catalog Card Number 68-19929

Editorial office
for Great Britain:
Gordon and Breach, Science Publishers, Ltd.,
8 Bloomsbury Way,
London, W.C. 1

Editorial office for France:
Gordon & Breach
7–9 rue Emile Dubois
Paris 14e

Distributed in France by:
Dunod Editeur.
92 rue Bonaparte, Paris 6e

Distributed in Canada by:
The Ryerson Press,
299 Queen Street West,
Toronto 2B, Ontario

Foreword

This volume contains a selection of papers presented at the Bionics Symposium 1966 which was held at Dayton, Ohio, May 3–5, 1966 under the sponsorship of the Aerospace Medical Division and the Research and Technology Division of the Air Force Systems Command.

Bionics is concerned with the investigation, modelling, and technological realization of principles and performance characteristics which are of engineering interest. Although there are a great variety of technologically useful properties in living organisms e.g. the investigation of biological propulsion systems, boundary layer control, principles of energy conversion, and many others, this symposium was devoted to the cybernetic aspects of bionics and, accordingly, the papers presented concern themselves with problems in communication, control and information processing by living and technological systems.

It is this area, encompassing the most intricate functions of intelligent behavior, as for instance pattern recognition, concept formation, optimum decision making, self-organization and learning, in which nature through millions of years of evolution is in many respects far ahead of today's technology, and in which research is particularly challenging and rewarding.

The approaches for the derivation of technologically useful knowledge from the study of living organisms are many and there is no single magic formula. In some cases a physiological, biochemical or biomechanical investigation may directly result in a clear-cut model revealing a mechanism of engineering interest, but as a rule it takes considerable additional mathematical, physical and engineering effort and ingenuity to design and develop models which display properties comparable to those in living organisms. This development may well deviate to a large degree from the original biological prototype and result in a system which exceeds

the performance of the prototype in some respects. Sometimes, particularly when attempting to model the highly sophisticated information processing capabilities of the central nervous system, biology may provide not much more than the quantitative definition of the problem and proof for the existence of a solution. The living prototype shows it can be done.

Great progress has been achieved during the past decade in the design of biologically inspired, sophisticated information and control systems. New important concepts have been introduced as a result of a better general understanding of the whole area and many specific problems have been solved. However, in dealing with such extremely complicated problems as, for instance, understanding of the structure and functions of the central nervous system it becomes apparent that the classical methodology will not suffice and the search for more powerful methods to cope with problems, which are by orders of magnitude more complex than the classical problems of science, becomes imperative.

The papers presented in this volume cover the whole spectrum of cybernetics from physiological and biophysical studies to mathematical models and hardware. For the convenience of the reader, they are loosely grouped into four sections:

I. General Aspects of Bionics and Cybernetics

II. Biological Foundations

III. Mathematical Models, Theories, and Methods

IV. Engineering Models and Applications

However, there are no clear lines of demarcation and in order to get the full flavor of this interdisciplinary meeting, the reader should look at the table of contents to be guided to his area of specific interest.

The editors wish to express their appreciation to all those who helped to organize the symposium and prepare this volume. First of all, our thanks are due the authors and the session chairmen, especially John C. Lilly for his highly interesting after-dinner lecture "Research with the Bottlenose Dolphin", W. Ross Ashby, John E. Keto, M. E. Maron, Warren S. McCulloch, Bernard Widrow, and I. C. Whitfield, who chaired the invited paper sessions of the symposium. Gratefully acknowledged are the untiring efforts of Mr. James M. Miller and his staff

from the Research Institute of the University of Dayton who so efficiently organized the meeting and much of the work connected with the preparation of this volume.

Hans L. Oestreicher

Darrell R. Moore

Editors

Contents

I. General Aspects of Bionics and Cybernetics

II. Biological Foundations

III. Mathematical Models, Theories and Methods

IV. Engineering Models and Applications

SECTION I

General Aspects of Bionics and Cybernetics

F. H. GEORGE

University of Bristol and Educational and Scientific Developments Limited

Bristol, England

Logic, Language and the Nervous System

The aim of this paper is to present a general analysis of cognitive behaviour. The analysis is to be essentially cybernetic, and attempts to contribute to our understanding of the way people think and behave, and it is hoped by such means to further our understanding of the way in which human brains work. The paper is also, in a sense, a survey of what has been done in various related disciplines.

It is difficult to define precisely the domain of the survey, since it involves in broad terms, automata theory, hardware models, computer programs as well as philosophical and theoretical attitudes to human behaviour and the underlying mechanisms of the nervous system. We can at least say that a study of meta-mathematics, of switching circuits, of computer hardware and of chemistry of all kinds is omitted. Furthermore what is included represents a selected attempt to provide a systematic approach to the study of the brain. The selection is most obvious where we talk briefly of recent research in neurology of all kinds; this section is only meant to point up a particular attitude and no more.

A POINT OF VIEW

Our understanding of the way in which cognitive functions are performed by the human brain is steadily improving, and this in turn helps, however indirectly, to throw some light on the structure and function of the nervous system.

It is easy enough to formalize some part of what we know of cognition by representing the relevant data as models in the form of finite automata or computer programs and this in turn helps us to build the beginnings of a conceptual model of the nervous system which is a help both to neurophysiologists and neuropsychologists.

Our difficulties spring from three main considerations:

(a) The sheer size of the human nervous system, which, while being a finite automaton, may nevertheless include units with as many as 100,000 inputs and only one highly ramified output; there may also be as many as 100,000 different units in the human nervous system (Cowan, 1962).

(b) Finite automata, whether in theory or in practise, cannot easily be assembled to treat of such an enormous modelling problem; this in turn means that to test a hypothesis, or possible solution to the more general problem, or even sub-problems of the general problem, is very difficult.

(c) The languages we use for scientific model-making are inter-connected with the so-called facts we try and mirror in our theories and models; so it is that the conceptual nervous system of the theorist is only different in degree from the actual nervous system of the human being.

This last point, when fully grasped, can, of course, be an advantage to us, and if we are correct in believing that whatever the complexity of the units, the principles on which the nervous system works may be relatively simple at some level, then there is some hope that we shall continue to progress towards a solution to our problem.

Let us now pursue the problem of methodology a stage further. First let us enquire more as to what our problem is. The author's answer to this is that the general problem for all biologists is a complete functional blueprint of the nervous system (indeed the whole of all organisms for *all* of biology). By the word "functional" we mean to emphasise that had it been an automobile we were studying, we should need more than just the wiring diagram and the plan of the structure, we should also need a precise predictive account of the full range of the automobile's behaviour in a whole range of environments.

The *blueprint* can take various forms, and cyberneticians, physiologists and psychologists are all producing a range of blueprints (models, theories, etc.) in many different ways and at many different levels of complexity. An example of these differences of levels and complexities is brought out by a statement by Warren McCulloch at the N.P.L. Symposium held in London in 1958. He said that a neuron can properly be described by an eighth-order differential equation rather than as a binary unit. The point, of course, is that just as we can *explain* an electric light's behaviour in terms of the throwing of a switch, we could explain it in terms of "current

flow" or "quantum theory", etc. The level of complexity of the explanation should be appropriate to some specific goal. We shall argue that the goal for most purposes of understanding the brain may be assumed at this stage to be satisfied by taking the neuron as a binary unit.

More important perhaps than the level at which such problems are attacked is the realisation that they can be attacked at many different levels of abstraction.

It is no part of my aim here to propose a "philosophy of science" but it is important, if we are to make sense of the kaleidoscope of the many different approaches to biology, to have some idea as to their merging with language and logic. The point is that the cybernetician and the theorist are producing a conceptual nervous system, while the neurophysiologist observes an actual nervous system. But when the *actual* nervous system is *described* (modelled, etc.) it is itself conceptual. It is for this reason that we should increasingly emphasise the similarities between the work of the cybernetician with his synthesis and simulation models, and the various biologists with their "actual" models.

It is now necessary to take the argument one step further. Logic and language are ingredients, with mathematics, statistics, etc., of our descriptive methods. But if we approach the problems of cognition and brain function at the highest level in the human being, we are forced to include the ability to use language and logic within that model. This is a merging of the study of semantics, logic, linguistic philosophy, psychology and physiology, at least, all in the context of the science of *artificial intelligence*. These in turn, when taken together, will represent a hierarchical model, which requires testing at various levels and by various means.

Some of the assumptions underlying the above comments are:

i. The study of the brain is a team study undertaken by a large range of scientists, mathematicians, logicians and philosophers.

ii. It is believed that an understanding of the brain in terms of its electrical, chemical and physiological properties alone is insufficient. It is believed that such an understanding supplements and integrates into our "molar" understanding of the brain's function; both together allow us to build the actual models for testing.

iii. The emergence of integrated teams of scientists will permit of intelligent inter-testing of ideas, and help to eliminate the present wasteful type of isolated brain-modelling.

One of the main points of all this is that we believe that a new use is being found for the "classical" problems of modern philosophy. The analytic-synthetic problem, meaning and truth, synonymity, extension and intension, etc., all take on a new, and we believe hopefully, a highly significant meaning in the new context of *artificial intelligence*.

We must now change our own level of discussion, and turn from the general viewpoint taken so far to examine in more detail some examples of the problems faced by different workers in the field and note, in passing, some of their achievements.

It is impossible to survey such a field as this in detail in less than a book, and that would need re-editing annually; it is difficult even to define it roughly, but we shall hope to achieve a little of each. The ultimate aim must be to create the flavour of what is in some ways a new (cybernetic) look at the problem of the human brain and its relation to behaviour.

METHODS AND MODELS

We have already argued that the methods of description of the scientist are more or less inextricably bound up with the "facts" being described. We can also argue that the features of nature observed, under all conditions of experimental control, are bound up with the interpretations (beliefs, biases, etc.) of the observer. We are thus, from the start, confronted with a fairly complex methodological problem, which, while it can be ignored for some purposes, cannot be ignored altogether.

Crawshay-Williams (1957) had underlined the difficulty in pointing out the philosopher's demand for some sort of "absolute certainty", which contrasts with that type of scientist who operates wholly within a philosophy of (naive) realism, of which he is often unconscious, or at least forgetful. The idea that language (the vagueness of language, etc.) does not affect a branch of science is almost always false, and where it is most true one may often suspect a selection of data for experiment precisely because of its non-linguistic nature. For example, a study of studies of human behaviour shows that the most important feature of human behaviour, which we deem to be language and logic, has been given least detailed analysis.

But let us now look briefly at the more obvious examples of the methods used by the various people who might be deemed to be concerned with a study of the brain. Let us take them briefly in turn:

1. Finite Automata

The main point about the field of finite automata, certainly in its tape and scanner form, such as in the Turing machine, is that it is primarily concerned with mathematics and meta-mathematics and not with behaviour or brain function.

We should say secondly that the primary search is for an effective procedure or algorithm, and not primarily with heuristics. The results of Shepherdson (1959), Rabin and Scott (1959), Wang (1961) and many others can be taken to be typical.

McNaughton (1961) has summarised well the field of automata, and makes a number of distinctions which we would find helpful to follow. Growth automata can become arbitrarily large, but some (*partial*) *growth automata* are limited in the upper limit to which they can grow and turn out to be no more powerful than a *fixed automaton.* We can think of growth automata as being *potentially infinite* in that they can always add on more tape indefinitely as and when they need it, even though at any time t, the amount of tape they possess is clearly finite.

We accept the obvious meaning of synchronous and non-synchronous, deterministic and probabilistic as modifying the noun "automata", and accept as a starting point, McNaughton's definition of an automaton:

> "An automaton is a device of finite size at any time with certain parts specified as inputs and outputs, such that what happens at the outputs at any time is determined by what has happened at the inputs."

McNaughton accepts this as both broad and vague as a definition, and proceeds to refine it. We shall however here quote Rabin and Scott (1959).

Rabin and Scott supply the following definition of a finite automaton:

> "A finite automaton over a finite alphabet E is a system $R = (S, M, s_0, F)$ where S is a finite set (the internal states of R), M is a function of $E \times S$ with values in S (the table of moves of R), s_0 is an element of S (the internal state of R), and F is a subset of S (the designated final state of R).

A theorem that can be shown to follow, after other definitions of class of automata and equivalence etc., are given, is:

Theorem. If $R = (S, M, s_0, F)$ is a non-deterministic automaton, then R is equivalent to the ordinary (deterministic) automaton (T, N, t_0, G)

where T is the set of all subsets of S, N is that function of $E \times T$ such that $N(r, t)$ is the set of all s in S for which there is an s' in t such that s is in $M(r, s')$. $t_0 = s_0$, and finally G is the set of all subsets of S that contain at least one element of F.

Results here are only intended as indicative of what is being attempted and they lead to important points about automata which are requiring of mathematical proof.

The most important point about these results is that they represent an attempt to find mathematically interesting problems, often in an attempt to generalize existing results. So we find multi-tape automata, one way tape automata, Turing machines with only two-state symbols, attempts to provide probabilistic solutions to problems in the field of recursive function theory, and so on and so forth.

The definition supplied by Rabin and Scott varies little from the slightly more general set-theoretic definition (McNaughton, 1961):

> "A finite automaton is an ordered quintuple $\langle S, I, U, f, g \rangle$ where S, I and U are elements, f is a function mapping $S \times I$ into S, and g is a function mapping $S \times I$ into U. Then for every element s in S and i in I, $f(s, i)$ is an element of S, and $g(s, i)$ is an element of U."

We mean this definition to be interpreted so that S, I, U, f, and g are the set of states, set of input values, set of output values, the transition function and the output function respectively.

One specific question of interest that follows from our study of automata is the relative power of a probabilistic and a deterministic automaton. It has proved difficult to make precise comparisons and in one case (de Leeuw *et al.*, 1956) it has been shown that a probabilistic automaton can do no more than a deterministic automaton.

Bearing in mind our own interest in this survey, which is primarily in behaviour and brain models, we would not be especially surprised that this is so. And although, as McNaughton (1961) says, in a broad sense, probabilistic automata can do more than deterministic ones, we would accept, for brain models, the idea that probabilistic automata are automata that approximate to deterministic ones.

We must also be very careful to recognise the fact that the search for a mathematically interesting problem is by no means the same as the search for scientifically interesting facts which can be expressed succinctly in mathematical form. A failure to understand this difference has lead

in the past to a lot of wholly irrelevant criticism from both parties. It can of course, be that some development can be important to both parties, but generally speakng this has not so far turned out to be the case. Since our main search is for brainlike mechanisms and methods for evolving and describing such mechanisms, we may expect, in general, work in meta-mathematics to be of minor (although definite) interest to ourselves.

2. Neural Nets

Since the pioneer work of McCulloch and Pitts (1943) and the developments of Kleene (1951), von Neumann (1952), Culbertson (1950) and Stewart (1966), there has been a lot of activity connected with the neural net version of automata, but perhaps nothing has been done so far that has substantially confirmed early hopes of this as a method for constructing effective tests of modelling concepts.

The work on reliability has some importance, Moore and Shannon (1956), Blum (1962), Cowan (1962), Verbeek (1962), Lofgren (1962a, 1962b), have all written on the subject and have provided a variety of different approaches to the problem. Von Neumann had carried out pioneer work, but had assumed better components than the brain is likely to possess and this has led to a reconsideration of the whole problem of unreliable nets. This has indeed raised the specific question of whether an arbitrarily high reliability can be achieved in a neural net that uses unreliable components and where the nets are not completely redundant.

The solution of the above problem is not yet forthcoming, but the search, significantly, has been among the various logics that have been developed by people with vastly different interests from those here mentioned in the search for brain models (Lewis and Langford, 1932; Post, 1921).

Another aspect of neural nets that has received a great deal of attention is that of reproduction and self-repair. Lofgren's results seem to suggest a finite lifetime for any such system. Indeed he shows that a species can survive longer than an individual member, and that a member can survive longer than all its components and that a self-repairing organism can survive indefinitely if some of the automata's normal input-output conditions are relaxed.

E. F. Moore has shown that if a machine can grow to only a finite maximum size and each of its components has a probabilistic rate of decay, then such a system must be "mortal". But this is again, in view of Lofgren's work, a discovery apparently without far-reaching biological significance. This matter begins now to grade over into biological simulation by automaton theory and is considered further in the next subsection.

Stewart, R. M. (1962), Stewart, D. J. (1960) and others have considered the dimensions of neural nets, and this work has led to the realisation that a "complete" classification system or a complete conditional probability system of the kind envisaged by Uttley (1954, 1955) cannot possibly be biologically realistic.

These considerations have led in turn to the belief (Chapman, 1960; George, 1961), that while classification and conditional probability nets remain realistic ways of modelling the human brain, they need to be severely modified by further principles such as those of *adaptation* and *self-classification*. Indeed the question now arises as to whether the automata or nets need to be *growth automata* or can still be *fixed* and prewired?

It has been argued that automata that grow are representative of organic and other biological processes and are hence essential to models of behaviour. However, it has also been shown that growth automata which do not have unlimited and indefinitely fast growth may be more restricted in their capacities than fixed automata. In either event the capacity to simulate growth on fixed automata suggests that herein lies the greatest potential for the immediate future.

Landahl and Runge (1946) among others have shown how neural nets can be computerised, and this appears to hold out the major hope for the future of this approach. The reason for this state of affairs is simply that unless neural nets can be computerised they cannot generate sufficiently large models to make our essential test realistic. Efforts have been made to simulate anatomically complex, as well as functionally complex, models, but so far little of advantage has accrued.

Another approach to these same, or similar, matters has come through abstract sequential automata (Glushkov, 1962; Ginsberg, 1962; Rosen, 1964) and these hold out promise that perhaps a further set of principles might be so demonstrated. George (1965) has demonstrated that scientific induction can be achieved by conditional probability computing, and

in essence this is a further result derived from the field of sequential automata. This particular work will be referred to again in the discussion of Section 4.

3. Finite Automata and Biological Simulation

There have been a number of studies carried out explicitly in the field of biological simulation and finite automata, and much of this work is admirably summarised by Apter (1966).

Apter is concerned with simulating the process of development and he uses for this purpose both a computer model and a Turing machine type model. The aim is to study morphogenesis; it is recognised that the principles of organisation and control governing the development of complex organisms such as humans are likely to be vital factors in biological development. The work by von Neumann (1951, 1952), on self-reproduction is well known and emphasises, as do Burks (1957), and Wang (1961), the vital factor of growth to automata theory as well as to biology. Burks and Wang regard Turing machines and self-reproducing automata as special cases of growing automata. This raises the question of the importance of growth models generally. The present writer's view is that much of human behaviour and much of the organisation of the brain can be studied and modelled by "non-growth" automata whereas equally it is true to say that if the subject to be studied is one such as self-reproduction or morphogenesis, then growing automata may prove vital. However, it should be remembered that "growth" can be simulated on a computer; much, here however, depends upon the accepted meaning of the word "growth".

Von Neumann (1952), Penrose (1959), Moore (1961), Jacobson (1958) and others have all constructed, either in hardware or software, models of self-reproduction, but mostly the methods are inadequate to the process of growth and development. As Apter (1966) says of von Neumann's work:

> "... the first machine is making the second machine, not the second machine making itself from instructions given to it by the first."

Apter's own work involves development algorithms, as he calls them. These are expressed in Turing machine terminology with the additional

conventions that R means that the set of quadruples (machine) is itself reproduced, that the output of a machine is the input at the next moment of any machine reproduced by that machine, that there is a continued input of $S_1 s$ to the original machine, and that the tape moves automatically in one direction at the rate of one square of tape per moment.

For example, given the quadruples

$$q_1 s_1 R q_2$$

$$q_2 s_1 s_1 q_2$$

where R means "replicate the automaton", we can easily achieve linear growth as depicted in the following instantaneous description: –

$q_1 s_1 \rightarrow q_1 q_2 s_1 \rightarrow q_1 s_1 q_2 s_1 \rightarrow q_1 q_2 s_1 q_2 s_1 \rightarrow q_1 s_1 q_2 s_1 q_2 s_1 \rightarrow q_1 q_2 s_1 q_2 s_1 q_2 s_1$
etc.

From these simple beginnings, and using the same familiar "Turing machine" terminology, he develops various patterns and growth shapes. Apter also proposes a general model for pattern regulation in the organism, and puts the model in terms of the recently developed concepts of Jacob and Monod (1961) concerning gene action. We shall however not pursue this subject further here.

We can summarise the position thus – biological models hold out high hopes for the development of conceptual models of a wide range, not just of the nervous system. They also hold out hope that either as computer programs or as paper-and-pencil models they will mirror successfully what may turn out to be the vital property of *growth*.

One should add though that as in the case of finite automata and neural nets, there is likely to be little more than the demonstration of relatively isolated principles until such time as we find and prove methods for large scale simulation.

4. Hardware Neurons and Neuronlike Systems

The problem of simulating a neuron in hardware has existed for a very long time, and Lillie's nerve model is still a reminder of this fact. More recent attempts to provide such models focus attention on the work of Harmon. Harmon's (1961) neuromimes are well known applications of "black box theory" in the neurophysiological field.

Harmon's (1961) work involves the construction, by electronic circuits, of a system which simulates many of the known properties of the neuron. These neuromimes are then made the subject of further experimentation both as individual units and as collections. Von Bergeijk (1961) has shown how such neuromimes can simulate different features of the cochlea, while Levinson and Harmon (1961) have shown how the psychological and physiological properties of flicker-fusion can be brought together in one model.

The models of Zemanek, Kretz and Angyan (1961) come nearer to bridging the gulf between logical or neural nets and the hardware neurons. They constructed these models in hardware and they were able to show a degree of adaptive behaviour. At this point we are reminded of a whole mass of models which can produce different degrees of adaptation (Walter, 1953, Ashby, 1952, George, 1961, Uttley, 1954, 1955, Shannon, 1951, Chapman, 1959, Stewart, 1959, and many others besides). More recently Scott (1964), among others, has analyzed the properties of nerve tissue in mathematical terms.

What we should say at this point is that there is no evidence that these models can carry us beyond the first stage of demonstrating that adaptation, of a more or less sophisticated kind, can occur and can take many different forms.

We still need a design for a neuron that is cheap and capable of being the basis for a system large enough to be behaviourally and neurologically interesting. Research in chemical stores and chemical growth (MacKay and Ainsworth, 1964, Pask, 1958) holds out hope for the future. Microminiaturised circuitry holds out further hope, but so far it seems clear that the gap between achievement to date and desired achievement is still very wide.

5. General Hardware and Software Models

More general models, both in software and hardware, are numerous in cybernetics and we shall mention but a few.

Farley and Clark (1961) have demonstrated some of the electrical characteristics of the nervous system by suitably programming a computer. This is an attempt to increase our understanding of the functional picture of sets of neurons or nets.

Beurle (1962) has studied these same functional characteristics rather differently and has studied random nets of actual neurons. There is again some progress from such a standpoint, and Beurle is able to show a great deal of humanlike trial-and-error and learning behaviour, and even more speculative behaviour.

Turing also wrote on randomly connected nets and discussed the manner in which organisation gradually emerged. This is presumably one stage removed from the development of the individual since his starting point is by no means random.

Hebb formulated a theory of cell assemblies (1953) which was later modified by Milner (1957) and this represented an attempt to show an effective growth process could account for learning and other modifications in the brain which were correlated within cognitive activities. Rochester (1959) simulated the cell assembly in hardware and showed its inadequacies and since then Minsky (1963) has criticized the cell assembly as being irrelevant to our understanding of the brain's activities. It should be noted that Hebb's work has had a great appeal to psychologists and it should have been the case that its widespread appeal would be transferred to the field of neural nets and automata theory. By and large this has not been the case and it is unfortunate that psychologists have missed what seemed to be a major opportunity for advancement. They have, however, pursued the simulation methods of Newell, Shaw and Simon (1959, 1960, 1963) and this has followed a somewhat similar course what might have been followed by applied finite automata, although it is more concerned with processes and less concerned with neural mechanisms.

NEUROPHYSIOLOGICAL EVIDENCE

It would be absurd to attempt a survey of all the evidence that has been built up recently regarding our understanding of the function of the nervous system. It is too vast a subject for this, and we shall attempt no more than to illustrate with a few examples that seem to represent important trends.

In the field of conditioned reflexes, Doty (1965) showed that electrical stimulation of brains in macaques had systematic results of an interesting

kind. He used 0.2–1.0 m-sec. pulses as conditioned stimuli which evoked lever pressing conditioned responses. He found that if a point of the striatal cortex was stimulated, then the lever pressing started. He then found that this conditioned response could also be elicited by stimulation of other points in the striatal cortex, even from the contralateral area.

Sharlock, Neff and Strominger (1965) showed that cats could be trained to discriminate differences between sequences of tones. They then carried out bilateral ablation of the auditory cortex and followed it with a retention test. They worked out a map showing that certain areas of the cortex when destroyed, not only destroyed the retention but stopped relearning.

Whitfield and Evans (1965) studied the effect of frequency-modulated stimuli on the auditory cortex of the cat. Among many other results they showed that some units had a much wider range of response for frequency-modulated tones than for steady tones. They also noted that the behaviour of the majority of units could not be predicted from the steady-state data.

Simmes and Mishkin (1965) studied the effect of ablation of the sensory-motor region on monkeys. They were tested on a tactile discrimination. When using the ipsilateral hand they found their learning of difficult form discrimination was retarded, and they were less sensitive to rough surfaces, although not to different sizes of object.

These experiments have been cited more or less at random, but they have important consequences methodologically and indeed for our understanding of the brain itself. It goes without saying that the various people working in the field of neurophysiology and allied fields of electroencephalography, clinical neurology and the like should attempt an interpretation of their results. Indeed they do so and the picture emerging is becoming clearer all the time.

An example of another recent development is that of Adey (1961) who implanted electrodes in the temporal lobes of cats and analyzed records of their changes on a computer. He discovered certain characteristic wave patterns—one he called an "approach rhythm" which he associated with learning. Adey has suggested that learning is associated with spatial pattern changes in neuronal currents, which reminds us once again of Hebb, Milner and the work on cell assemblies—surely the time has come for another concerted attack to build this sort of model—and build it and test it on a computer.

We have no space to discuss the reticular formation (Schade and Ford, 1965), but its importance to human behaviour is well known, even if the details are not well understood.

We now have some idea of the relation between the behavioural state of man, the state of awareness and the electroencephalographic record (Schade and Ford, 1965). For example, in a state of alert attentiveness, where the subject describes himself as concentrating, there is an associated EEG record of characteristic partially synchronized low-amplitude waves. In deep sleep on the other hand, where there is no consciousness, there is an EEG record of large and very slow waves, with random irregular patterns.

In a state of strong emotions, which we know can be artificially stimulated, there are variously described states of confusion, divided "attention", etc., with an associated EEG record of a desynchronized kind and of low to moderate amplitude. There are also fast mixed frequencies which accompany these highly emotional states.

Schade and Ford (1965, page 317) summarise in very general terms the simple interaction of the drive system and conscious behaviour. This is illustrated by a diagram due originally to Stellar (1960). One of the interesting features of this diagram is the essentially cybernetic point of view adopted.

Stellar views the concept of drive as that of lowering of thresholds for response patterns. But he also assumes that the degree of motivated behaviour varies directly with the activity of certain excitatory centres in the hypothalamus. Stellar assumes the interaction of inhibitory and excitatory hypothalamic centres. The external world can, of course, modify this hypothalamic state through sensory stimuli, and internal states influence this hypothalamic state through the vascular system. Finally, it is assumed that cortical and thalamic influences exert further excitatory and inhibitory influence on the hypothalamus.

We know there are certain changes in neurophysiological states which are correlated with learning. Thinking too is related to neurological changes in the temporal, parietal and frontal areas. Furthermore, as Schade and Ford put it:

> "In general one can say that the speech mechanisms probably form a condition for certain kinds of thinking processes and provide a special way for handling of information."

Thinking and problem solving are known to be related, but perhaps it is too much to expect that we can yet hope to distinguish one from the other at the neurological level, even if they are significantly distinguishable at all, except at the verbal level.

There is still much to be done to relate neurological states to behavioural states and consciousness, but the picture continues to develop in time.

LANGUAGE AND LOGIC

The main field in which the writer has operated over the last five years is that of language and logic and their relation to behaviour, so it is natural that this section should turn out to be the most detailed.

Language behaviour has been the most neglected of all aspects of behaviour by experimental psychologists. It is a field that has been largely left, until very recently, to philosophers and that group of philosopher-scientists such as Peirce, Morris, Dewey and others who have tried to reconstruct philosophy—including logical and linguistic problems—within the context of science. It is in some measure in their tradition that we hope to approach problems of logic and language.

As a preliminary to our own efforts in providing computer programs for logic and language, as an aspect—indeed a central aspect—of artificial intelligence, we should mention a number of other workers in the field. Chomsky is well known for his work on the phase structure of language and his reduction of statements to a series of kernel types is most relevant to our own procedures.

Edmundson (1963), Garvin (1963), Hayes (1963), Swanson (1963), and many others have also contributed to this picture of language as a type of measurable human activity. A form of human activity that must also reveal something of the nature of the organisation of the human brain.

We should now, in leading in to the linguistic aspects of artificial intelligence, pay due respect to the work of Bruner, Goodnow and Austin (1956), Banerji (1964), Minsky (1963), Newell, Shaw and Simon (1963) Samuel (1963), Gelernter (1963), Windeknecht (1965) and a whole host of others.

Snediker (1964) has suggested methods for information retrieval which bears closely on the fundamental techniques of concept formation.

Maron and Kuhns (1960), Maron (1961), have also, among others, made contributions which are synthesised in our own work in Bristol.

The outline given in this section of linguistic analysis and concept formation on a digital computer represents the author's own approach to these problems. It should be mentioned that other similar approaches have been developed by other members of the Bristol group (Sarkar, 1964, 1966; Berger, 1966). Also the recent paper of Levien and Maron (1965) represents an approach which is extremely similar to the one tackled by the Bristol group.

We shall be drawing attention to various points of difference as we carry out our overall summary and we shall now proceed to this end.

1. Language and Artificial Intelligence

We now consider the problem of a computer C_1 asking a question or making a statement, and the problem of C_1 being asked a question and having a statement made to it.

The forms of the language must include:

a. Ordinary English
b. Canonical forms of English
c. Empirical Logic and
d. Machine Code

We shall not here attempt to go into detail of coding, but merely illustrate how language processing can occur. We start with the problem of synthesis.

Given a statement S_1, such as "I will see you tomorrow and tell you about Jack then". We must first assume lists of words L_1, L_2, ..., L_6 made up as follows:

L_1 = Nouns
L_2 = Pronouns
L_3 = Adjectives
L_4 = Verbs
L_5 = Adverbs
L_6 = Connectives

So we process S_1 by identifying I from L_2 as a pronoun and substituting C_1 for I. "will see" is (future tense) of see and is identified in L_4. "you" is from L_2, and we substitute a proper name (e.g. C_2), Jack is from L_1,

and "tomorrow" is from L_1 and "and" is a connective, so we substitute a full stop, and we have the canonical form:

$$C_1 \text{ (will) see } C_2 \text{ (tomorrow),}$$

or

$$C_1 \text{ (will see tomorrow) } C_2$$

which is symbolized in our empirical logic:

$$S_7^* C_1 C_2$$

where S_7^* = "will see tomorrow".

We can now store $S_1 (\equiv S_7^* C_1 C_2)$ and we store it in lists such as temporary stores dealing with a plan for tomorrow. We store it (or cross reference this) with respect to C_2 and have all (relevant) statements concerning C_2 available in our temporary store. After we reduce the rest of S_1 to S_1^2, we shall have our criteria of relevance, since we shall have:

$$C_1 \text{ (will tell you then about) (Jack)}$$

we assume "Jack" is known and is sufficiently identified. We shall then have a logical version such as:

$$T_8 C_1 G_1$$

and all statements incorporating G_1 will also be brought into temporary store.

It is clear that some class name or individual name such as "game" or "C_2" will be the principal feature of each canonical form which will have one verb only, and we may (nearly) achieve a standardization by reduction of each system to one of Chomsky's eight kernel types and to do so requires only that the computer can recognise the type of word (e.g. "he" is a pronoun) that occurs in a statement and by a comparison with its own fixed lists.

We shall now outline the computer operation in compiler terms.

We need first a REDUCTION—this is our compiler word which means "reduce all statements form register X to END to a canonical kernel", so

$$\text{RED(X)}$$

is our shorthand code word. Similarly canonical forms are TRANSLATED into logical form. So we have

TRA(m, Y)

which is read as "translate m canonical forms starting at address Y".

Store these m forms, and the store is either selected by a search for the term name in each kernel, or by a list-processing technique (Wilkes, 1965, McCarthy, 1961).

We can now indicate very generally how the Computer C_1 uses the language in our four cases:

Case 1.: C_1 makes statement
Case 2.: C_1 asks question
Case 3.: C_1 told statement
Case 4.: C_1 asked question

C_1 makes a statement in answer to a question, where the translated question suggests statement to be made:

e.g. "Where is X?"
gives "X is (there)"

C_1 asks a question, when asked and does not know the answer, or when computing and has incomplete knowledge.

When C_1 is told something, it will check source reliability (and ideally also motive) and will store the statement. If C_1 is asked a question it will give an answer; it could be made to lie or prevaricate, but this becomes complex in a computer; we must certainly have criteria for prevarication. The very need for criteria suggests that the computer is operating, if like a human being, in having a hierarchy of goals and subgoals. These goals and subgoals must, of course, provide all such criteria.

Without considering bluffing, prevarication and the like, computer programs have been written which perform all the other operations.

But let us cut short this brief account of the synthesis of natural language and ask about the simulation problem. No programs have been written by the author with a view to simulating linguistic behaviour, but simulation has been used as a basis for synthesis, and synthesis sheds some light on simulation.

We could perhaps at this early stage recognise very clearly the vital importance of recognition and categorization. The computer C_1 must recognise how to process an input by recognising what the input actually is. It must recognise what the appropriate mode of processing is at each stage of the total operation.

It is clear that for synthesis, an input can be formally recognised as linguistic or otherwise by the use of a single digit, and it can similarly be so recognised, given that it is linguistic, as being either a statement or a question. But for simulation, it is essential for the person (simulated by the computer) to detect whether a message symbolizes or signifies something or not. In a formal sense, it is easy enough to understand when you are being spoken to, but more difficult to decide in such a case, as hearing a tapping on the wall or the ceiling, whether what is heard is *intended* to symbolise or signify.

Gestures from human sources can amplify meaning in a behavioural context, and such symbolization and signification can be absolutely clear in general, if often vague and ambiguous in detail.

After the input has been recognised as such, there still remains further "recognition stages" in the processing. The computer or human must recognise an appropriate response and must recognise when an appropriate response is not available. It must recognise how to tackle a problem, and must recognise that one problem is similar to (or different from) another. It is difficult at this point not to accept readily the overall plan of Miller, Galanter and Pribram (1959), and say that this is a problem of programming.

If we accept, as we do, the general edict of the above authors, we must argue in some measure about the detail, since the TOTE operation still seems too limited a view of the process of selective stimulation and reinforcement.

We must continue to probe the essential differences between our synthesis programs and the simulation programs that are specifically needed. We need also to be aware of what if anything occurs to help us understand the probable workings of the brain.

Separate research work has been done in these allied fields (George 1962, 1965a, 1965b, 1966a, 1966b) which shows different aspects of logical cognitive behaviour and suggests more of the possible mode of operation of the human brain. It should, however, be added that this was in no case the motive behind the work.

2. Compiler Languages and Language Behaviour

By way of illustration of a different approach to the problem of language, we should mention the work of Michie (1965), Foster (1965), Brooker and Morris (1963), Napper (1965). We shall say a few words about the last named who has taken the most different approach of all the above-named to exactly the same problem as the author.

Brooker and Morris developed a compiler-compiler language to assist compiler writers in writing compiler languages. Napper then provided yet another higher order compiler language which allows the writing of a program in any language at all, provided you define your language at the same time as you use it. This means that he is using the compiler-compiler language in a dynamic fashion and expanding his vocabulary as the need arises.

Napper arranges for the format to be read into the computer when translating. The formats can be made up by a set of further formats, and the compiling or compiler-compiling is switched off while this defining (or redefining) is taking place. The format routines help the translation by substituting one set of instructions for another.

He arrives at forms such as

set (PROPERTY) (OF) (Item) (=) (item/1)

where the property is any capital letter word and item and item/1 are two lower case words. In square brackets we have *Class Words*, and *Phrases* are used for describing the language, such as in the use of class words, so that, for example, we can handle synonyms, e.g.

PHRASE (IS) = is, are, was, were, will be.

a *Format* is a phrase or unit of language (sentence or clause) which is laid out in a particular form, and each different format starts with a different *imperative*, while a *Format Routine* gives you instructions for the appropriate use of Formats.

This type of approach is not directly concerned with the language-data relations which are a part of the author's approach. It is rather concerned with the development of language, by definition in further language, which relates all verbal activity to computing as such. It seems to the author that his own approach is concerned more with developing a relationship between computer operations (logic), language, and the na-

ture of the *external* environment to which much of the language will refer. In a sense it seems that these two methods are complementary and must (or could) lead to the same sort of end state. From our point of view in this article, it at least underlines the vital importance of language to intelligent behaviour and brain function.

Let us now finally look at a few tentative generalizations which might be drawn from the forgoing work and let us pretentiously call it a "brain model".

A BRAIN MODEL IN OUTLINE

It may seem ridiculous to make any general statements about the brain on the basis of this brief and rather cursory summary article. We shall, however, try and draw up a sort of outline and make some suggestions as to how it might be tested and generally followed up.

The brain is assumed to be a data-processing system with a very large store and is in essence working on coded information travelling along the pathways of the nervous system. The sensory pathways are primarily concerned with input and this input activity is integrated into the complex selectively reinforced processing of the brain. It seems clear that sensory inputs are independent classification systems (no doubt, partial classification only, occurs) which act as filters to a higher integrated partial classification system which is the store.

Information stored in the brain occurs at various levels of generality and we may expect to distinguish short term from long term memories.

There is in existence more and more evidence that suggests the cortical locations for these memory stores, and we are acquiring new information all the time. Our view is that we shall return to the notion of a high level of cortical localisation, but think of it in dynamic terms. Information (cf. list processing) is stored in overlapping manner and detailed information, like data placed in computer registers, may vary from person to person and even from time to time in the same person.

The reticular formation seems clearly to be associated with motivation which is itself a complicated activity concerned with selection, activation, priorities, emergencies, etc., all within the compass of a homeostatic principle subserving the organic needs of the body for survival. The limbic system is almost certainly especially connected with this motiva-

tional type of activity. The limbic system probably regulates the disposition of organisms by use of neural homeostats (Pribram, 1960) and the notion of motivation must occur on levels other than the merely immediate reinforcement of a food need for example.

The reflex arc or the simple conditioned reflex is something of an artefact as far as the human brain is concerned, although there is now a great deal of research (e.g. Doty, 1965) which, while using conditioning terminology, draws attention to greater detail of the workings of associations and classifications in store. This work also suggests the very close association between equivalent areas in different hemispheres.

If visual information is processed initially in the occipital areas, auditory information in the tempo-parietal areas, etc., the regions we would expect to find processing language are the so-called "speech areas". Here we have a problem since speech for the brain is complex. Human beings can "vocalize" as a motor activity of a relatively simple kind. They symbolize as a very high level conceptual activity. Sounds can be formalized as words and sentences and their utterance leads to an auditory response, which leads to translation into symbols.

Language is not only motivated, but itself motivates. Words may remind us of, or initiate needs which we had forgotten or suppressed!

Language is also associated with imagery. When we "image" (cf. *imagine*) we somehow stimulate some subset of the total set of sensory inputs which are involved in the actual sensory experience.

One thing seems clear, we learn to associate noises (words) with objects, relations, etc. and we must assume that the human brain stores language and data separately, but intimately associated by something like a list-process or a complex cross-reference system.

We can "image" an event from the past and we can make statements about (describing) that event. In fact we tend to speak events as we "image" them, and can hardly speak about something without "imagnig" it.

Conceptualizing is intimately bound up with language and is thought to be a function closely associated with the frontal areas of the brain. The hierarchical nature of the brain suggests that the generation of new principles, recursion formulae, meta-rules, etc., which allow the solution of new problems is primarily a frontal lobe activity.

Cybernetic research, particularly from the domains of computer synthesis and simulation suggests the brain is a complex store with the

ability to make inferences and compute, where the computational activity of mathematics has to be learned and conceptualized. The question is why is there not a higher correlation between cortical damage and impairment of function. Search for memory traces of a dynamic character have been largely abandoned so we must, it seems, accept instead the fact that the detailed information contained in any one cortical area varies according to the order in which it has occurred in the history of the owner, or at least partly by this and partly by his cross-reference system. In other words, information is contained in the detailed structure of the nervous system in a way which allows complicated overlapping of both functions and detail.

One way of fitting in the detail and testing the essential rightness of the views expressed is for a team of scientists (nothing less would do) to build a large and detailed simulation of the brain on a computer (or series of computers), filling in all the detail that is known with some measure of confidence, and adding in plausible details until the model is fully connected and effective. The test is then that of seeing whether it, as a brain (control) system, can successfully program its environment – perhaps in the same way that human brains do?

REFERENCES

Adey, W. R. (1961). Studies of hippocampal electrical activity during approach learning. *Brain Mechanisms and Learning* (Ed. J. F. Delafresnaye), Blackwell, Oxford.

Apter, M. J. (1966). *Cybernetics and Development*, Pergamon Press, Oxford.

Ashby, W. R. (1952). *Design for a Brain*, Chapman and Hall, London.

Banerji, R. B. (1964). "Computer Programs for the Generation of New Concepts from Old Ones." In *Neuere Ergebnisse der Kybernetik*. (Eds. K. Steinbuch and S. W. Wagner) Oldenbourg, pp. 336–343.

Baumgartner, G., Brown, J. L., and Schulz, A. (1965). Responses of single units of the cat visual system to rectangular stimulus patterns. *J. of Neurophysiology XXVIII*, **1**, 1–18.

Berger, S. Heuristic Programs for Planning (to be published).

Beurle, R. L. (1962). Functional Organisation in Random Networks. In *Principles of Self-Organization*, 291–314.

Blum, M. (1962). Properties of a neuron with many inputs. In *Principles of Self-Organization* (Ed. H. von Foerster and G. W. Zopf), pp. 95–120.

Brooker, R. A., Morris, D., and Rohl, J. S. (1963). The Compiler-Compiler. *The Annual Review in Automatic Programming* (Ed. R. Goodman), Academic Press, New York.

Bruner, J. S., Goodnow, J. J. and Austin, G. A. (1956). *A Study of Thinking*, John Wiley, New York.

Burks, A. W., and Wang, H. (1957). The Logic of Automata. *J. Assoc. Computing Machinery* **4**, 2; **4**, 3.

Chapman, B. L. M. (1959) A Self-Organising Classifying System. *Cybernetica*, pp. 152-161.

Cowan, J. (1962). Many-Valued Logics and Reliable Automata. In *Principles of Self-Organization* (Ed. H. von Foerster and G. W. Zopf), pp. 135–180.

Crawshey-Williams, R. (1957). *Methods and Criteria of Reasoning*, Routledge and Kegan Paul, London.

Culbertson, J. T. (1950) *Consciousness and Behaviour*, Brown Dubuque, Iowa.

Dotey, R. W. (1965). Conditioned reflexes elicited by electrical stimulation of the brain in macaques. *J. of Neurophysiology XXVIII*, 623–640.

de Leeuw, K., Moore, E. F., Shannon, C. E., and Shapiro, N. (1956). Computability by probabilistic machines. In *Automata Studies* (Eds. Shannon, C. E., and McCarthy, J.), pp. 183–212. Princeton University Press, Princeton, New Jersey.

Edmundson, H. P. (1963). A Statistician's View of Linguistic Models and Language-Data Processing. In *Natural Language and the Computer* (Ed. Garvin, P. L.), McGraw-Hill, New York.

Farley, B. A., and Clark, W. A. (1961). Activity in networks of neuron-like elements. In *Information Theory* (Ed. Cherry, E. C.), 242–251. Butterworth, London.

Foster, J. M. (1965). Interrogation Languages. Paper read at *Conference on Computers*, Edinburgh.

Garvin, P. L. (1963). A Linguist's View of Language-Data Processing. In *Natural Language and the Computer*, McGraw-Hill, New York.

Garvin, P. L. (1963). The Definitional Model of Language. In *Natural Language and the Computer*, McGraw-Hill, New York.

Gelernter, H. (1963). "Realization of a Geometry-Theorem Proving Machine." In *Computers and Thought*. (Eds. E. A. Feigenbaum and J. Feldman), McGraw-Hill New York.

George, F. H. (1961). *The Brain as a Computer*, Pergamon Press, Oxford.

George, F. H. (1962). Simple Adaptive Programs for Computers. Paper read at *Conference on Cybernetics*, U.C.L.A., October.

George, F. H. (1965a). The Use of Models in Science. Paper read at *Conference on Scientific Method*, University of Cambridge, July.

George, F. H. (1965b). Computer Applications in Decision Taking and Process Control. Paper read at *International Symposium on Long Range Planning for Management*, Paris, September.

George, F. H. (1966a). The Development of Computer Assisted Instruction and Problems of Programming. Paper read at *National Programmed Learning Conference and Exhibition*, Loughborough, March.

George, F. H. (1966b) Hypothesis Confirmation on a Digital Computer. Paper read at Bionics Symposium, *Dayton, Ohio.* This volume pp. 714–723.

Glushkov, V. M. (1962). *Synthesis of Digital Automata,* Moscow.

Ginsberg, S. (1962). *An Introduction to Mathematical Machine Theory, Reading,* Mass. Addison-Wesley.

Goldacre, R. J., and Bean, D. A. (1959). Electronic Models in the Study of all Interaction. *Discovery,* **20** (7), 277.

Harmon, L. D. (1961). Studies with artificial neurons I. Properties and functions of an artificial neuron. *Kybernetik* **1**, 3, 89–101.

Hayes, R. M. (1963). Mathematical Models in Information Retrieval. *In Natural Language and the Computer,* McGraw-Hill, New York.

Hebb, D. O. (1949). *The Organisation of Behaviour,* Wiley, New York.

Jacob, F., and Monod, J. (1961). On the Regulation of Gene Activity. In *Cold Spring Harbor Symposia on Quantitative Biology,* **26**, pp. 193–211.

Jacobson, H. (1958). On Models of Reproduction. *American Scientist,* **46** (3), 255–284.

Kleene, S. C. (1951). Representation of Events in Nerve Nets and Finite Automata. RM-704.

Landahl, H. D. and Runge, R. (1946). Outline of a Matrix Calculus for Neural Nets. *Bull. Math. Biophys.* **8**, 75–81.

Levien, R., and Maron, M. E. (1965). Relational Data File: A Tool for Mechanized Inference Execution and Data Retrieval. RM-4793-PR.

Lewis, C. I., and Langford, C. H. (1932). *Symbolic Logic,* The Century Co., New York.

Levinson, J., and Harmon, L. D. (1961) Studies with Artificial Neurons III. Mechanisms of Flicker-Fusion. *Kybernetik,* **1**, 3, 107–117.

Lofgren, L. (1962a). Limits for Automatic Error Correction. In *Principles of Self-Organisation* (Eds. H. von Foerster and G. W. Zopf), pp. 181–228.

Lofgren, L. (1962b) Kinematic and Tessellation Models of Self-Repair. In *Biological Prototypes and Synthetic Systems,* pp. 342–369, Plenum Press, New York

MacKay, D. M., and Ainsworth, W. A. (1964). Electrolytic Growth Processes with Applications to Self-Adjusting Automata. In *Neuere Ergebnisse der Kybernetik* (Eds. K. Steinbuch and S. W. Wagner), pp. 326–335.

Maron, M. E. (1961). "Automatic Indexing: An Experimental Inquiry." *Assoc. Computing Machinery,* **8**, 3, pp. 404–417.

Maron, M. E. and Kuhns, J. L (1960). "On Relevance, Probabilistic Indexing and Information Retrieval." *Assoc. Computing Machinery,* **7**, pp. 216–244.

McCarthy, J. (1961). A Basis for a Mathematical Theory of Computation. *Proceedings of the Western Joint Computer Conference,* p. 282.

McCulloch, W. S., and Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Neuron Activity. *Bull. Math. Biophys.* **5**, 115–133.

McNaughton, R. (1961). The Theory of Automata, A Survey. In *Advances in Computers* (Ed. F. L. Alt), Vol. II, Academic Press, New York.

Michie, D. (1965). Scope and Limitations of Computers for the Medical User: The MINI-MAC Project. *Experimental Programming Report,* No. 6, Edinburgh.

Miller, G. A., Galanter, E. H., and Pribram, K. H. (1961). *Plans and Structure of Behaviour*, Holt, New York.

Milner, P. M. (1957). The Cell Assembly: Mark II. *Psychological Review*, **64**, 242–252.

Minsky, M. L. (1963) Steps Towards Artificial Intelligence. In *Computers and Thought* (Eds. E. A. Feigenbaum and J. Feldman), pp. 406–452. McGraw-Hill, New York.

Moore, E. F., and Shannon, C. E. (1956). Reliable Circuits Using Less Reliable Relays *J. of Franklin Institute*, **262**, 191–208 and 281–298

Moore, E. F. (1961) Machine Models of Self-Reproduction. Paper presented at *Symposium on Mathematical Problem of Logical Sciences*, New York.

Napper, R. B. E. (1965). The Third-Order Compiler: A Context for Free Man-Machine Communication. *Conference on Computers*, Edinburgh.

Newell, A., Shaw, J. C., and Simon, H. A. (1959). Report on a General Problem Solving Program. Proceedings of the *International Conference on Information Processing*, Paris.

Newell, A., Shaw, J. C., and Simon, H. A. (1960). A Variety of Intelligent Learning in a General Problem Solver. In *Self-Organising Systems* (Eds. M. Yovitts , and S. Cameron), Pergamon Press, Oxford.

Newell, A., Shaw, J. C., and Simon, H. A. (1963). The Theorem Proving Machine. In *Computers and Thought* (Eds. A. Feigenbaum and J. Feldman), McGraw-Hill, New York.

Pask, A. G. S. (1959). Physical Analogues to the Growth of a Concept. In *Mechanisation of Thought Processes*. Her Majesty's Stationery Office, London.

Penrose, L. S. (1959). Self-Reproducing Machines. *Scientific American*, **11**.

Post, E. L. (1921). Introduction to a General Theory of Elementary Propositions. *Amer. J. of Math.* **43**, 163–185.

Pribram, K. H. (1960). Theory in Physiological Psychology. *Annual Review of Psychology*, **11**, 1–40.

Rabin, M. D., and Scott, D. (1959) Finite Automata and their Decision Problems. *I.B.M. J. Res. Develop.* **3**, 114–125.

Rochester, N., Holland, J. H., Haibt, L. H., and Duda, W. L. (1956). Test on a Cell Assembly Theory of the Action of the Brain using a large Digital Computer. *I.R.E. Transactions on Information Theory*, IT-2(3) 80–93.

Rosen, R. (1964). Abstract Biological Systems as Sequential Machines. *Bull. Math. Biophys.* **26**, 103–111.

Samuel, A. L. (1963). "Some Studies in Machine Learning Using the Game of Checkers." In *Computers and Thought* (Eds. E. A. Feigenbaum and J. Feldman). McGraw-Hill, New York.

Sarkar, P. (1963). *The Cybernetics of Business Systems*. Thesis, The College of Aeronautics, Cranfield.

Sarkar, P. *Automata Theory and Heuristic Computer Programming*. (To be published).

Scott, A. C. (1964). Analysis of a myelinated nerve model. *Bull. Math. Biophys.* **26**, 247–254.

Semmes, J., and Mishkin, M. (1965). Somatosensory Loss in Monkeys After Ipsilateral Cortical Ablation. *J. of Neurophysiology XXVIII*, 473–486.

Schade, J. P., and Ford. D. H. (1965). *Basic Neurology*, Elsevier, Amsterdam.

Shannon, C. E. (1951) Presentation of a Maze-Solving Machine. In *Cybernetics*. Transactions of the Eighth Conference of the Josiah Macy Jr. Foundation (Ed. H. von Foerster), pp. 173–180.

Sharlock, D. P., Neff, W. D., and Strominger, N. L. (1965). Discrimination of Tone Duration after Bilateral Ablation of Auditory Cortical Areas. *J. of Neurophysiology XXVIII*, pp. 673–681.

Shepherdson, J. C. (1959). The Reduction of Two-Way Automata to One-Way Automata. *I.B.M. J. of Research and Development*, **3**, 2, 198–200.

Snediker, J. M. (1965). "A Generalized Program for Information Retrieval." Systems Research Center Report SRC 77–A–65–29, Case Institute of Technology, Cleveland, Ohio.

Stellar, E. (1960). Drive and Motivation. *Handbook of Physiology*, Section 1. Neurophysiology III (Eds H. W. Field, H. W. Magoun, V. E. Hall). Washington D.C.

Stewart, D. J. (1960). *Automata and Behaviour*, Ph. D. Thesis, University of Bristol.

Stewart, D. J. *Automata Theory and Neural Nets*. (To be published.)

Stewart, R. M. (1962). Theory of Structurally Homogeneous Logic Nets. In *Biological Prototypes and Synthetic Systems* (Eds. E. E. Bernard and M. R. Care), pp. 370–380, Plenum Press, New York.

Swanson, D. R. (1963). The Formulation of the Retrieval Problem. In *Natural Language and the Computer*, McGraw-Hill, New York.

Turing, A. M. (1952). The Chemical Basis of Morphogenesis. *Phil. Trans. Roy. Soc.* **237**, 37–72.

Uttley, A. M. (1954). The Classification of Signals in the Nervous System. *Electroenceph. Clin. Neurophysiology*, **6**, 479.

Uttley, A. M. (1955) The Conditional Probability of Signals in the Nervous System. *RRE. Memo No.* 1109.

Van Bergeijk, W. A. (1961). Studies with Artificial Neurons II. Analogue of the External Spiral Inervation of the Cochlea. *Kybernetik* **1**, 3, 102–107.

Verbeek, L. (1962). On Error Minimizing Neural Nets. In *Principles of Self-Organization* (Eds. H. von Foerster and G. W. Zopf), pp. 121–134.

Von Neumann, J. (1951). The General and Logical Theory of Automata. In *Cerebral Mechanisms in Behaviour* (Ed. L. A. Jeffress), Wiley, New York.

Von Neumann, J. (1952). *Probabilistic Logics*, California Institute of Technology.

Walter, W. G. (1953) *The Living Brain*, Duckworth, London.

Whitfield, J. C., and Evans, E. F. (1965). Responses of Auditory Cortical Neurons to Stimuli of Changing Frequency. *J. of Neurophysiology XXVIII*, 655–672.

Wilkes, M. V. (1965). Lists and Why They Are Useful. *Computer J.*, **7**, 278–281.

Windeknecht, T. G. (1964). A Theory of Simple Concepts with Applications. *Systems Research Center Report SRC* 53-A-64-19, Case Institute of Technology, Cleveland, Ohio.

Zemanek, H., Kretz, H., and Angyan, A. J. (1961) A Model for Physiological Functions. In *Information Theory*. (Ed. E. C. Cherry), Butterworth, London.

HEINZ VON FOERSTER,

ALFRED INSELBERG,

and PAUL WESTON

University of Illinois

Urbana, Illinois

*Memory and Inductive Inference**

INTRODUCTION

Bionics, if I understand it correctly, is the art of applying principles inherent in living organisms for the construction of systems of—hopefully—comparable sophistication. In the words of Warren McCulloch, who always finds a brilliant formulation, Bionics is to consider "Living Models for Lively Artefacts" (1).

In my opinion one of the fundamental principles that sustain life is an organism's ability to infer by induction or, as it is commonly put very loosely—to infer from a course of events in the past the course of events in the future. In this formulation the principle of inductive inference comes close to magic, but it makes one point very clear, namely, that to think of induction without memory of past events is an absurdity. Although in the last couple of years possible mechanisms of physiological memory and the logical problems associated with the principle of induction have received considerable attention by not the smallest minds amongst us (2) (3) (4) (5) (6), I still believe that on these topics the last word has as yet not been spoken. I hasten to add that you will not hear it today uttered by me.

My intentions are modest. I simply wish to raise a few points with regard to the kind of memory that is prerequisite for induction. I hope to be able to show that a memory that merely stores *propositions* leads to technological, or organic, monstrosities and frustrates, rather than

* This work is sponsored in part by the US Air Force, Systems Engineering Group, Contract AF 33 (615)-3890; the US Air Force Office of Scientific Research, Grant 7-66; and the US Department of Health, Education and Welfare, Grant GM-10718.

facilitates, inductive operations. On the other hand, I hope to make it at least plausible that a memory that provides—so to say—the proper "impedance match" for inductive inference is one that stores methods for computing *relations.*

I shall develop this thesis along four points, beginning with a brief account on our somewhat meager knowledge of an elusive mental faculty, namely, recall and recognition. Second, I shall refresh your memory about the various debates on the principle of induction with a short summary of the present state of affairs. The logical machinery necessary to express this principle I will try to describe in my third point "Logical Calculi", and finally I shall suggest some possible implementations of the required calculus by nonlinear networks.

MEMORY

In general-purpose digital computer systems two kinds of memory can be clearly distinguished by their difference in structure and in function, namely, the storage system and the program. Unfortunately, some anthropomorphically inclined quipsters have dubbed the storage system "memory", and henceforth some naive psychologists—and, alas, some neurophysiologists—who believed the engineers to know what they were talking about, kept looking for analogues of magnetic tapes, discs, cores and drums within the nervous system. This is reflected for instance in the consistent application of terms like "trace" or "engram" in learned papers on physiological memory, and the persistent search for physiological manifestations of these traces and engrams in the neural tissue of the brain.

With this I do not wish to belittle the problem of physiological memory. On the contrary, I believe this problem to be more profound than meets the eye, and this for several reasons.

First, I think, the faculties of recall and recognition are interwoven into the totality of a personality, and to contemplate these faculties independent from others is, if at all possible, a difficult job which has to be done with great caution to avoid a host of semantic traps.

Second, as Plato already knew, "memory" is met on at least three different levels which include a person's genetic make-up, his cultural background and his personal past. As you may recall, Socrates shows to Menon that his uneducated slave can be made to "remember" how

to prove the Pythagorean Theorem if only asked the right questions. This kind of memory, he points out, is a faculty that is common to all men. Moreover, the slave understood Socrates' questions, for he remembers the use of the Greek language, a faculty that is common to all Greeks. Finally, the slave may have memories of his personal experience which are common to none, known only to the salve.

Today, we still do not know where to draw the boundaries between these faculties as shown by the persistent discussions which have at their core the dichotomy "Nature versus Nurture".

I wish to mention still another profound difficulty in studies of human memory, namely, the wide gap in our knowledge of what goes on in the intermediate level between the microscopic observables on the one hand, i.e., the individual neurons and their subcellular constituents whose activity can now be studied by ingenious neurophysiological techniques or by incredibly skillful microanalyses, and, on the other hand, the macroscopical observables, i.e., the gross behavioral "outputs" of an experimental subject under controlled conditions.

Since observations in living brain tissue with microelectrodes or in extracted tissue on microscopic preparations are necessarily confined to local phenomena, it is only too understandable that irreversible changes in some of the observed functions or structures, say, changes in the efficacy of synaptic transmission or in the numbers of terminal buttons, or in the mass ratio of nucleotides, etc., which may be associated with certain experiences of the experimental subject or animal, will be interpreted as local manifestations of physiological memory.

I do not doubt for a moment that macroscopic changes in behavior induced by experience must ultimately be accounted for by changes on the microscopic level. The question I wish to pose myself is simply how these manifestations should be interpreted. Are these changes on the micro-level analogues to the local changes of, say, the magnetization in the tape of a tape recorder? Are they to be interpreted as facilitations of pathways which will, with continuous use, become well-trodden roads to centers that quickly initiate desired actions? Or are they to establish circular connections in order to hold specific pulse sequences indefinitely within this reverberating circuit?

Of course, I could continue this array of rhetorical questions, but at this point a short allegory may suggest the way in which I shall attempt to answer them.

Assume I possess two nicely packaged, small automata, called Alpha and Omega, which have some musical talents. One evening I take them along to a concert where we all listen to a magnificent performance of some of Torelli's concerti grossi. Home again, I ask Alpha—by pushing the appropriate interrogation buttons—to tell me what we heard. Miraculously, he responds to my question by playing back to me tone by tone Torelli's concerti as we heard them, even with the disturbing cough of the chap two rows behind me.

Now I turn to Omega with the same question. After some silence, interrupted by static, he tells me in a heavy German accent: "These were concerti grossi by a pre-Bach Italian composer, perhaps Corelli or Scarlatti."

For one who is uninitiated in the intricacies of constructing automata, clearly Omega missed the boat. Not only did he wrongly guess the composer—it was Torelli, of course—but he could not produce even a single tone.

In spite of these shortcomings, I believe, we would be impressed by Omega's performance if compared with that of his pedestrian brother Alpha. Now, the question arises how are we to justify our preference for Omega in the face of his apparent imperfection.

To me the crucial difference in the performance of these gadgets is that while Alpha is producing, admittedly with high fidelity, a replica of his input by mapping the input states represented in one physical modality into states of another physical modality, Omega is performing on his inputs complex operations that ultimately lead to a symbolic representation of the observed external events. Or, in other words, while Alpha remains with his one-to-one mapping operation in the same domain of representation, Omega is confronted with the task of transforming the information given in one domain of representation say, the "pressure wave function" $F(t)$, into an entirely different domain of representation, namely, the symbols of a natural language.

It seems, from this description, that a quantitative argument may be applicable. Since Alpha performs a one-to-one mapping, clearly the amount of information, H, regarding Torelli's concerti grossi must have been preserved. However, in the case of Omega it appears that H(IN) is overwhelmingly larger than H(OUT), as one can easily see by the measly 100 bits of Omega's verbal output*. One may interpret this as a reduction

* There are 14 words to approximately 7 bits per word (7).

in uncertainty (entropy) H that is accomplished by Omega. Consequently, Omega may be accredited with considerable organizational powers. However comparison of the two quantities H(IN) and H(OUT) has to be made with some caution, for these two entropies are not compatible. The set of events from which the input message is selected to give H(IN) is not the same as the set of events from which the output message is selected to give H(OUT). Of course, this observation does not diminish Omega's organizational powers, it only suggests where to look for finding the source of this power. It must, of course, reside in Omega's ability to carry out the transformation from one domain into another as mentioned before, or – to put it more forcefully – to compute on his input certain invariants – or abstracts – and to establish the appropriate relations amongst these abstracts in the given input situation.

Clearly, all this is easier said than done. But there are strong constraints in Omega's external universe, the world of music, and his internal universe, the structure of language, which come to his aid. A concerto grosso has strict rules that are substantially different from, say, an oratorio, and these rules have a name, they are called "concerto grosso". Similarly, there are rules of concatenation that control the sequence of the linguistic symbols as well as their relationships. An utterance: "These were Corelli or Scarlatti by concerti grossi" is not a false statement, it is nonsensical.

Let us assume for the moment that the constraints in Omega's external and internal universe are indeed very strong so that the difficulties of an implementation of Omega's abstraction and transformation abilities are not out of this world. I even venture to say that probably most of you have in the meantime thought of a nice solution for Omega either in form of a computer program or of a neat piece of transistorized hardware. Whatever this solution might be, it is clear that what we may now call Omega's "memory" is his stored ability to compute on his input the desired abstracts and to establish the appropriate relations amongst these abstracts in a given input situation. Since he is supposed to determine invariants in temporal sequences, it is clear that he has to "hold" a sequence of sufficient length to make, at least, a tentative decision. He needs, therefore in addition a "short term memory" or better, he has to be able to make a "short term record" which has not to be longer than to go into one of Omega's ears and to go out by the other.

With this rough knowledge of some of Omega's basic functional and anatomical units we can now appreciate his superiority over Alpha. While we may play to Omega hours and hours of music to which he calmly makes his more or less cogent remarks about what is being played, Alpha will sooner or later run out of storage space and will quit listening to whatever is played. This points to another fundamental difference between these two chaps which manifests itself when we wish to determine the capacity of their memories. While there is no difficulty whatsoever to establish quantitatively Alpha's memory capacity, or better, his "storage capacity" in "bits", "bytes", "words" or whatever units are preferred, we run into considerable difficulties to associate with Omega's stored ability to compute abstracts any quantitative concepts that are compatible with those that are rigorously applicable in Alpha's case. Even the capacity of Omega's "short term record" is almost impossible to assess, although the quantitative concepts applicable to Alpha would be well suited to this case. The difficulty here is, of course, the inaccessibility of the transfer points from record to computations which may vary from instant to instant depending upon the information needed to carry out the initial tentative decisions. If this short-time memory is tested after processing by interrogation, we are unable to decide whether the results are affected by computational limitations or by limitations of the short-term record. Of course, one may restrict the question of the capacity of this fleeting record to obvious time constants in the input apparatus of Omega. Then the answer is trivial.

With this somewhat lengthy narrative on Alpha and Omega one may ask what can be learned from this allegory to understand better John or Jim or cats and worms.

Let us assume that these were idealized Alpha-type spherical creatures with radius R, their body all brains, densly filled with small spherical cells with radius r, and their surface densly covered with a single layer of sensory cells also of radius r. With this simple geometry the ratio between the Number, N_S, of sensory cells and the number, N_B, of brain cells is

$$N_S/N_B = 3r/R.$$

The information absorbed by this creature through its sensors after a time t is:

$$H_S = t \cdot N_S/\tau.$$

Where τ is the refractory period of a sensor or, its reciprocal $1/\tau$, the number of bits processed per second per sensory cell.

The storage capacity of its brain is

$$H_B = kN_B$$

if k is the storage capacity in bits per cell.

This creature quits looking into the world when $H_B = H_S$, or after

$$t_{\max} = (\tau k R)/(3r).$$

Let me be realistic for a moment and let me use some numbers from physiology. Assume our creature to be of elephantine size with radius $R = 3$ meters, its cells $50\,\mu$ in size, that is $50 \cdot 10^{-6}$ m, their refractory period 50 ms, and $k = 1000$ bits per cell. With these numbers we have

$$t_{\max} = \frac{50 \cdot 10^3 \cdot 3}{3 \cdot 50 \cdot 10^{-6}} = 10^9 \text{ ms} = 10^6 \text{ sec} = 10 \text{ days}.$$

This means that after its short intellectual life of ten days this elephantine creature is saturated with a record of all it could sense in its environment during this time, but it cannot do anything with this information unless, of course, to its brain is attached a super-brain which figures out some regularities and structures from this incredibly large hodgepodge of data. This, clearly, is an almost impossible task and it may take so long that most probably in the meantime little, clever creatures that have learned to compute „Large Motionless Sphere = EATS" may long before have chewed up this giant—sensors, brain, super-brain and all.

In this competitive situation it seems to pay off to be a good computer rather than to be a faithful recorder. Particularly, since a good computer can operate "on line" on the most faithful record imaginable, namely, on the actual events themselves (8). But these computers are Omega-type systems.

It is not difficult to imagine that systems that can compute universals from particulars are also able to carry out something like an inverse operation of these computations, that is, given a universal, say, "home town", to decompose it into some of its symbolic constituents like "barber shop", "corner drug store", "school" and "church". If this decomposition is done with great detail it is sometimes referred to as "total recall". However, one should not forget that these recollections

are, of course, not presented in form of events, but in form of symbols representing events*.

Perhaps I have succeeded with these examples to make it at least plausible that the essential features of physiological memory are its various abilities to manipulate symbols, first inductively, by computing generalities from particular, and then deductively, to reconstruct the particular from the structure of the generalities. It seems also to be plausible that during "learning", that is when these specific abilities are developed, changes will take place in the functional or structural make-up of these systems predisposed to such changes. However, such changes, which must also manifest themselves on the microscopic level, cannot be interpreted to be "traces" or "engrams" of events in the exact sense of these words. They should be understood as changes in the structure of distributed computers that facilitate inductive operations.

In order to see more clearly the particular requirements of such structures, let me briefly recapitulate some ancient and modern views on the principle of induction.

INDUCTION

In discussing the principle of inductive inference and the logical problems that are associated with this principle we shall cross various domains of philosophical thought that range from the skeptical Wittgenstein who warned: "That the sun shall rise tomorrow is a hypothesis; that is, we do not know whether it will rise." (9), to the faithful believers in the "Laws of Nature" who predict the onset of eclipses within fractions of second and, most of the time, are right.

In order that in the forthcoming discussion on induction we shall not be bothered by an oblique reference of this modality of logical inference to "Laws of Nature" particularly to temporal aspects of these Laws regarding relations of the form Past-Future, let me first investigate the legal structure of these Laws, and second let me dispose of the concept of Time altogether, for this convenient mental construct has no place in induction.

* The eidetic seems to be an exception to this. But he is rare and his ability is far from being understood.

It is clear that in human affairs laws are established to regulate human behavior. He, who violates these laws, is punished. He is usually called the culprit. In contrast to these established relations between law, culprit and punishment, when it comes to Laws of Nature it is the lawmaker who is the culprit if something is observed that violates his laws. It is not Mercury who is punished when he does not "obey" Newton's laws of planetary motion; it is Newton who has to go to the doghouse, and it was Einstein who put him there. Now it is his turn to be sent to the doghouse, but as yet nobody can really show he was wrong.

This suggests that these laws are figments of our imagination or, as Wittgenstein puts it, "We cannot predict future events from the present. Faith in causality is superstition" (10).

These apparently outrageous propositions may lose their edge if we realize that Time itself is a doubtless useful but otherwise unnecessary, parameter in our description of the physical universe (11). I hope that my proposed method to dull an outrageous proposition is not taken to be a mere substitution of one outrageous proposition by another one.

Time, in my opinion, is only introduced to keep track of the simultaneity of events belonging to two or more spatially separated chains of events. However, it is very inconvenient to refer to the occurrence of one event by invoking the occurrence of another event which, in turn, may require reference to still another event, and so on. Consequently, it is convenient to introduce a standard sequence of events, for instance the angular displacements of some observable selected celestial body, to which simultaneity of other events may be referred. Indeed, this is the method by which the unit of "time" is established, namely, by picking a convenient amount of the mean angular displacement of the sun with respect to a stationary observer on earth. The customary choice is 15 angular seconds, also called "one temporal second." If we call this angular unit φ_0, then all so-called temporal expressions can be rewritten in terms of this angular unit which, more or less faithfully, permits one to estimate the positions of the sun coincident with the events under consideration. I believe it is the charming invention of clocks which caused the belief in the "flow of time" or in similar notions. It seems to me that the conceptual construct of Time is here confused with the reality of a sequence of events represented in a Clock.

Perhaps the absence of Time in the Laws of Nature is most clearly seen in the canonical representation of their Lagrangeian or Hamiltonian

formulation using generalized co-ordinates and momenta*). Here the Laws of Nature appear as $(n-1)$-dimensional fixed hyper-surfaces, or hyper-sculptures, which are defined by the constraining equation (for instance, constant total energy) in an n-dimensional phase space which, in turn, is defined by the n degrees of freedom of the system under consideration. The instantaneous state of the system is associated with a point on the surface of this sculpture, and past and future have no relevance. Should we be forced to change the laws because we do not find the representative point of the actual system on the surface of our sculpture, we change the form of the sculpture so as to conform again to the actual positions of the representative point.

I turn now to a formulation of the principle of induction which is devoid of any reference to time, and particularly of a reference to a relation between past and future events. I use the formulation by Katz(12) who discusses this principle in connection with generalizations. "[Generalizations] say of anything that has a certain property P_1 that it also has a certain other property P_2. They cover all things to which the property P_1 applies, and so they cover not only those which we have observed to possess P_2 but an unlimited number of unexamined things as well."

Temporal aspects enter this argument only if slightly modified. Again, according to Katz: "In an inductive inference we argue that, since all cases of P_1 in the past have been cases of P_2, we may conclude that all future cases of P_1 will likewise be cases of P_2."

If I understand it correctly David Hume was the first to draw our attention to a peculiar circularity in argument if we wish to justify this mode of inference. For if one justifies this principle by invoking its success in the past and hence to forecast its success in the future, one employs the *justificandum* as the *justificans*, and thus committing a fallacy known as *circulus vitiosus*. However, Katz (13) subjects this fallacy to a brilliant analysis and shows that it cannot be refuted on trivial grounds. Everybody interested in a fine logical argument may enjoy this treatise.

For my purpose it is sufficient to point out that the difficulty in justifying this principle stimulated many thinkers to attack the problem. There are in fact two aspects of it. The one concerns the problem of justification *per se*. In my opinion Katz indeed solved this problem as he

* Dotted quantities as, e.g., $\dot{q}$, $p, \dot{p}$, etc., refer in this argument, of course, to derivatives with respect to the angular quantity Φ_0.

properly announced in the title of his book: *The Problem of Induction and Its Solution.*

The other side of the problem is to find measures of trustworthiness of a particular generalization, or "degrees of confirmation" for the hypotheses under consideration. Carnap (14) attacked this problem with the full power of the logical machinery to his disposal, while Reichenbach (15), (16) hoped to settle it with the calculus of probability.

In the following I shall give three examples in which probability is applied in induction. I hope that these examples will illustrate again the peculiar position this principle assumes in the domain of logical inferences.

1. The Urn Paradox

Figure 1 shows $(n + 1)$ urns with labels, k, running from 0 to n $(0 \leqq k \leqq n)$. Each urn is filled with precisely n balls of which, according to its label k are white and $(n - k)$ are black. After filling, the labels are erased.

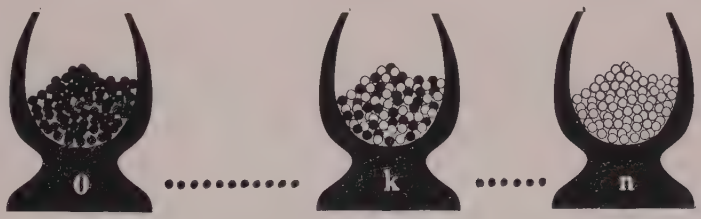


Fig. 1. $(n + 1)$ urns with k balls white and $(n - k)$ balls black.

One urn is now selected at random and m balls are drawn but not replaced. Assume they all turned out to be white. We may ask for two probabilities:

1) What is the probability for having selected the urn with white balls only:

$$\Pr(\text{white urn}) = ?$$

2) What is the probability for the next ball drawn to be white:

$$\Pr(\text{next ball white}) = ?$$

The answers are (17):

1) $\Pr(\text{white urn}) = (m + 1)/(n + 1)$
2) $\Pr(\text{next ball white}) = (m + 1)/(m + 2)$

Clearly, if the number of urns is very large ($n \to \infty$), the probability for having selected a particular one, e.g., the "white" urn, becomes vanishingly small. Hence, in the first case our intuition guides us correctly. However, in the second case when the number of balls drawn is very large ($m \gg 2$), the probability for the next ball drawn to be white approaches certainty

$$\lim_{m \to \infty} [\Pr (\text{next ball white})] = 1,$$

independent of the number of urns at disposal and, *a fortiori*, independent of our *knowledge* of how many urns are at disposal. Consequently, from consistently drawing white balls we are led to believe with overwhelming probability that we have selected an urn containing white balls only, although this probability is vanishingly small as was shown in the previous case.

So much for those who believe that consistent data indicate "truth".

2. Problematic Sunrise

An accepted way to compute probabilities is through counting frequencies. If S and F are the numbers of "success" and "failure" respectively in a specified set of events, the probability of, say, success is

$$\Pr (\text{success}) = S/(S + F).$$

We wish to challenge Wittgenstein's contention regarding our ignorance of tomorrow's sunrise. Sure enough, if the probability for tomorrow's sunrise is calculated from the number of successful and unsuccessful sunrises, certainty of success is indicated:

$$S/(S + 0) = 1.$$

However, the absence of failures leaves us with a skeptical feeling as to whether or not the probability concept is legitimately applied to this case. The interesting thing is that we would feel much more comfortable if, on one or the other occasion, the sun did not rise on the next morning.

3. Extraterrestrial People

The probable error (P.E.) in a statistical sample is associated with n, the size of the sample, by the following relation:

$$\text{P.E.} = 0.6745/\sqrt{n - 1}.$$

The sample consisting of cases in which life is found on celestial bodies consists of precisely one case ($n = 1$). Hence, statements about life on other planets are associated with an infinite probable error, in other words, nothing can be said about extraterrestrial life and, *a fortiori*, nothing (of a higher order) can be said about extraterrestrial people.

Nevertheless, considerable efforts are made to establish "communications" with these "people" who are not even "hypothetical" or, to be more precise, whose existence is supported by a hypothesis with zero degree of confidence.

I picked this last example to show that inductive inference is connected with our actions not in an obvious or straightforward way. The domain of the principle of induction is a purely logical one, and the implications derived in particular cases may or may not initiate actions. These depend on the outcome of at least two kinds of identifiable operations that are sandwiched between implication and action.

The first kind involves selective operations on a set of implications which are now associated with hedonistic values or "payoffs". The theory of games (18) (19) is concerned with the logic of this situation.

Second, when a decision is ultimately made it becomes a command. The logical structure of the relations that exist between a command and its execution has lately been given a superb treatment by Reschler (20).

Although these topics hold great fascination, they are beyond the purpose of my presentation. I shall address myself now to the logical structure of the inductive inference scheme and, first of all, to the question of the appropriate logical calculus which permits adequate representation of this scheme.

LOGICAL CALCULI

According to Hilbert (21) several levels of logical calculi may be considered which correspond to levels of complexity in the logical structure of the systems under consideration. He distinguishes, in ascending levels of complexity, (i) the calculus of propositions, (ii) the predicate calculus, and (iii) the calculus of relations. Some other authors prefer to call the calculus of relations calculus of predicates of n-th order, which includes Hilbert's predicate calculus as the calculus of predicates of first order.

Let me recapitulate briefly some of the ground rules of these calculi, just enough to express the principle of inductive inference in one of its notations.

1. Calculus of Propositions

Propositions, as for instance "all men are mortal", "bionics is a science", "snow is black" etc., are lumped into a package denoted by A, B, C, etc. and connected by logical particles (& = (and)), (v = (inclusive or)), (→ = (implication)) etc. to form logical functions

$$A \,\&\, B; \quad A \mathrm{v} C; \quad C \rightarrow A; \quad \text{etc.},$$

whose truth values are dependent upon the truth values of their constituent propositions.

For instance if the proposition "all men are mortal" is represented by A and the proposition "bionics is a science" is represented by B, then the connection

$$A \rightarrow B$$

appears to be true, for an implication is only false if the implicatum is false, but I believe we all agree that the proposition B is true.

This calculus is very limited indeed. This can be sensed when we realize that an atomic proposition as e.g. "all men are mortal" has in fact a complex finestructure. Let me consider for a moment several levels of coarseness.

(1) "All men are mortal" is taken as an elementary totality being either true or else false. This interpretation belongs to the calculus of propositions.

(2) "All men are mortal" may be taken as a definition. For instance, if I use Aristotle's unflattering circumscription for men and gods as being "featherless bipeds", then I may use the above definition to separate men from gods: "featherless bipeds that are immortal are gods."

(3) "All men are mortal" may be taken to represent a hypothesis with a certain degree of confidence. In this particular case the truth of this hypothesis cannot yet be asserted. Moreover this hypothesis stands on pretty weak feet if we use a probabilistic argument. The total number of men ever to have seen the light of this world is estimated to be 70 billions (22), 3 billions of which are alive today. Consequently for approximately 5 out of one hundred the mortality hypothesis is not confirmed.

In order to accomodate such possibilities of interpretation in proper notation the calculus of proposition has to be extended.

2. Predicate Calculus

One considers all things to have or to have not certain predicates Pr. For instance

Me ... "all things are men"
Mo ... "all things are mortal".

The proposition "all men are mortal" which earlier was an unstructured elementary entity symbolized by "A" reveals in this notation more subtelties:

$\overline{\text{Me}}$ v Mo ... "all things are not men or are mortal"*

i.e., to use another example "some men are immortal":

$\overline{\overline{\text{Me}}}$ v Mo ... "it is false that all things are not men or are mortal".

With this extension of the propositional calculus it is now possible to accommodate all Aristotelian syllogisms; however, this extended calculus is still too weak to allow for a workable representation of even the simplest mathematical or other logical relations between two or more variables. This, however, is accomplished in the

3. Calculus of Relations

Here one considers predicates Pr $(x_1, x_2, \ldots x_n)$ of an arbitrary number of variables $x_1, x_2, \ldots x_n$ and confines the applicability of these predicates to all or to some values of these variables x_i by so called "quantifiers" (x) or (Ex) which are prefixed to the predicate. For instance

(x) Pr(x)... "all x have predicate Pr"

(Ex) Pr(x) ... "there is at least one x for which Pr is the case."

Let me illustrate the power of this notation on an example using dyadic relations, i.e., predicates with two variables, or predicates of second order. I define the dyadic relation expressing equality of two variables x and y by the following notation:

$= (x, y)$ ... "x equals y"

* A bar on top of a predicate indicates negation of this predicate.

where the equal sign " = " replaces the general "Pr". I define another dyadic relation expressing immediate succession of one variable by another by the notation:

$$S(x, y) \ldots \quad \text{"the immediate successor of } x \text{ is } y\text{."}$$

One may now express for instance one of Peano's axioms

"each number has one and only one successor"

in the following manner:

$$(x)\,(Ey)\,\{S(x, y)\ \&\ (z)\,[S(x, z) \rightarrow =(y, z)]\}$$

This can be read as follows:

> "for each x there exists at least one y which is immediate successor of x and which, if z is immediate successor of x, is equal to all such z."

We are now prepared to put the principle of inductive inference into this formalism. Let $P_1(x)$ and $P_2(x)$ represent two properties of things x, and $\neq (x, y)$ the dyadic relation expressing inequality of things x and y. Generalization of the concomitance of P_1 and P_2 is achieved if we conclude from some cases of concomitance to all cases:

$$(x)\,(Ey)\,\{P_1(x)\ \&\ P_2(x) \rightarrow [\neq (x, y) \rightarrow P_1(y)]\} \rightarrow (y)[P_1(y)\ \&\ P_2(y)]\}$$

Or in words "if for all x for which P_1 and P_2 are concomitant there exists some y, different from x, to which P_1 applies then P_1 and P_2 are universally concomitant."

Clearly, this statement is not universally true, hence it must say something about the world and, thus, can be found to be true or else false. This, however, is the fate of all hypotheses. The crux of it, of course, is the inference from particularity (Ey) to universality (y) regarding concomitance of the properties P_1 and P_2. The variables x and y bind these two properties together, consequently the whole expression establishes a relation between these properties:

$$R(P_1, P_2).$$

The important point to note here is that the principle of inductive inference postulates a relation in which predicates appear as arguments rather than propositions. Since these predicates may themselves be predicates of higher order, i.e. n-th order relations, induction may lead to the establishment of relations.

After this brief *tour de force* into logical calculi let me back track for a moment to the point where we discussed Omega's way of remembering the past by first quickly computing generalities from particulars and then to "hold" these computational operations for the purpose of reconstructing the particulars by utilizing their mutual relationships. I proposed to use this paradigm for at least one form of physiological memory, namely, its "long range" manifestation.

We see now that at least the generalizing portions of memory are essentially inductions over and over again, and since induction establishes relations, the possibility of reconstructing particulars by utilizing their mutual relationships may arise as a natural consequence of the basic functions of such systems.

The question now arises does physiology permit such possibilities or are these speculations far out of the ballpark. It appears to me that with the notions of abstracting networks as they have been developed by experimental neurophysiologists as well as by engineers and mathematicians, the bridgeheads have been established to link eventually their notions with the notions of a memory that functions as an adaptive inductive inference computer.

In the following I shall attempt to sketch these ideas.

NETWORKS

He who talks to an interdisciplinary audience has to watch his terms. Particularly if these terms are common to all disciplines, but have different meaning in each of them. As long as a specialist works in his own corner, differences in meaning go unnoticed and are harmless. Tension mounts when it is discovered that the same term is used by the man across the fence.

In spite of the integrating action of our meetings there exists an undeniable antagonism between neurophysiologists and engineers when they talk about networks, for both contemplate such structures for apparently similar purposes, but with a decidedly different approach. The neurophysiologist, occupied with analysis, approaches his job with scepticism and caution, while the engineer, occupied with synthesis, approaches his job with naivité and confidence. When he comes up with a sophisticated toy that performs some of the functions of living organisms,

and shouts: "Heureka, I got your system explained!" the physiologist shakes sadly his head and points to the numerous other functions which are not performed by the imitating gadgets.

It seems to me that "mutual fertilization" of these sciences is not yet quite at its heights. Not because of sterile gametes in these potential partners; it is because the copulatory mechanisms are still in a prepubescent state.

What is known about networks today, and how can this knowledge be formulated so that physiologists and engineers may not raise objections? I believe both will agree that, in a general sense, networks are collections of arbitrarily branching, unidirectional transmission lines which interact with each other at certain discrete, localized regions in space.

I believe also that to call that which is transmitted over these lines a "Signal" and to call that which takes place at these localized regions "Information Processing" has become common practice in both camps. The regions at which information is processed are reduced by engineers to functional boxes with inputs and outputs. and by mathematicians to the network's nodal elements of either finite or infinitesimal extensions are bestowed with mysterious faculties of manipulating spatio-temporally distributed variables.

Neurophysiologists find it more difficult to identify these localized regions of interaction, for they find them crammed with structural and functional details as, e.g., the neuron's perikaryon with all its dendritic ramifications, and with all the interstices and interfaces that are produced by the transition or termination of axons arising from other neurons.

Even more difficulties are encountered if effects of interaction are to be specified (23). Here the situation is so complex, the possible responses so numerous, and the measurements so difficult that, in spite of the tremendous knowledge gained in the last decade, neurons are still very enigmatic.

Nevertheless, concepts as, e.g., „facilitation", inhibition", "temporal integration", etc., appear to be generally applicable to these regions.

Consequently, it was hailed as a major triumph when about seven years ago in spite of all these difficulties the operational modalities of some neural networks close to the sensory layer were established (24). It appears that these nets process information in parallel by performing

distributed operations on their inputs so as to extract certain invariants from a given set of stimulus configurations.

The theoretical foundations for these abstracting operations were laid down about twenty years ago in great generality by McCulloch and Pitts (25) (26), and more specifically later by others (27), (28), (29). It can be safely said today there are few spatio-temporal abstracts for which—at least in principle—corresponding abstracting networks cannot be established.

However, these findings leave the neurophysiologists cold and, I believe, for some good reasons. They argue that the connectivities employed in these theories or the transfer function suggested for the elements have little, if anything at all, in common with the physiological entities they claim to describe. Their complexities are insufficiently represented in the naive theoretical models.

I have given some thought to this apparent schism between neurophysiologists and model makers, theoretical or practical, and I think that this schism may eventually be resolved if theoreticians succeed in finding just the right level of generalizations. They should not be so general as to say almost nothing, and not so specific as to say too much. Physiologists, I trust, will in time shed their reservations about the motives of their theoretical and engineering colleagues. I am convinced that they are not out to trivialize the magnificent organization of living things, and their attempts to comprehend these organisms are very honest indeed.

Since I am supposed to be a theoretician, I will follow the recipe which I suggested before. In the remaining portion of this presentation I shall briefly touch upon a few recent developments which, hopefully, aim at the right level of generalization. However, at this early stage of development there is no more than the sketch of the skeleton of an idea.

I shall now list the various premises and postulates regarding memory in nonlinear networks and I shall take them up point by point, developing them in more or less detail as the necessity may arise.

PRELIMINARY LIST

(1) There are parallel networks without loops that are known to compute invariants (abstracts) in a set of stimulus configurations.

(2) Cascades of such parallel networks are known to compute abstracts of ascending order of generality.
(3) It is known that nets with loops, once excited, may continue their activity around these loops and, under certain conditions, may maintain reference to the past for an indefinite period of time.
(4) It is known that nets with loops are equivalently represented by indefinite temporal cascades of virtual parallel networks without loops.
(5) From points (2) and (4) it is clear that networks with loops may in time compute abstracts of ascending order of generality.
(6) It is known that linear networks with loops, once excited, will continue their activity around these loops, however, eventually without reference to the past whatsoever, save for the fact of having once been excited.
(7) No such decay occurs, if linear superpositions are replaced by superpositions defined in nonlinear algebraic systems.
(8) Reference to algebra only, without specification of function, suffices to show the possible decomposition of present activity into representations of elements belonging to the remote past.

DETAILED ACCOUNT

I shall now discuss these eight points in some greater detail.

1, 2. Abstracting Networks

I would rather carry coal to Newcastle than to elaborate on abstracting nets before this audience. This topic is before us now for more than twenty years, its bibliography is overwhelming and at this very symposium there are at least a dozen papers that will add to the knowledge in this subject. Hence, taking your permission for granted, I shall proceed to the next point.

3. Nets with Loops

Nets with loops were first seriously considered by McCulloch and Pitts (25). In part III of their paper, they clearly established the significance of loops in nets with regard to continued, reverberating activity around these loops with possible "reference to past events of an indefinite degree of remoteness."

In networks with loops one has to distinguish between maintaining activity *per se* and maintaining reference to earlier states. Maintenance of activity *per se* in such nets is secured if they are conservative or if outside sources replenish possible internal energy losses. Clearly, these conditions are fulfilled in neural nets, and are feasible for all artefacts if so required. However, maintenance of references to earlier states within a network with loops is not a natural consequence of the loops contained in the network. In addition, it is necessary that the nodal elements of the net process their afferent signals by some nonlinear superposition rule. Pitts and McCulloch could forego this observation, for their nodal elements compute logical functions which are *ipso facto* highly nonlinear. The necessary *and* sufficient conditions regarding the nonlinear superpositions which will safeguard references to earlier states in the network are not known to me. The insufficiency of linear superpositions in this respect I shall discuss in point (6).

4. Unloop the Loop

Let any network of n nodes (elements) be represented by a directed graph, i.e., connections between nodes are oriented by "arrows" leading

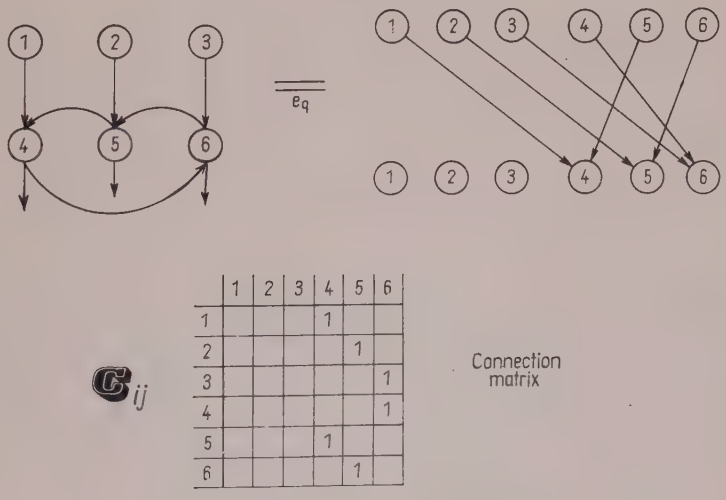


Fig. 2

Fig. 2. Equivalence of a net with loops (interaction net) with a cascaded net without loops (action net).

from a "source" to a "target". Consequently, an arbitrary element ε_i may be both, source for ε_k and target for ε_j. We denote such an oriented connection that leads from ε_i to ε_j by $c_{ij} = 1$; but $c_{ij} = 0$ if there is none. Clearly, the c_{ij}'s represent the $n \times n$ entries of a square matrix C which determines uniquely the connection structure of this network (Fig. 2). However, by interpreting the n rows and the n columns of this matrix as representing n sources and n targets respectively, and the entries c_{ij} as indicating presence or absence of connection, the same matrix C represents uniquely a network of $2n$ elements, n of which are sources only, and n of which are targets only. This network has no loops and signals flow in one direction only ("parallel" or "action" network).

Let element ε_j compute the function $F_j[x_1, x_2, \ldots x \ldots x_n]$ on its inputs $(x_1, x_2, \ldots x \ldots x_n) = X$ and let this computation be of finite duration τ. Then, in both nets, after one period of τ the output of element ε will be at time $(t + \tau)$:

$$y_j(t + \tau) = F_j[X(t)].$$

Since in the first net any element ε_j may be the source of other elements, the output $Y(t + \tau)$ of this net becomes input to elements of the second net:

$$X(t + \tau) = Y(t + \tau).$$

This step is accomodated in the second net by repeating it in cascade (Fig. 3), and again in both nets, after two periods of τ, the output of element ε_k will be

$$\begin{aligned} y_k(t + 2\tau) &= F_k[X(t + \tau)] \\ &= F_k[Y(t + \tau)] \\ &= F_k[F_j[X(t)], \end{aligned}$$

and by induction, using subscripted subscripts i_λ to denote steps of time or levels of cascade, in both nets after m periods of τ the output of element ε_i will be

$$Y_{i_m}(t + m\tau) = F_{i_m}[F_{i_{m-1}}[F_{i_{m-2}}[\ldots F_{i_1}[X(t)]]]]$$

Consequently, a cascade of m virtual parallel networks is functionally equivalent to a network with loops after m cycles of computation (29).

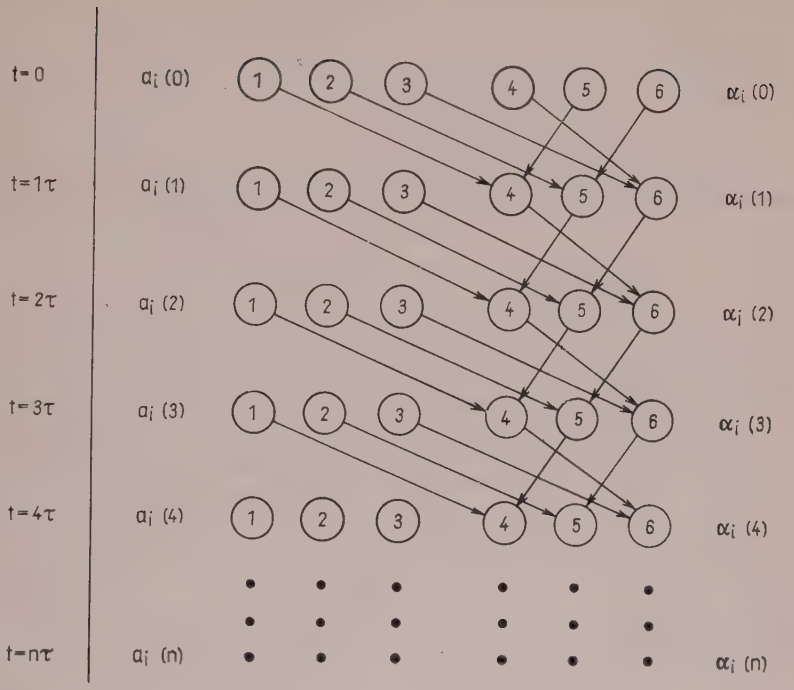


Fig. 3

Fig. 3. Temporal cascades of virtual nets without loops representing reverberations of a net with loops. a_i $(i = 1, 2, 3)$ and α_i $(i = 4, 5, 6)$ represent the activities x_i of the elements in the nets of Fig. 2.

5. Iteration of Process

It appears to be clear from the foregoing that iteration of process, whether by cascades or by reverberations, will enhance, maintain or destroy some original features of X. This, of course, depends on the process $F = \{F_i\}$ which is defined by the net's structure and function.

6. Additive Compositions

I shall now give specific meanings to the functions F_i by postulating that each element ε_{i_λ} weights its input from $\varepsilon_{i_{\lambda-1}}$ according to $v_{i_{\lambda-1}, i_\lambda}$

and then performs algebraic summation on the weighted inputs. Moreover, the net is conservative. Thus:

$$F_{i_\lambda}(X) = \sum_{i_{\lambda-1}} w_{i_{\lambda-1}, i_\lambda} \cdot x_{i_{\lambda-1}}$$

with

$$\sum_i x_i = \overline{X}$$

and

$$\sum_{i_\lambda} w_{i_{\lambda-1}, i_\lambda} = \sum_{i_\lambda} v_{i_{\lambda-1}, i_\lambda} \cdot c_{i_{\lambda-1}, i_\lambda} = 1$$

Consequently, according to the paradigm given in point (4), the output of element ε_{i_m} after m iterations

$$y_{i_m}(m\tau) = \sum_{i_{m-1}} w_{i_{m-1}, i_m} \sum_{i_{m-2}} w_{i_{m-2}, i_{m-1}} \sum \cdots \sum_{i_0} w_{i_0, i_1} x_{i_0}(0)$$

But, because of distributivity of factors w_{ij}, we simply have

$$y_j(m\tau) = \sum_i w_{ij}^{(m)} x_i(0),$$

where $w_{ij}^{(m)}$ are the entries of the iterated stochastic matrix W . If we let $m \to \infty$, this matrix degenerates to the row vector w_j and

$$y_j(\infty) = w_j \sum_i x_i(0) = w_j \overline{X}.$$

In other words, all traces of the original stimulus configuration $x_i(0)$ are gone, the network is in a state which is solely determined by properties of the network w_j, and by the fact that a stimulus $\overline{X}$ was applied.

7, 8. Non-Additive Superpositions

The principal idea of non-additive superpositions and their algebraic properties has been worked out by my colleague Dr. Alfred Inselberg (30) who kindly provided me with an appendix to this paper in order to complement the brief narrative that limited space and time will permit me. As you will see in a moment, the charm of these superpositions lies in the fact that the hopelessly unwieldy expression of iterated functions $F_i[F_j\,[F_k \ldots]]]$ for the output of the net after a large number of cascades can be brought into tangible forms without reference to particular functions by generalizing the algebra of superpositions only. It appears to

me that such a calculus may be desired in a situation where the computation of algebraic sums or the computation of logical functions are too strong assumptions for an appropriate description of superpositions of neural events.

The crucial point in these superpositions is to consider compositions that also have separations. Let me illustrate this on binary operations which shall be called $*^1$ and $*^2$.

Consider an operator T such that

$$T(x*^1y) = T(x)*^2 T(y).$$

The left-hand side represents the composition, the right-hand side separation for the operator T, and the pair $*^1$, $*^2$ is a superposition. If the binary operations on both sides are alike, say, $*^1 = *^2 = *$, the corresponding operator T will be called "uniform", otherwise, as in the case above, "non-uniform".

Furthermore let us introduce the corresponding scalar operators ∇^1 and ∇^2 such that

$$T(a\nabla^1 x) = a\nabla^2 T(x)$$

for the non-uniform case and for uniform operators simply $\nabla^1 = \nabla^2 = \nabla$.

The interpretation of these binary and scalar operations and the corresponding operators T in "black box" language is to associate with the mutual effect that two input signals may have on each other a binary composition $*^1$. The outcome of this composition $(x *^1 y)$ is subjected to the operator T which produces an output that, in turn, has the structure of a binary composition $T(x) *^2 T(y)$.

A different, but equivalent, notation may further illuminate this point. Let F and G be two functions corresponding to $*^1$ and $*^2$ respectively, then the composition-separation postulate reads:

$$T[F(x, y)] = G[T(x), T(y)]$$

The scalar operations ∇^1, ∇^2 and ∇ play in this calculus the same role as scalar multiplication does in the case of additive superpositions. This permits us to accommodate connections between elements or gains at each element in the case that they are not absorbed in the operator.

The application of this machinery to the input-output relationship of an arbitrary element ε_j that is located in a layer λ and receives its inputs

from all, or part, of the elements ε_i in the preceding layer $(\lambda - 1)$ will make it more transparent.

First, we have to realize that by picking a particular element ε_j, we have to associate with this element not only the appropriate operator T_j, but also the corresponding binary operations $*_j^1$ and scalar operations ∇_j^1. Since we wish to account for all n elements of the previous layer we introduce as before (point (6)) a weighted connection matrix w_{ij} and a symbol (read C for composition)

$$\mathop{\mathbf{C}}_{i=1}^{n} *_j^1$$

which corresponds to Σ for additive composition and is a shorthand notation for a string of variables composed by the operation $*_j^1$. The subscript over which this composition has to be taken will be indicated in the usual form by placing it below this symbol. Here we wish to compose all afferent signals originating from elements ε_i, hence i is the running index going from $1 \to n$.

With these preliminaries we are now in a position to write the output for ε_j:

$$y_j(t + \tau) = T_j \left[\mathop{\mathbf{C}}_{i=1}^{n} *_j^1 w_{ij} \nabla_j^1 x_i(t)\right]$$

But with the beautiful properties of this algebra we can slip the operator T_j under the composition C, realizing, of course, that all operations flip from $*^1$, ∇^1 to $*^2$, ∇^2:

$$y_j(t + \tau) = \mathop{\mathbf{C}}_{i=1}^{n} *_j^2 w_{ij} \nabla_j^2 T_j[x_i(t)].$$

It appears from this expression as if cascading were immediately possible by recursion. However, a difficulty may arise from collision of operators T_i and T_j referring to elements in different layers and having unspecified composition rules for $T_i T_j$. A compatibility criterion, given in the appendix eliminates this difficulty and the output of the element ε_{i_m} after m iterations is

$$y_{i_m}(m\tau) = \mathop{\mathbf{C}}_{i_0=1}^{n} *_{i_m}^2 \mathop{\mathbf{C}}_{i_1=1}^{n} *_{i_m}^2 \cdots \mathop{\mathbf{C}}_{i_{m-1}=1}^{n} *_{i_m}^2 \left[\mathop{\Phi}_{\lambda=1}^{m} w_{i_{\lambda-1} i_\lambda} \nabla_{i_m}^2 T_{i_m} \cdots T_{i_1}\left(x_{i_o}(0)\right)\right]$$

where $\mathop{\Phi}_{\lambda=1}^{m} w_{i_{\lambda-1} i_\lambda}$ represents the mth application of the connection-gain matrix w_{ij} and is a scalar. This corresponds to the mth power of

the matrix $\|w_{ij}\|$ for the appropriate multiplication rules as shown in the Appendix.

Perhaps, at first glance, this expression appears to be a greater mess than anything we had encountered before. Hopefully, this impression disappears at second glance, for the real beauty of it is that the different roles played by the structure of the network as given through its connections, and its function as given by the collection T_i of operators associated with the elements, is now neatly separated. The structured part is represented by

$$\overset{m}{\underset{\lambda=1}{\Phi}} w_{i_{\lambda-1}i_\lambda}$$

while the functional part is lumped together into the iteration of operators

$$T_iT_jT_k \ldots$$

We are presently studying the effects of various choices of T_i on the modification of the original signal configuration $x_i(0)$ when processed in successive cascades. I wish I could show you some computed results. However, the work takes not weeks as I had hoped, but months. I can only ask for your understanding and patience to wait for the next installment.

Appendix

NONLINEAR CASCADED NETWORKS

1. Uniform Operators

Consider a cascaded network without loops, (Fig. 3), each layer of the cascade being functionally identical to all other layers. The input-output relation for a given layer is represented by an operator T.

At first we consider operators which are in a class C_F [Ref. (30), p. 10], that is if T is an operator from a set D into itself ($T: D \to D$) there exists a function F such that

$$F: DXD \to D$$

and

$$T[F(x, y)] = F(T(x), T(y)). \tag{1}$$

Alternatively letting:

$$x*y = F(x, y)$$

relation (1) can be written as:

$$T(x*y) = T(x)*T(y). \tag{2}$$

Such operators are called uniform.

To gain a feeling for the variety of such operators some examples are provided:

Example 1. T a linear operator that is $T: D \to D$ and there exists an operation $+$ on D (with $+$ having the usual properties) such that

$$T(x + y) = T(x) + T(y) \tag{3}$$

Also, D is a linear space over some field Γ with "$\cdot$" denoting the scalar multiplication. T being such that:

$$T(\alpha x) = a \cdot T(x), \ a\varepsilon\Gamma, \ x\varepsilon D \tag{4}$$

Example 2. T a star-linear operator. That is $T: D \to D$ and there exists a binary operation $*$, $(* \neq +)$, D such that:

$$T(x*y) = T(x) *T(y).$$

The set D with the operation $*$ is an abelian group.

Furthermore, D is a star space [Ref. (3), p. 11–15] over some field Γ with "∇" denoting the scalar operation and T such that:

$$T(a\nabla x) = a\nabla T(x), \ a\varepsilon\Gamma, \ x\varepsilon D \tag{5}$$

In particular if there exists an $f: D \to D$ with f^{-1} and

$$x*y = f^{-1}(f(x) + f(y)) \tag{6}$$

then

$$a\nabla x = f^{-1}(af(x)) \tag{7}$$

For specific examples consider

(i) $f(x) = \log x$, then

$$x*y = e^{\log x + \log y} = xy$$

$$a\nabla x = e^{a \log x} = x^a$$

A star-linear operator for such a star would be any operator T of the form:

$$T(u) = e^{L(\log u)} \tag{8}$$

where L is a linear operator.

(ii) $f(x) = e^x$
$x*y = \log(e^x + e^y)$
$a\nabla x = \log(ae^x) = \log a + x$
$T(u) = \log L(e^u)$.

(iii) $f(x) = x^c$, $f^{-1}(x) =$ the principal value of $x^{1/c}$.
$x*y = (x^c + y^c)^{1/c}$
$a\nabla x = (ax^c)^{1/c} = a^{1/c}x$
$T(u) = [L(x^c)]^{1/c}$ (principal value).

Example 3. Generalized translations of star-linear operators. [Ref. (30), p. 33–41]

(i) Let $T(u) = L(u) + v \quad T: D \to D$

where L is a linear operator from D into D and u is a fixed element in the range of L.

$$x*'y = \frac{x + y}{2}$$

It is possible to define a scalar operation, "∇'", for such an operator but it needs to be defined on a *pair* of scalars and a *pair* of elements of D, i.e.:

$$(a, b)\,\nabla'(x, y) = \frac{ax + by}{a + b}, \quad a + b \neq 0$$

We have then, that

$$T((a, b)\,\nabla'(x, y)) = (a, b)\,\nabla'(T(x), T(y))$$

(ii) A generalized translation of a star-linear operator S.

If S is a star-linear operator (as in example 2) with the, "$*$", represented by:

$$x*y = f^{-1}(f(x) + f(y)),$$

define

$$T(u) = f^{-1}(f(S(u)) + v) \tag{9}$$

for u a fixed element in the range of S. Then,

$$T(x*'y) = T(x) *' T(y) \tag{10}$$

for

$$x*'y = f^{-1}\left(\frac{f(x) + f(y)}{2}\right) \tag{11}$$

Furthermore, if

$$(a, b)\,\nabla'(x, y) = f^{-1}\left(\frac{af(x) + bf(y)}{a + b}\right), \quad a + b \neq 0 \tag{12}$$

then

$$T((a, b)\,\nabla'(x, y)) = (a, b)\,\nabla'(T(x), T(y)) \tag{13}$$

The space $(D, *', \nabla')$ that is the set D with the binary operation $*'$ and a scalar operation ∇' is called a star-extension space. It does not have the properties of a linear space and a operator T defined by (9) is not a linear operator in *any sense*. In particular the properties of $*'$ are:

(1) associativity

$$x*'(y*'z) = (x*'y) *'z \tag{14}$$

(2) commutativity

$$x*'x = y*'x \tag{15}$$

(3) idempotent property

$$x*'x = x \tag{16}$$

The properties of ∇' are:

(1) scalar commutativity:

$$(a, b)\,\nabla'(x, y) = (b, a)\,\nabla'(x, y) \tag{17}$$

(2) $*'$-distributivity:

$$\begin{aligned}&(a, b)\,\nabla'(x_1*'x_2, y_1*'y_2) = \\ &= ((a, b)\,\nabla'(x_1, y_1)) *'((a, b)\,\nabla'(x_2, y_2))\end{aligned} \tag{18}$$

(3) ∇'-"distributivity":

$$\begin{aligned}&(c, d)\,\nabla'\{[(a_1, b_1)\,\nabla'(x, y)], [(a_2, b_2)\,\nabla'(x, y)]\} = \\ &= (ca_1(a_2 + b_2) + da_2(a_1 + b_1), cb_1(a_2 + b_2) + db_2(a_1 + b_1))\\ &\nabla'(x, y)\end{aligned} \tag{19}$$

(4) scalar multiplicative identity:

$$(l, l)\,\nabla'(x, y) = x*'y, \tag{20}$$

where l is the multiplicative identity of the scalar field Γ

(5) scalar additive identity:

$$(0, 0)\,\nabla'(x, y) = e, \tag{21}$$

where 0 is the additive identity of Γ and e is the $*$ (not $*'$) identity in D.

Example 3 (i) is a special case of such a situation. Another example is: S a multiplicative operator [Example 2 (i)]. Let

$$T(u) = v \cdot S(u)$$

with v a fixed element in the range of S. Then

$$T(x *' y) = T(x) *' T(y)$$

with

$$x *' y = \sqrt{xy} \qquad \text{and}$$

$$T[(a, b)\,\nabla'(x, y)] = (a, b)\,\nabla'(T(x)T(y))$$

with

$$(a, b)\,\nabla'(x, y) = (x^a y^b)^{\frac{1}{a+b}}; \quad a + b \neq 0$$

With the provision of these examples we are now in a position to discuss the cascade shown in Figure 3. Let us assume that T (the input-output relation) is such that there exists a binary operation $*$, on the domain and the range of definition of T, with

$$T(x * y) = T(x) * T(y) \tag{2}$$

To preserve generality we require only that $*$ be commutative [Equation (15)] and associative [Equation (14)]. These properties are shared by the examples (1, 2, and 3).

Furthermore it is assumed that there exists a scalar operation ∇, defined on a single (or pair) element of the field Γ and a single (or pair) element of the domain D^* and such that:

$$T(a \nabla x) = a \nabla T(x)$$

or

$$T[(a, b)\,\nabla(x, y)] = (a, b)\,\nabla(T(x), T(y)]$$

for a pairwise ∇.

* The words, or pair, in parenthesis are inserted to preserve generality and so that generalized translations (Example 3) may also be considered. In case a pairwise definition of ∇ is required the restriction imposed on the elements of the cascade is that each element receives stimulation from at least two other elements or no elements at all.

The only properties of ∇ which will be employed here are scalar commutativity:

$$a\nabla(b\nabla x) = b\nabla(a\nabla x) \tag{22}$$

[or Equation (17)] and $*$-distributivity:

$$a\nabla(x*y) = (a\nabla x)*(a\nabla y) \tag{23}$$

[or Equation (18)]. Also it is required that there exist a function Φ, $\Phi : \Gamma \times \Gamma \to \Gamma$ such that:

$$a\nabla(b\nabla x) = \Phi(a, b)\,\nabla x \tag{24}$$

The notation

$$\Phi(a, b) = a \cdot b \tag{25}$$

will be alternatively employed.

Likewise, for a pairwise ∇, a $\Phi : \Gamma^6 \to \Gamma$ such that*:

$$\begin{aligned}&(c, d)\,\nabla\{[(a_1, b_1)\,\nabla(x, y)],\ [(a_2, b_2)\,\nabla(x, y]\} = \\ &= (\Phi(c, d, a_1, a_2, b_1, b_2), \Phi(c, d, b_1, b_2, a_1, a_2))\,\nabla(x, y)\end{aligned} \tag{26}$$

[See Ref. (30), p. 38]

It should be pointed out that the properties of $*$ and ∇, which are employed *do not imply linearity of T.*

Consider now the cascade having m layers each layer having n elements. The operator T is represented by an n-tuple:

$$T = \{T_1, T_2 \ldots, T_n\} \tag{27}$$

where T_i, an operator, represents the input-output relation of the *ith element* of a layer.

Define $\bigcirc\!\!\!\!*$ by:

$$\bigcirc\!\!\!\!*_{i=1}^{n} x_i = x_1 * x_2 * \ldots * x_n \tag{28}$$

then the output, $y_{i2}(2\tau)$ of the ith element in the *second* layer is given by:

$$T_{i_2}\left[\bigcirc\!\!\!\!*_{i_1=1} w_{i_1 i_2} \nabla y_{i_1}(\tau)\right] = y_{i_2}(2\tau) \tag{29}$$

where $w_{i_1 i_2}$ denote the gain and connectivity of the i_2th element in

* The notation Γ^2 means $\Gamma \times \Gamma$, the Cartesian product of Γ taken twice. Similarly for Γ^6.

the second layer from the i_1th element in the first layer and $y_{i_1}(\tau)$ denotes the response of the i_1 element in the first layer*.

Applying conditions (2) and (5) for (10) and (13) for pairs inductively to (29) one obtains:

$$y_{i_2}(2\tau) = \overset{n}{\underset{i_1=1}{\mathbf{C}^*}} w_{i_1 i_2} \nabla T_{i_2}(y_{i_1}(\tau)) \tag{30}$$

The response, $y_{i_3}(3\tau)$, of the i_3th element of the third layer is likewise given by:

$$y_{i_3}(3\tau) = T_{i_3}\left[\overset{n}{\underset{i_2=1}{\mathbf{C}^*}} w_{i_2 i_3} \nabla y_{i_2}(2\tau)\right] = \overset{n}{\underset{i_2=1}{\mathbf{C}^*}} w_{i_2 i_3} \nabla T_{i_3}(y_{i_2}(2\tau)) =$$

$$= \overset{n}{\underset{i_2=1}{\mathbf{C}^*}} w_{i_2 i_3} \nabla T_{i_3}\left[\overset{n}{\underset{i_1=1}{\mathbf{C}^*}} w_{i_1 i_2} \nabla T_{i_2}(y_{i_1}(\tau))\right]$$

$$= \overset{n}{\underset{i_1=1}{\mathbf{C}^*}} \overset{n}{\underset{i_2=1}{\mathbf{C}^*}} w_{i_1 i_2} \cdot w_{i_2 i_3} \nabla T_{i_3} T_{i_2}(y_{i_1}(\tau)) \tag{31}$$

Iterating $(m-1)$ times the output, $y_{i_m}(m\tau)$, of the i_mth element of the mth layer is given by:

$$y_{i_m}(m\tau) = \overset{n}{\underset{i_1=1}{\mathbf{C}^*}} \overset{n}{\underset{i_2=1}{\mathbf{C}^*}} \cdots \overset{n}{\underset{i_{m-1}=1}{\mathbf{C}^*}} \left[\overset{m-1}{\underset{\lambda=1}{\Phi}} w_{i_\lambda i_{\lambda+1}} \nabla T_{i_m} T_{i_{m-1}} \cdots T_{i_2}(y_{i_1}(\tau))\right] \tag{32}$$

where

$$\overset{m}{\underset{\lambda=1}{\Phi}} w_{i_\lambda i_{\lambda+1}} = w_{i_1 i_2} \cdot w_{i_2 i_3} \cdot w_{i_3 i_4} \cdot \ldots \cdot w_{i_m} w_{i_{m+1}} \tag{33}$$

[see Equation (25)].

It may be of value to point out the strong resemblance of an expression like (32) to the expression resulting from the nth power of the matrix $\|w_{ij}\|$ with $\mathbf{C}$ playing the role of $\sum$ and "$\cdot$" playing the role of the common multiplication of scalars.

The response $y_{i1}(\tau)$ may be expressed in terms of the initial input $x_{i0}(0)$ at element i_0 namely:

$$y_{i_1}(\tau) = T_{i_1}(x_{i_0}(0))$$

Substituting this expression into (32) yields:

$$y_{i_m}(m\tau) = \overset{n}{\underset{i_1=1}{\mathbf{C}^*}} \overset{n}{\underset{i_2=1}{\mathbf{C}^*}} \cdots \overset{n}{\underset{i_{m-1}=1}{\mathbf{C}^*}} \left[\overset{m-1}{\underset{\lambda=1}{\Phi}} w_{i_\lambda i_{\lambda+1}} \nabla T_{i_m} T_{i_{m-1}} \cdots T_{i_2} T_{i_1}(x_{i_0}(0))\right] \tag{32a}$$

* For a pairwise ∇, Equation (29) could be written in terms of pairs of w's and y's inside the brackets.

2. Non-uniform operators and compatible cascaded networks

Consider again the cascade shown in Fig. 3. Each layer being functionally identical to the other layers. The operator T is again represented by an n-tuple (each layer has n elements)

$$T = (T_1, T_2, \ldots, T_n) \tag{27}$$

where T_i, an operator, represents the input-output relation of the *ith element* of layer. It is required that T_i belong to some class C_{F_i, G_i}. [Ref. (30), Chapters I, II and III] that is if $T_i : DE$ there exists a pair of functions:

$$F_i : DxD \to D$$

$$G_i : ExE \to E$$

such that

$$T_i[F_i(x, y)] = G_i[T_i(x), T_i(y)] \tag{34}$$

Letting

$$F_i(x, y) = x *_i^1 y$$

$$G_i(u, v) = u *_i^2 v$$

relation (34) may be written as:

$$T_i(x *_i^1 y) = T_i(x) *_i^2 T_i(y) \tag{35}$$

Furthermore, it is required that two scalar operations ∇_i^1, ∇_i^2 exist such that:

$$T_i(a\nabla_i^1 x) = a\nabla_i^2 T_i(x) \tag{36}$$

with ∇_i^1, ∇_i^2 defined over the *same* scalar field Γ.

The operations $*_i^k$, ∇_i^k, $(k = 1, 2)$ have the same properties as those required for $*$ and ∇ in the previous section. Namely, the $*_i^k$ are commutative and associative and the ∇_i^k are scalar commutative and $*_i^k$ distributive. In addition it is required that:

$$a\nabla(b\nabla x) = \Phi(a, b)\,\nabla x \tag{25}$$

for Φ some scalar valued function (e.g. $\Phi(a, b) = ab$, field multiplication).

For such operators T_i the output, $y_{i_2}(2\tau)$, of the i_2th element of the second layer is given by:

$$T_{i_2}\left[\mathop{\mathrm{C}}_{i_1=1}^{n} *_{i_2}^{1} w_{i_1 i_2} \nabla_{i_2}^{1} y_{i_1}(\tau)\right] = y_{i_2}(2\tau) \tag{37}$$

where $C*^{1}_{i_2}$ is defined by equation (28), for $*^{1}_{i_2}$ the $w_{i_1 i_2}$ denote the connectivity and gain of the i_2th element in the second layer from the i_1 element in the first layer.

Applying conditions (35) and (36) inductively on (37) one obtains:

$$y_{i_2}(2\tau) = \underset{i_1=1}{\overset{n}{C}} *^{2}_{i_2} w_{i_1 i_2} \nabla^{2}_{i_2} T_{i_2}(y_{i_1}(\tau)) \tag{38}$$

The response, $y_{i_3}(3\tau)$, of the ith element of the third layer is likewise given by:

$$y_{i_3}(3\tau) = T_{i_3}\left[\underset{i_2=1}{\overset{n}{C}} *^{1}_{i_3} w_{i_2 i_3} \nabla^{1}_{i_3} y_{i_2}(2\tau)\right] = \underset{i_2=1}{\overset{n}{C}} *^{1}_{i_3} w_{i_2 i_3} \nabla^{2}_{i_3} T_{i_3}(y_{i_2}(2\tau)) \tag{39}$$

where

$$y_{i_2}(2\tau) = \underset{i_1=1}{\overset{n}{C}} *^{2}_{i_2} w_{i_1 i_2} \nabla^{2}_{i_2} T_{i_2}(y_{i_1}(\tau)) \tag{40}$$

It is at this stage that a *compatibility condition* will be imposed on the network. Namely if element i of layer 2 is connected with element j of layer 1 we will suppose that T_i and T_j are such that if T_i is in the operator class C_{F_i, G_i} and T_j in the class C_{F_i, G_i} then $F_j = G_i$, (Fig. 4). By a known

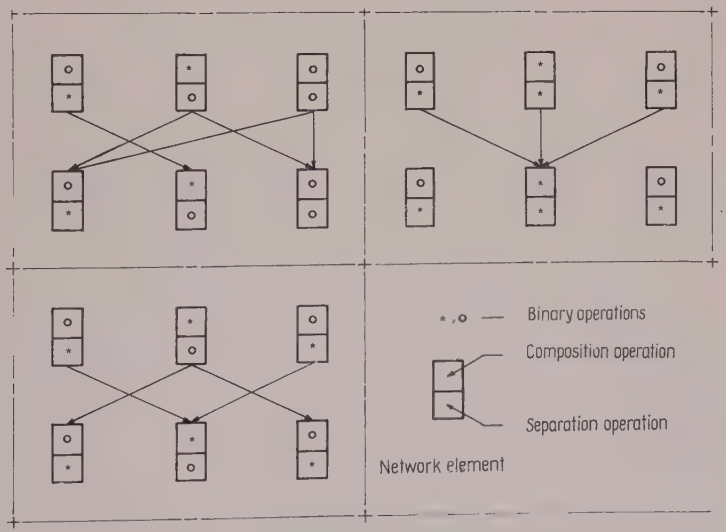


Fig. 4. Examples of cascaded compatible networks with maximal connectivity.

theorem (Ref. 30), Theorem 3b pp. 9 and 10), the composition of the operators, T_iT_j are $*_i^1$ and $*_j^2$ so that:

$$T_iT_j(x*_i^1 y) = T_iT_j(x) *_2^i T_iT_j(y) \tag{41}$$

Similarly for the scalar operations. With such a compatibility condition, substituting (40) into (39) and applying (41) inductively one obtains:

$$y_{i_3}(3\tau) = \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_3} w_{i_2 i_3} \nabla^2_{i_3} T_i \left[\overset{n}{\underset{i_1=1}{C}}{}^{2}_{i_2} w_{i_2 i_1} \nabla^2_{i_2} T_{i_2} (y_{i_2}(\tau)) \right] =$$

$$= \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_3} w_{i_2 i_3} \nabla^2_{i_3} \overset{n}{\underset{i_1=1}{C}}{}^{2}_{i_3} w_{i_2 i_1} \nabla^2_{i_3} T_{i_3} T_{i_2} (y_{i_1}(\tau)) =$$

(By compatibility)

$$= \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_3} \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_3} \left(w_{i_2 i_3} \cdot w_{i_1 i_2} \right) \nabla^2_{i_3} T_{i_3} T_{i_2} (y_{i_1}(\tau)) \tag{42}$$

Proceeding in this fashion, $y_{i_m}(m\tau)$, the output of the i_mth element of the mth layer is given by:

$$y_{i_m}(m\tau) = \overset{n}{\underset{i_1=1}{C}}{}^{2}_{i_m} \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_m} \cdots \overset{n}{\underset{i_{m-1}=1}{C}}{}^{2}_{i_m} \left[\overset{m-1}{\underset{\lambda=1}{\Phi}} w_{i_\lambda i_{\lambda+1}} \nabla^2_{i_m} T_{i_m} T_{i_{m-1}} \cdots T_{i_2}(y_{i_1}(\tau)) \right] \tag{43}$$

where Φ is given by (33) for $\nabla^2_{i_m}$.

Finally, in terms of the initial input $x_{i_0}(0)$, the response y_i (m) is given by:

$$y_{i_m}(m\tau) = \overset{n}{\underset{i_1=1}{C}}{}^{2}_{i_m} \overset{n}{\underset{i_2=1}{C}}{}^{2}_{i_m} \cdots \overset{n}{\underset{i_{m-1}=1}{C}}{}^{2}_{i_m} \left[\overset{m-1}{\underset{\lambda=1}{\Phi}} w_{i_\lambda i_{\lambda+1}} \nabla^2_{i_m} T_{i_m} \cdots T_{i_1}(x_{i_0}(0)) \right] \tag{44}$$

Examples of $C_{F,G}$ operators

1. $T(u) = \operatorname{Log} u$

$$x*_1 y = xy,\ u*_2 v = x + y$$
$$a\nabla_1 x = x^a,\ a\nabla_2 u = ax$$

2. $T(u) = g^{-1} L(f(u))$

where L is some linear operator

$$x*_1 y = f^{-1}(f(x) + f(y))$$
$$u*_2 v = g^{-1}(g(u) + g(v))$$
$$a\nabla_1 x = f^{-1}(af(x))$$
$$a\nabla_2 u = g^{-1}(ag(u))$$

REFERENCES

1. McCulloch, W. S.: (1965) "Living Models for Lively Artefacts," in *Science in the Sixties*, D. L. Arm (ed); University of New Mexico Press, Albuquerque, pp. 73–83.
2. Bigelow, J. H.: "Theories of Memory" *ibid.*, pp. 84–95.
3. (1965) *The Anatomy of Memory*, D. P. Kimble (ed); Science and Behavior Books, Inc., Palo Alto.
4. Rosenblatt, F. *et al.*: (1966) "The Transfer of Learned Behavior from Trained to Untrained Rats by Means of Brain Extracts." *Proc. Natl. Acad. Sci.* **55**, 548–555.
5. Gordon, M. W. *et al.*: (1966) "RNA and Memory: A Negative Experiment." *Am. J. Psychiatry* **122**, 1174–1178.
6. Katz, J. J.: (1962) *The Problem of Induction and its Solution.* University of Chicago Press, Chicago.
7. Shannon, C. E.: (1951) "Prediction and Entropy of Printed English." *The Bell System Tech. J.* 30, 1, 55–64.
8. Foerster, H. von: "Memory Without Record," in *Ref.* (3): pp. 388–433.
9. Wittgenstein, L.: (1963) *Tractatus Logico-Philosophicus.* Humanities Press, New York, Proposition 6.36311.
10. *Ibid*, Proposition 5.1361.
11. Foerster, H. von: (1966) "Time and Memory," in *TIME*, BCL Report No. 3.1, Biological Computer Laboratory, University of Illinois, Urbana.
12. See Ref. (6): p. 3.
13. *Ibid*, p. 36ff.
14. Carnap, R.: (1950) *Logical Foundations of Probability.* University of Chicago Press, Chicago.
15. Reichenbach, H.: (1938) *Experience and Prediction.* University of Chicago Press, Chicago.
16. Reichenbach, H.: (1949) *The Theory of Probability.* University of California Press, Berkeley.
17. Russell, B.: (1948) *Human Knowledge.* Simon and Schuster, New York, p. 350.
18. Luce, R. D. and Raiffa, H.: (1951) *Games and Decisions.* John Wiley & Sons, New York.
19. Pask, G. and Foerster, H. von: (1961) A Predictive Model for Self-Organizing Systems." *Cybernetica* **3**, 258–300 (1960); and **4**, 20–55.
20. Reschler: (1965) *The Logic of Command.* Dover, New York.
21. Hilbert, D. and Ackermann, W.: (1928) *Grundzüge der theoretischen Logik.* Springer, Berlin.
22. Foerster, H. von: (1966) *Numbers of Man, Past and Future*, Biological Computer Laboratory, University of Illinois, Urbana.
23. Szentagothai, J.: (1964) "Anatomical Aspects of Functional Transformation," in *Information Processing in the Nervous System*, R. W. Gerard and J. W. Duyff (ed); Excerpta Medica Foundation, Amsterdam pp. 119–138.
24. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. and Pitts, W.: (1959) "What the Frog's Eye tells the Frog's Brain". *Proc. I.R.E.* **47**, 1940–1951.

25. McCulloch, W. S. and Pitts, W.: (1943) "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bull. Math. Biophys.* **5**, 115–133.
26. McCulloch, W. S. and Pitts, W.: (1947) "How We Know Universals; The Perception of Auditory and Visual Forms." *Bull. Math. Biophys.* **9**, 127–147.
27. Babcock, M. L. *et al.*: (1960) *Some Principles of Preorganization in Self-Organizing Systems.* Tech. Rep. No. 2, ONR Contract 1834(21), Electrical Engineering Research Laboratory, Engineering Experiment Station, University of Illinois, Urbana.
28. Foerster, H. von: (1962) "Circuitry of Clues to Platonic Ideation," in *Artificial Intelligence*, C. A. Muses (ed); Plenum Press, New York, pp. 43–82.
29. Foerster, H. von: (1967) "Computation in Neural Nets", *Currents in Modern Biology*, **1**, 47–93.
30. Inselberg, A.: (1965) *On Classification and Superposition Principles for Nonlinear Operators.* Tech. Rep. No. 4 U.S. AF-OSR Grant 7-64, Electrical Engineering Research Laboratory, Engineering Experiment Station, University of Illinois, Urbana.

W. ROSS ASHBY

University of Illinois

Urbana, Illinois

Some Consequences of Bremermann's Limit for Information-processing Systems

Many limitations stand in the way of advanced computing. Some of them, such as the limitation by budget, may be removed at any moment. Others, such as that due to the structure of the language by which scientists communicate with one another, could perhaps be removed with sufficient effort. I wish here, however, to consider the consequences of a particular limit established by Bremermann (1962, 1965): "The capacity of any closed information transmission or processing system does not exceed mc^2/h bits per second." (m is the mass of the system, c the velocity of light, and h Planck's constant). Putting m equal to 1, and giving c and h their known values, the limit is numerically about 10^{47} bits per g per sec.

Centuries of time and tons of computer merely raise the quantity to about 10^{80} bits. Beyond this quantity nothing made of matter as we know it can go.

The fact that the limit allows up to about 10^{80} bits may seem to be practically no restriction at all. In fact, however, the processes that we hope to use in advanced (brain-like) computing encounter this limit only too quickly. To establish the fact, I give two typical examples:

1: A screen of lamps, 20 × 20, each of which is only lit or unlit, presents pictures that we want to group into those that have some property and the remainder. Suppose we ask "what is the best grouping?" This unpretentious question asks for one from a set. Of lamps there are 400, of pictures there are 2^{400}, i.e. 10^{120}, and of groupings there are 2 raised to this power. So the selection of a particular grouping from this set demands (unless other restrictions intervene) no less than 10^{120} bits. The apparently simple question thus makes a demand that goes far beyond the limit.

The work on which this article is based was supported partly by Contract AF 33(615)-3890 with the Research and Development Division, Air Force Systems Command.

2: An artificial retina has a million sensitive units, each of which can only be excited or not-excited. It acts through a net that produces, as output, only a 1-bit move or not-move. Suppose we ask "what is the relation between input and output?" The question asks, essentially, for the mapping from the set of input states ($2^{1,000,000}$ in number) to the set of output states (2 in number). The number of mappings is the output number raised to the power of the input number. So the selection of a particular mapping from the $2^{(2^{1,000,000})}$ demands (unless other restrictions intervene) no less than $10^{300,000}$ bits. Again, an apparently simple question has demanded a quantity of information-processing that goes far beyond the limit.

These examples may suffice to show how easily we may ask questions, or demand computational processes, that go far beyond Bremermann's limit. Instead of being a remote and almost imaginary curiosity, it stands right in our way as soon as we attempt the more advanced forms of information processing.

The consequences of this limit are various. Here I shall mention only a few that seem to me to be outstanding in the context of bionics.

As "regulation and control" are of the highest practical importance, let us first apply the limit here. A simple example may help to get the basic ideas clear. Let us suppose that a fleet, just as it is about to leave port on active service, discovers that its communication devices for ship-to-ship coordination have failed; as a result, it has to put to sea with only some human signallers equipped with hand-operated flash lamps. Here we have a dynamic system, with a goal clearly defined (by current naval strategy), and that is subject to a limit on the amount of communication that can occur internally, determined by the capacity of the signallers. Now it is clear that the admiral may dispose of his signallers in various ways, and there may be no single maneuver of the whole fleet that can be declared impossible, yet common sense tells us that this fleet's maneuvers are going to show special features that, for instance, the enemy admiral may well notice after a time.

"Achieving coordination in maneuver" means that the total set of all possible combinations of movement (including those that lead at once to collision) are to be restricted to a special subset of the combinations (those combinations approved by naval strategy.) Achieving the restriction *demands* the corresponding quantity of transmission (by Shannon's tenth theorem or by the law of requisite variety.) Thus, to be

more definite, suppose that there are 100 ships, that the only requirement in maneuver is that all ships shall turn in the same direction, and that the signaller's total capacity as a channel provides 200 bits per course-setting. Such a fleet can coordinate its directions to the degree of choosing between port, starboard, and ahead (for $99 \log_2 3$ is less than 200) but no distribution of signallers or arrangement of coding can refine the selection of direction to adding half-to-port and half-to-starboard (for these would require $99 \log_2 5$ bits, which is greater than the 200 bits available). Thus, *the existence of a limit on the total quantity of information transmissible puts an absolute limit to the amount of regulation or control achievable.*

The arithmetic of this example shows that Bremermann's limit is no immediate threat in the case of regulations that are direct. A million ships, all having to move correctly to one part in a million would only demand $10^6 \log_2 10^6$ bits per course-setting, i.e. about 2×10^7 bits—nowhere near the limit. But this smallness does not mean that the limit can be forgotten when we change to the bionic sciences. Here the regulation and control is often directed at some *complexly patterned* event, with strong interactions between the parts (or all statements highly conditional). In such cases the quantities of information tend to increase (when the number of components is increased) at the explosive exponential rate rather than at the moderate multiplicative.

A well-known example of the effect of a complex goal is given by the mechanical chess-player. The goal ("achieve mate") looks simple, but at the present time the only sure method known for specifying what this means in individual plays is to write out all possible plays and to label each as "good" or "bad". Since the number of plays is at least 10^{120}, Bremermann's limit is an impassable barrier. Since the game of chess is simpler than the battle of life, we may expect that his limit, far from being a mere numerical curiosity, will impose itself frequently in real and practical situations.

The reason for the sudden jump from the moderate quantity of information used by the fleet to the immoderate quantity demanded by chess is, of course, due to the *combinatorial* quality of chess. Whether a piece's position is good or bad *is conditional on* where the other pieces are. The conditionality makes the variety grow combinatorially (often exponentially), where the simpler forms grow only additively or at a simple multiplicative rate. Since in bionics, and in advanced computing, we are

specially concerned with these combinatorial processes, it is particularly in our science that we are likely to encounter the limit early in our work. The topics that are specially likely to imply a major degree of interaction between the parts are especially those involving the concepts of:

System	Order
Organization	Subset
Pattern	Property
Net	Relation
Automaton	Constraint

all of them highly relevant to "advanced information processing" and "the mechanical brain".

We are thus likely to encounter the limit at an early stage in our researches, especially in bionics and artificial intelligence. But the theme has much wider implications in philosophy, at which I would like to glance.

The most obvious fact is that we, and our brains, are themselves made of matter, and are thus absolutely subject to the limit. Not only are we subject as individuals, but the whole cooperative organization of World Science is also made of matter, and is therefore subject to it. Thus both the total information that I can use personally, and the information that World Science can use, are limited, on any ordinary scale, to about 10^{80} bits. Whatever our science will become in the future, all will lie below this ceiling.

We cannot claim any special advantage because of our pre-eminent position in the world of organisms. We have been shaped, and selected to be what we are, by the process of natural selection. As a selection, this process can be measured by an information-measure; it is therefore subject to its limits. In any type of selection, under any planetary conditions, a planetary surface made of matter cannot produce adaptation faster than the rate of the limit. However good we may think we are, 10^{80} measures something that we do not exceed. The science of the future will be built by brains that cannot have had more than 10^{80} bits used in their preparations, and they themselves will advance only by something short of 10^{80}. This is our informational universe: what lies beyond is unknowable.

We can see something of what will be unknowable. Sometimes nature's laws have a simple informational structure. The law of gravity, for in-

stance, has been found to relate the attractions between two particles, i and j say; and this relation is *unconditional* on the positions of other particles, k, l, m, ... etc. This unconditionality means that the complexities go up, as more particles are added, in a more or less additive way (the potentials do, in fact, combine simply by addition). Contrast this case with (say) a social system, in which the relation between two variables i and j may itself depend on other variables. This would be as though, in gravity, the law of attraction between i and j were altered by the position of k. Here the complexity goes up in some manner approximately exponentially. Thus the existence of the limit tells us that our achieved science will always be one of the world in its simpler interactions. If there are complex natural laws, we shall never know them.

The limit is thus likely to be specially obstructive in the sciences of the complex. One of these is sociology, just referred to as an example. The other is our own science of bionics, especially when we attack the problems of artificial intelligence. What should we do?

One reaction to the limit is simply to ignore it, noticing it only when we must. But the history of science has shown repeatedly that when an awkward limitation appears the science tends to become sterile until it has actually made the limitation a part of its working conceptual structure. The early microscopists, for instance, treated the limitations imposed by light's finite wave-length as a mere nuisance. Seeing was believing, until Abbe and Helmholtz developed the new microscopy, in which the wave features of diffraction and interference became *intrinsic* working parts of the theory. Atomic physics, too, ran into evergrowing troubles until it recast its basic ideas and constructed a new theory with the basic limitations, due to quantum restrictions and indeterminancy, built into it. Thus, there seems good reason to suggest that our best way, in the face of this limit, is to study it and to make it an integral part of our working ideas.

How is this integration to be achieved? I can here offer only a slight suggestion, in the hope that it will be found useful. Most of this work lies in the researches of the future.

First, we know that the mathematicians and engineers have derived great advantage from their development of the "linear" processes: matrix algebra, the Laplace transform, etc. With these processes they can work extensively in the linear world without the danger of breaking, at each operation, into the far more complex world of the non-linear.

This example shows that an extensive set of operations can be developed such that a great deal of worthwhile work can be done within the set, with the operations themselves *automatically* preventing the worker from wandering into the "forbidden" regions. Bremermann's limit specifies just such a region.

Now Minsky (1963) has summarized the essence of the problem of "artificial intelligence" in words with which I entirely agree: "The real problem is to find methods which significantly delay the apparently inevitable exponential growth of search trees." So far as the system studied is genuinely combinatorial, so far is the exponential growth inevitable, and Bremermann's limit acts with maximal intensity. But a large proportion of our problems in bionics are in fact subject to strong internal constraints, (most of them derived ultimately from the intense redundancy and repetitiveness shown at the atomic level.) One of the most general and wide spread constraints is that the system is to some degree reducible, i.e. capable of being studied piecemeal. When it is so, a system that seems to demand excessive information-processing may in fact allow its study to be achieved with less. (The essential reason is that if a quantity that increases exponentially, as a^n, can be treated in k stages, the branches fall to the order of $ka^{n/k}$. When n is large, the effect of k on the exponent, by dividing it, is far more powerful than its effect as a multiplier.) The method "consider the problem a piece at a time" is so widespread and so powerful that it may well be worthwhile to attempt the development of *all those operations that do not destroy reducibility*. When we know the set, the operations in it will form a calculus like those of the linear systems—such that we may do what we like within the set without fear of converting the problem from one solvable under the limit to one no longer solvable under it. A start in this direction has been made by the formulation of "cylindrance", (Ashby, 1965) which measures, for any relation between n variables, the degree to which it can be treated as if made of sub-relations, each on only some subset of the variables. It treats not only the fairly obvious case in which the relation consists of k wholly independent sub-relations but also the much more interesting case in which the whole relation has something of the simplicity of a k-fold division while being in fact still connected. (An elementary example is given by a country's telephonic communications, in that although all subscribers are joined potentially to all, the actual communications are almost all by pairs.)

The limit (of about 10^{80} bits) implies that we can never study the fully general relation between more than about 270 variables. (10^{80} bits allows us to pick one arbitrary subset from 10^{80} elements; 270 binary variables provide this number.) Since the cylindrance (a measure of intrinsic complexity) cannot exceed the number of variables, the limit implies that we can never study the fully general relation whose intrinsic complexity (if measured by cylindrance) exceeds 270.

If therefore we intend to study a system (a living brain perhaps) in which the relations do not have a cylindrance exceeding 270 we have a system that is potentially studiable. If now we unwisely ask questions or perform operations that raise the cylindrance above this number our very method of study has rendered it unstudiable. It is now known that cylindrance is safe under the operation of intersection (when the relations are treated as subsets of a product space) but that it may readily be raised by union.

This work is still in progress, but it already shows that there may exist methods, specially suited to the study of the complex system, whose operations do not lead us to the humiliating situation in which we discover that it is our own methods that have turned a potentially studiable system into one that, under the limit, is now essentially unstudiable.

SUMMARY

That nothing made of matter can transmit or process information faster than 10^{47} bits per g per sec may seem of small practical importance. In fact, many of the processes that have been proposed for machines with artifical intelligence require transmissions far in excess of this limit. Examples are given to show that large-scale processes of combinatorial richness run into the limit only too easily.

Not only are our machines so restricted, but the scientist's brain, made of matter, is also so restricted. Thus our personal knowledges, our philosophies, and our science are also limited to the same degree.

Some of its consequences in science are discussed. If our science is to be realistic, our theories must be structured so that this limit becomes an integral part of them. A suggestion is made of one way in which this incorporation might be achieved.

REFERENCES

Ashby, W. Ross (1965). Constraint analysis of many-dimensional relations. In: *Progress in biocybernetics*, edited by N. Wiener and J. P. Schade, Elsevier Publishing Co., Amsterdam; pp. 10–18.

Bremermann, H. J. (1962). Optimization through evolution and recombination. In: *Self-organizing systems* 1962, edited M. C. Yovits *et al.*, Spartan Books, Washington, D. C., pp. 93–106.

Idem (1965). Quantum noise and information. *5th Berkeley Symposium on Mathematical Statistics and Probability*; Univ. of California Press, Berkeley, California.

Minsky, M. (1963). Steps towards artificial intelligence. In: *Computers and Thought*, edited E. A. Feigenbaum and J. Feldman, McGraw-Hill Book Co., New York; pp. 406–450.

R. L. BEURLE

R. B. GUY

D. C. H. LONDON

Department of Electrical and Electronic Engineering

at the University of Nottingham,

Nottingham, England

Artificial Intelligence and the Nature of the Environment

INTRODUCTION

Although one of us pointed out some of the possibilities arising out of randomly connected networks in the early days of the recent upsurge of interest in "intelligent devices", there has been an intervening gap of more than ten years. During this time, considerable advances have been made in empirical approach to neurophysiology and in the results this has produced. Moreover, a large amount of effort has been deployed on the development of a variety of models to represent various "intelligent" processes.

Models have been made to represent processes which appeared to take place in the central nervous system of various species, and to represent basic processes such as the conditional response, association, the assessment of conditional probability and so on. These models have been both conceptual, consisting of theoretical arguments or proofs, and practical, comprising electronic and other devices built into networks of various descriptions.

When the opportunity arose recently to re-examine the present state of study in this field, a number of basic facts seemed to emerge, namely:

1. That most attempts to study intelligent behaviour, only consider situations that have been grossly simplified by contrast with the "real life" situations encountered by even the more modest of intelligent animals.

2. That there are a number of different facets to the phenomena which are for convenience lumped together under the description "intelligent behaviour".

3. That individual authors have tended to concentrate on one "aspect" or "facet" of intelligence, whereas natural intelligence may well be an integrated "system" in which the interplay between the component parts is as important as the contribution of any one part.

4. That in natural intelligence, economy of "hardware" is as important as the sophistication of the tasks performed.

5. That there has been a tendency among writers in this field to take the nature of the problem to be solved by an intelligent device for granted. The assumptions regarding the environment in which the intelligent device is taken to operate are often over-simplified, without being stated explicitly.

We propose to take this opportunity of looking a little more closely at the nature of the environment. In doing so, we recognise the necessity, at this stage of understanding of the subject, for some simplification, particularly with regard to time scale, and we shall stress the limitations that this imposes. We shall then comment on the relevance of our conclusions to the form of an "intelligent device" in which economy of logical units is an important factor.

DEFINITIONS

The word intelligence can be used in many different ways to cover various activities associated with the manipulation of information. It has become conventional to speak of intelligence as the ability to associate, the ability to think or reason, or the ability to discover an appropriate action. In discussing the subject here, we do not wish to exclude any process which allows the intelligent organism to optimise its response to a particular set of circumstances. An intelligent process thus implies an ability to adapt so that, if the initial response is found less than ideal, it can be modified according to experience. Moreover, if circumstances change, the response can change too. An intelligent device need not be limited solely to adaptive behaviour. If there is an ideal response which is known to start with, the device can be built to give this response in the first place. The adaptive ability gives the intelligent device an

advantage when the ideal response is not known to start with, or when circumstances require a changing response.

The intelligent device is thus one that can to some extent adapt to its surroundings and it is always implicit, if not expressly stated, that there is some task to be performed or goal to be achieved. With living organisms this ultimate goal may generally be regarded as survival, either of the individual or of the species. In general, the task is to achieve some specified aim in some specified environment.

This means that the task is defined in terms of the environment and cannot be dissociated from it. This, in turn, means that the optimum structure of the "intelligent" device is defined in terms of the environment and the task, and must thus reflect the nature of the environment. An "intelligent" machine which is ideal in one environment may be worse than useless in another. When we take into account the need for economy in units for which the intelligent device is built, it is particularly important to utilise all available information about the environment to optimise the structure of the device for the task it is to perform.

ADAPTIVE BEHAVIOUR – THE DEVICE AND THE ENVIRONMENT

Conventionally, intelligent or adaptive behaviour includes trial and error learning, association either without reinforcement or with it, various decision making procedures and methods of pattern recognition, and template fitting.

To clarify our approach, and to define the terms used, it is proposed to discuss the relationships between the environment and the device or organism interacting with it, for both trial and error learning, and association.

1. Trial and Error Learning and the Environment

Taking pure trial and error learning first, the essential elements in this situation are shown in Figure 1, and are

1. The environment (E)
2. The adaptive part (Q) of the device or organism

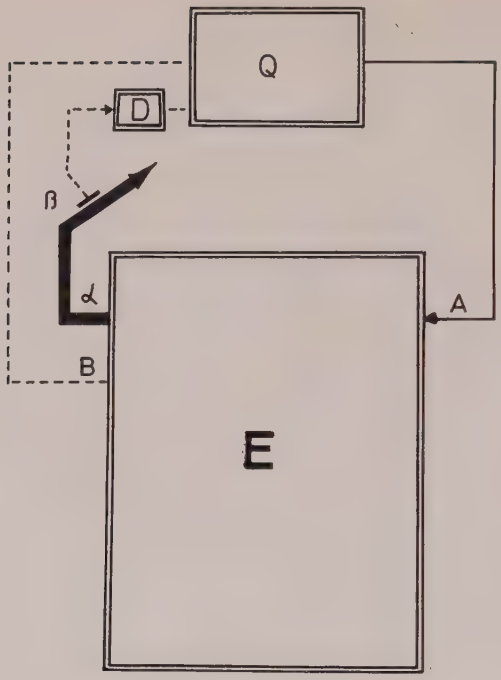


Fig. 1

3. Means to enable the device to act on the environment (A)
4. Means whereby the environment can react (α), and sensory means (β) with some means for discrimination D to enable the device to determine whether this reaction is favourable or unfavourable.
5. Means (B) to enable the device to make observations and so gather limited information on the state of the environment.

It will be obvious that the relationship between α and β is of critical importance. In a living organism, it is the reaction α of the environment directly or indirectly on the organism that is important. In an artificial device, α will be the performance, successful or otherwise, of an appointed task. The sensory input β to the discriminator D is the means which enable the device to sense whether or not it is succeeding in this task. If D embodies a precise measure of the success of this task, then the difference between monitoring β and α is of no significance. If D embodies an indirect measure, then to the device, success is defined as maximising whatever is defined by D. Here it is evident that foreknowledge is required

of the nature of the environment sufficient to ensure that what is defined by D is, for the purpose in hand, a good measure of the success of the task required of α.

In biological organisms, where one may take survival as the ultimate criterion, the success defined by D is usually a relatively short-term intermediate aim, e.g. taste, satisfaction of hunger, etc. It is obviously most important that the intermediate aim should be related to the environment in such a way that the ultimate goal is fulfilled.

The requirements that a device must meet in order to perform trial and error learning under such circumstances as these have been discussed many times. Let us look instead at the requirements imposed on the environment, In Figure 2, the action paths A and α, and information channels β and B have been continued inside the rectangle representing the environment to illustrate the minimum requirements of an environment in order to make possible a useful trial and error learning process.

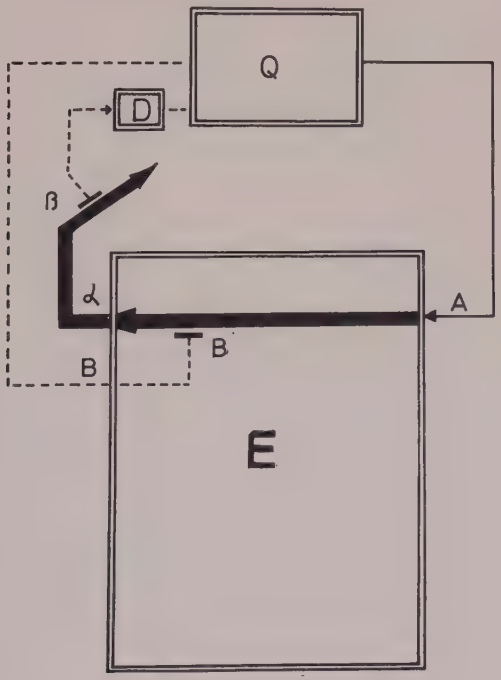


Fig. 2

If these connections must exist between the device and the environment, then they must also exist within the environment, and the only question is the exact form which they take. $A \rightarrow a$ is a path along which action can travel and it is assumed to exist in some but not necessarily all states of the environment. The $B - B$ information channel represents the fact that, for learning to be possible, the observations B which the device can make on the environment, must, at least, on some occasions reflect the presence or absence of the $A \rightarrow a$ path.

Given a number of action paths and information channels related in this way in the environment (Fig. 3), then in principle, trial and error

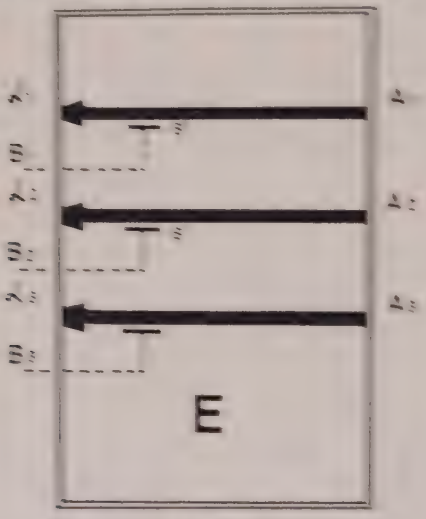


Fig. 3

learning depends only on the existence of a suitable device or organism. Without such paths and channels trial and error learning is impossible, and connections of any other type in the environment do not affect the practicability of pure trial and error learning. One may ask, from either a philosophical or a practical point of view, what is the relationship between the observation B and the $A \rightarrow a$ path about which it carries information. This question may be relatively simple to answer for any specific example in any particular environment, but is less easy to answer in general terms. There are several possibilities.

(a.) The observation B may be a direct observation of the existence of the physical means whereby A is to mediate a.

(b) It may be an observation of something that is likely to be caused by the pressure of the A—α path.
(c) It may be an observation of something that is likely to be initiated by the same cause as that which initiated the path.
(d) It may be an observation of the cause of the existence of the path.

These are only the simplest of the relationships that might exist. The possibilities are so numerous that, in general terms, it does not appear to be possible to do more than state that the existence of such action paths and information channels is a pre-requisite for trial and error learning. If these do not exist in the environment there can be no trial and error learning.

2. Direct Association

Now let us consider pure association in its simplest form. We do not propose to discuss how association is dealt with in the device or organism yet. This has been discussed from various points of view in a number of papers. Figure 4 illustrates the essential elements involved in this. There is an A—α path together with the B_1 channel that provides the information to which the device has an established response A. From the point of view of association, it is immaterial, when considering the environment, how the B_1—A—α response was established. It could have been learned by trial and error, or by previous association, or it could have been a pre-determined response. The β channel is not shown and is not involved in association.

B_2 signifies a second observable state of the environment which also reflects the presence or absence of the A—α path. For the device to be able to discover this B_2—A—α response without recourse to trial and error learning, there must be some recognisable similarity between the observable states B_1 and B_2. Otherwise, the device has no means of distinguishing B_2 as more likely to reflect the state of the A—α path than B_3, B_4, etc. The existence of this recognisable similarity is denoted by the symbol σ between the two B channels. Thus, the minimum requirement for simple association is the existence of action paths such as A—α, together with pairs of observable states such as $(A$—$\alpha)$—B_1, and TA—$\alpha)$—B_2 having sime observable similarity. Given pairs of observable states, linked in this way with an A—α action path, association

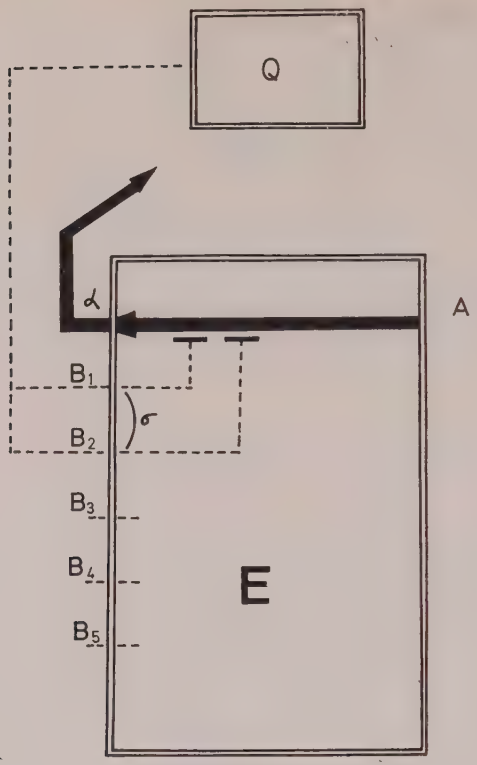


Fig. 4

can take place if a suitable device having the ability to form associations is available. Failing such related pairs, association is not possible.

The term recognisable similarity used in the last paragraph is of some significance. For two observable states to be recognisably similar, they must have something in common. Let us suppose that an observable state B contains many elements which we may, for convenience, designate as b_1, b_2, b_3, ..., taking these to be binary elements. Then, for example, B_1 and B_2 may have the form

B_1 11100101010101011110010

B_2 10110101010101010011001

The example has been chosen because the existence of a recognisable similarity is obvious.

If the order of b_1, b_2, b_3, etc., is of no significance, then similarities or coincidences of this general type are the only possible form of direct similarity that can exist between a pair of observations, and for convenience, examples of this type of association will be referred to as "Direct" or "1st Order" association type "*a*".

If b_1, b_2, b_3, etc., are ordered in some way which has an 'a priori' significance in relation to the information they convey about the environment, then a further class of similarity is possible. As an illustration if b_1, b_2, b_3, are ordered sequentially, so that a shift does not invalidate a similarity, we may have a similar pair

B_1 11100101010101011110010

B_2 00110110101010101010011

For such a pair to be recognisable, it would only be necessary to incorporate in the device some shifting arrangements to make all the necessary comparisons possible. There is an obvious example of this type of similarity. It is that the image of an object can be shifted in its position on the retina but can still be recognised as coming from the same object. It is not yet clear from empirical evidence to what extent this ability is learned, but the recent evidence from Hubel and Wiesel suggests that at least a part is innate. Whichever it is, it provides a simple means of accomplishing associations between similar visual images in slightly different positions in the visual field. This will be referred to as "Direct" or "1st Order" association type "*b*".

Leaving aside the rather obvious question of a simple shift in the visual field, what is the relationship between the A—α path and the observations B_1 and B_2 that provides the recognisable similarity between the latter? We noted earlier that four relatively simple relationships were possible between an observation B and A—α path. There is an obvious possibility of two observations having some similarity either:

(a) because they are observing the same thing, or
(b) because they are observing things that are related because they have a common cause,
(c) because they are observing things that are related as cause and effect.

It is not difficult to think of practical examples of these categories. Between them they cover such a wide range that it is impossible to be

more specific about the detailed relationship. Nevertheless, it is evident that for direct association to take place, there must be paired observable states having some recognisable similarity and both reflecting the presence or absence of a particular A—λ path. If these relationships do not exist in the environment there can be no learning by direct association.

3. Indirect Association

So far, we have only considered direct association where there is a recognisable similarity between the two B's, requiring at the most (for type "b") a knowledge of some "shift" or transform action which is a function of a known topological relationship between the elements. However, if we look at the environment again, we find that it is possible for there to be a relationship between two B's which is not evident from examination of these two B's alone. The two B's may also be associated indirectly in the sense that the relationship between them is the same as a relationship which has previously been found between other linked pairs of B's. This is illustrated diagrammatically in Figure 5 and will be referred to as second order indirect association.

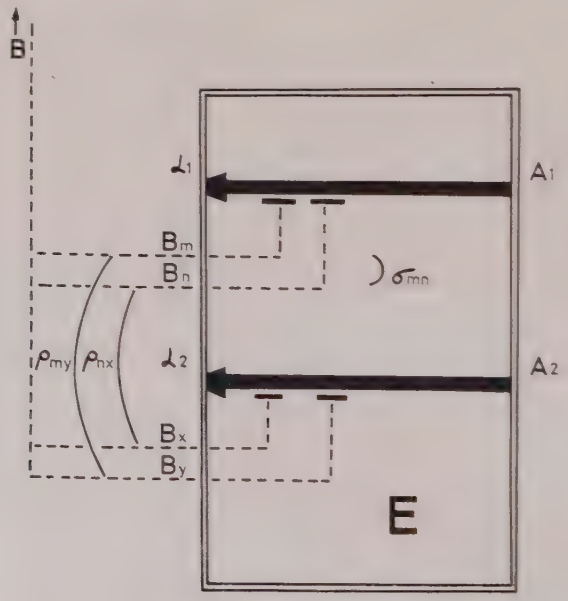


Fig. 5

In Figure 5 B_m and B_n are observations which have already been found to carry information about the same A_1—α_1 path. They may be said to be linked through the A_1—α_1 path, and this is indicated by the symbol σ_{mn} which in this case does not indicate any observable similarity. The observation B_x is already linked with an action link A_2—α_2 and we have the problem of finding a suitable action for observation B_y. If it is found that the relationship between B_m and B_x is the same as the relationship between B_n and B_y, then B_x and B_y may be said to be associated through B_m and B_n. That is, there could be an argument for using A_2 as a trial action for B_y unless some other trial action is justified on stronger grounds. Whether there is any strength in this argument depends on our knowledge of the environment. All we have postulated so far is that there are two observable relationships $B_m : B_x$ and $B_n : B_y$ which relate the linked pair, m, n with the heretofore unlinked pair xy. We are only justified in using this as a guide to the choice of action A_2 if our foreknowledge of the environment tells us that under such observed circumstances, there is an above average chance of B_y in fact carrying information about the presence of the A_2—α_2 path. If this is not so, then to pick action A_2 will not improve the chance of success over and above a purely random choice of action. We may sum this up by saying that there can be indirect association if there are pairs of observable B's linked through A—α paths, provided similar observable relationships exist between the two B's of one pair and the two B's of another pair. We propose to call this second order association because it involves comparison of two pairs of observable states.

The relationship implied in second order association may be written:

$$B_m : B_x :: B_n : B_y$$

The complementary relationship

$$B_n : B_x :: B_m : B_y$$

would of course have had equal significance. For simplicity, in Figure 5, the symbol ϱ has been used to indicate a relationship of the type we have been discussing and in terms of ϱ we could re-write the second equation:

$$\varrho_{nx} = \varrho_{my}$$

To examine the merits of an action A_2 for an observation B_y, we must examine relationships such as ϱ_{mx}, σ_{mn} and ϱ_{my}. If it possible to follow

round such a path and by virtue of the fact that $\varrho_{nx} = \varrho_{my}$ and return to x having started from y, then this constitutes an indirect association between x and y on which a trial of A_2, when B_y is observed, may be justified. In the illustration we have taken, σ_{mn} represented two B's linked by a common action. It could equally well represent two B's associated by a similarity in form.

With indirect association, by analysing the relationship ϱ, one can again identify two types. Type (*a*) involves identification of and positive or negative weighting of the elements b which are found to contribute to the similarity found significant in pairs of linked observations. In type (*b*) the information available from the environment is in the form of empirical cross-probabilities which may exist between every pattern of b's occurring in one B of a pair and every pattern of b's occurring in the other. These then provide a relationship which would enable a suitable transformation to be tried on subsequent pairs of B's to estimate the probability of their being linked to a common A—α link.

Second order association has been mentioned to contrast it with first order association. There is of course no reason to stop at second order.

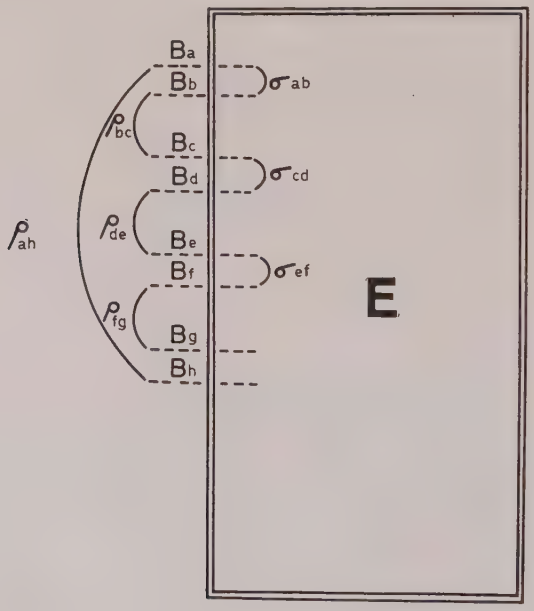


Fig. 6

Association information in some environments may well involve multiple relationships of higher orders and failure to use this information will give a less than ideal device. Figure 6 illustrates a situation of this type where a sequence of several relationships ϱ_{fg}, σ_{ef}, ϱ_{de}, etc., etc., is involved in associating B_x with B_y by virtue of the fact that there is a continuous series of such links between them.

4. The Nature of the Information

So far, we have not attempted to consider how an intelligent device would make use of this information, but only the nature of the information which is in principle available from the environment. We propose to follow this theme a little further by discussing the statistical form of the information available.

For direct association, is has been customary to think of Bayes' theorem as giving the ideal criterion by which trial responses should be chosen. Bayes' theorem is indeed the most satisfactory tool to employ, but it is important to bear a number of facts in mind. Bayes' theorem is derived on the assumption of a limited ensemble of observations of which the particular event under judgment is a fair sample. It assumes that "a priori" information is available regarding the relative probabilities of occurrence of all possible observations (the *B's*) and of the joint probabilities of occurrence of these with the events (or action paths) to be predicted. Moreover, when Bayes' theorem is stated in its normal simple form, the B's are mutually exclusive. In any real problem one usually does not have a set of mutually exclusive observations of the complete environment. One usually has a number of observations of elements (the *b*'s) of the environment, or of limited patterns of these elements and one has to estimate corresponding "*B*" probabilities. This in turn usually involves combining probabilities of elementary events, which can only be done if one knows the regression coefficient between these. Often the assumption of independence is made without adequate confirmation, and thus the precision which would otherwise come from the use of Bayes' theorem is lost. This is not to deny that Bayes' theorem can sometimes still turn out to be a useful guide. In fact it is the only sound guide, but it can only be used if we have the requisite information about the environment or empirical evidence of its success in the environment in which it is being used.

THE NEED FOR FOREKNOWLEDGE

We have been at pains to emphasise these points because it seems to us that for any intelligent device to operate in an environment, some distinct foreknowledge is required of the nature of that environment and of the task the device is to perform in it.

One may ask why the structure of the environment has been accorded so little attention in the design of various "universal" adaptive machines or logical machines that have been built or at least described. Our impression is that the originators of these machines have relied extensively on their own experience in designing them. Thus, perhaps without fully realising it, they have incorporated enough "knowhow" in the structure of the machine to enable it to operate successfully in the environment it was intended to operate in, or they have picked test environments which were of such a nature that they could demonstrate the qualities of the machine. The success of these machines is often a measure of the intuition of their designers. The general problem appears to us to be essentially one of knowing the statistical nature of the environment and this appears to be particularly important where economy is essential in the physical hardware in which the intelligent device is interpreted.

1. Acquisition of Foreknowledge about the Environment

How do we acquire the foreknowledge we need about the environment? There appear to be three basic methods. If it is an artificial environment we shall probably know something about its physical structure and from this its repertoire of responses may be computed. Thus we may be able to draw conclusions as to the nature of the associations likely to exist in it. If it is a natural environment such as we encounter in the real world, there are two possibilities. We may experiment with it for a sufficient length of time to give us some confidence that we know enough about its behaviour. This is what the automatic control people refer to as "identification" of the parameters of the transfer function. This is, however, not a complete answer, for how are we to judge whether the length or the nature of our experience is sufficient if we do not know something about the environment in the first place?

There is in fact no complete answer in a natural environment from which we can only extract information by experiment. However much

information we may acquire, we cannot be sure that there is not some higher order associative information, as yet undiscovered, waiting to upset the conclusions we have already reached. The possibility of some such "booby trap" is always present in an unknown environment. One can only give a precise description of an "optimum" adaptive device in relation to specific information about a known environment.

We have still to mention the third method, if it can fairly be called a "method". This is what happens in real life, where information is obtained by the simple expedient of natural selection as organisms interact with the invironment. Those organisms endowed with forms of intelligence more appropriate to the environment compete more successfully for the means for survival. Selection is a powerful method of transferring information, and those organisms that finally survive have embodied in them the foreknowledge which maximises their chances of success.

2. The Nature of the Foreknowledge Required

To return to the nature of foreknowledge required about the environment, this appears to be of two distinct types. Firstly, knowledge of the environment is essential to define the tasks to be performed, and to define a suitable intermediate goal or goals in the descriminator *D*. The nature of these tasks, and the relation of these to the intermediate goals is so dependent on the particular problem that it can hardly be discussed in the abstract.

Secondly, general knowledge of the nature of the associative relationships in the environment is required, in order to make possible an economical intelligent device. If information is predominantly in the form of high order association, then provision must be made to use this form of information efficiently. If a certain form or order of association is absent in the environment, then there is no value in making provision for this in the intelligent device. Not only is there no value in doing so, but environments may actually exist in which it is detrimental to do so. One might refer to these as "booby trap" environments.

These are the two types of information which are essential if we are to make an intelligent device to react successfully with the environment. There may sometimes be more than this minimum of information available. There may be some information about the details of actual asso-

ciative relationships in the environment. The question of using all three types of information will be discussed in the following section.

UTILISATION OF INFORMATION IN AN ARTIFICIAL INTELLIGENT DEVICE

1. Adaptive, Intelligent, and Pre-determined Behaviour.

Before discussing this subject further, it will be necessary to define the terms we shall need to use. In this discussion, a distinction will be made between an intelligent device and the adaptive part of such a device. We shall take it that it is the ability to acquire information from the environment and to modify its behaviour according to that information that is the distinguishing characteristic of an adaptive device or of the adaptive part of an intelligent device. It is this ability to extract detailed information which enables them to interact successfully in an environment about which there is insufficient detailed information to design devices with more specific responses.

Devices with a "pre-determined" response can, of course, be very useful and many are of great complexity, but they are characterised by an inability to extract associative information from the environment. Any information about the particular associations in the environment on which their operation depends, must therefore be built-in *in detail*. We shall take it that an intelligent device may incorporate adaptive behaviour as well as pre-determined behaviour according to the task set by the environment in which it has to operate.

We have already referred to the need for general knowledge about the information channels and the associative relationships in an environment. This information will usually be of a statistical nature, and we must contrast it with the detailed information about particular associations (i.e. the retationship of a particular B with a particular A–α path) which any organism must acquire by experience in order to produce the right reaction to each particular observed set of circumstances. It is this general information about the nature of associative relationships in an environment, that makes it possible, in principle, to design the adaptive part of an intelligent device to extract from the environment the detailed information required for the task it is to perform.

If some detailed information is available when the intelligent device is first designed, this may, of course, be built into the device too. It is indeed highly desirable that this detailed information should be built in, so as to conserve the time and effort which would otherwise unavoidably be spent by the intelligent device in extracting this information from the environment. There are various forms of detailed information which may be "built in" in this way. The most important are those that reduce the complexity of the adaptive operations that have to be performed.

Associations in the environment that cannot be predicted in detail, cannot be built in. The details of these associations must all be extracted from the environment, and only an adaptive device can extract this information. The device cannot react usefully until it has acquired information in this way. The situation is different if there are associations of practical significance that can be predicted in detail. The existence of such associations enables us to perform a part of the analysis of the incoming information without recourse to adaptive techniques.

If we perform this part of the analysis directly on the incoming information from the environment, we avoid overloading the adaptive part of the device. This is always a sound policy, for an adaptive mechanism, having a much more difficult task to perform, is more complex than a pre-determined mechanism. Moreover, the adaptive device, having to acquire information before it can use it, is slower. If we perform this analysis, which we may well call pre-analysis, before the adaptive part of the device, we then present a simpler task to the adaptive part. The remaining analysis is something which only an adaptive mechanism can perform. To employ the adaptive mechanism to perform the whole analysis, means that the adaptive mechanism must rediscover all the detailed associative information which is already known, and this is obviously inefficient.

We now have a picture of the basic relationship between the environment and the device, as shown in Figure 7. This emphasizes that the foreknowledge available about the environment can be of three distinct types. The knowledge of the basic nature of the environment determines the penultimate goals built into the discriminator *D*. The general knowledge of the form of associations inherent in the nature of the environment is built into the adaptive part *Q*. Any foreknowledge which may be available about actual details of particular associations in the environment will be built into the pre-analysis mechanism P.A.

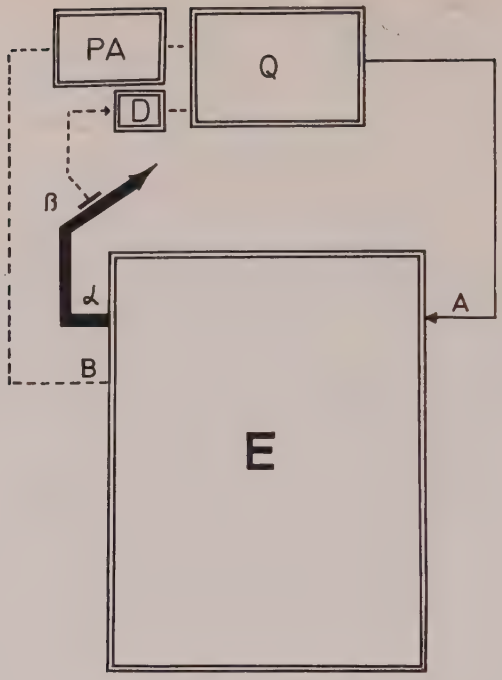


Fig. 7

2. The three Categories of Information Required to Formulate an Action

We have already spoken about the need for a general foreknowledge of the environment and the desirability of incorporating any detailed foreknowledge which may be available even though this could have been acquired by experience. The following table is intended to set this foreknowledge in perspective with the other categories of information with which an intelligent device is concerned, when it formulates an action in response to an event it has observed.

The first category comprises the information that is built into the structure of the device. From the second category the device selects the information which it builds in, during experience, in the form of adaptive modifications of its structure. The third category is information which may not be built in at all, but passes through the device in the

sense that it is immediately responsible for the initiation of an action. In truth, it is the category III information, interacting with the stored category II information, in a way partly determined by category I information, which really determines the action of the device.

Table 1. The Three Categories of Information with which an Intelligent Device is Concerned

I Foreknowledge of
(a) (i) Goals
(ii) Intermediate Goals
(b) General nature of associations.
(c) Detailed information available about associations.
II Experience from which detailed information about the environment may be extracted.
III The immediate observation or sequence of events that initiates an action.

To some extent there is an overlap between categories II and III. A certain item of incoming information may immediately trigger off an action thus falling into category III. At the same time, in responding to the favourable or unfavourable reaction of the environment, the device is adding to its experience and may modify its stored category II information.

At this point it may be appropriate to ask "how fundamental is this division of information into three categories?" There seems little doubt that category I information is of a different class. It is the information that defines the purpose and therefore the structure of the adaptive device. When we come to categories II and III it may be pointed out that when our device "learns" or "adapts" it is merely responding in a predetermined way to a stream of incoming information. If this is so why do we subdivide the incoming information into categories? Why indeed do we need an adaptive device at all, with feedback as its essential feature? Why do we not merely have a simple device with no feedback, designed in the first place to give the required predetermined response?

The answer is that the need for feedback in learning arises from the fact that information about A–α paths has to be extracted from the environment by experiment. This, in itself, does not necessarily predetermine the internal structure of the device except in the sense of determining the general nature of its outward behaviour. It is true that the

device could be built with a predetermined response to any specified input sequence. The argument for designing the device as an adaptive one based on feedback principles is a practical one which may have a major influence on the magnitude and complexity of the physical structure of the device. If the purpose of the device is defined in terms of specified goals then it will generally lead to a simpler device if we construct it to operate on a feedback basis. It can then test whether or not these goals are being attained and adapt accordingly.

The alternative involves a device in which every possible sequence of inputs can be recognised and can initiate an appropriate response. The category I information must of course still be incorporated, and in most cases the result is astronomical complexity. Thus, although nothing we have said so far presupposes the internal structure of the device, the discussion in the next section will be based on the concept of an adaptive response built into the unit Q.

It is when we think of the response as adaptive that it becomes convenient to distinguish between category II and category III information. Category III information is then the observation that produces a response and category II information the experience that causes this response to change with time as the device adapts to its environment. The distinction is not fundamental so much as one of convenience.

THE TIME SCALE AND LIMITATIONS OF HUMAN INTELLIGENCE

So far, nothing has been said about the time scale. Observations have been spoken of as if they were isolated events that could be considered independently. In fact observation is a continuous process, and one cannot define rigidly where an "observation" starts and finishes. The information available about the environment at any instant is the sum of all the information which has come in up to that point. Thus, instead of discussing observations as though they were isolated events we should comprehend in an observation "B" all the relevant incoming information up to the instant we are discussing. For an ideal device which can store and utilise all incoming information, no two such observations will be precisely alike. At the later instant there will be available all the experience of the earlier one, plus whatever has happened between. This means

that when we describe to observations as "similar" we are ignoring differences which we consider to be unimportant. The extent to which we are justified in doing this will depend on the form of the environment and on the general nature of the associations in the environment.

When one reviews some aspects of human interaction in the real world it is evident that on many occasions information from the distant past is of much less significance in determining a choice of action than information from the immediate past. When this is true one may be able to restrict observation to a limited period in the recent past in determining an appropriate reaction to an observation. Two instants then become for practical purposes identical if events during the period of observation immediately leading up to them are identical. These considerations often enable us to base actions on association which only involve relatively recent observations.

We have said that on many occasions observations extending into the distant past are of less importance, but we must stress that there is nothing fundamental about this. Whether the exclusion of more distant past observational data is valid, depends entirely on the structure of the environment. To justify it we must know something about the environment.

For the present, however, we propose to discuss the consequences of the assumption that we are dealing with observations which need not extend very far into the past. We do not exclude the possibility of some overlap in observations but we shall assume that the observation we are dealing with at any particular instant can be confined to a limited period in the immediate past. This can simplify the task to be performed by an intelligent device by a major factor because an observation (one of the "B"s in the diagrams) can be stored in a short term memory of limited capacity.

1. The Analysis of Short Term Observations

Under such circumstances the tasks to be performed by the adaptive part "Q" of the organism are

(i) short term storage of information corresponding to various sequences "B" of incoming information,

(ii) the analysis and identification of these sequences in terms of actions which have been found from past experience to be appropriate.

In Figure 8, the rectangle previously labelled "Q" has now been subdivided into two, to illustrate these two functions, which will be discussed separately. This subdivision has been made to illustrate a logical and

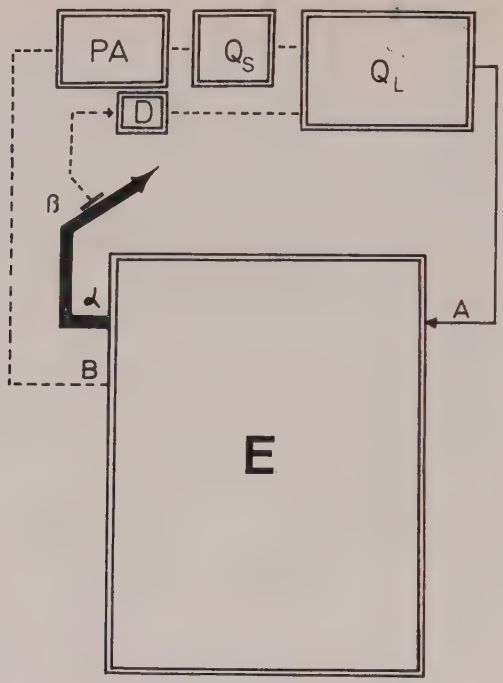


Fig. 8

functional relationship for the purpose of discussion. It does not necessarily follow that these two functions are performed in separate places in the human brain, or that they need be performed separately in any artificial intelligent device.

The function of Q_S

Q_S is the short term memory. Its function is now merely to store all the relevant information corresponding to an observational entity, so that it is available as a whole for analysis by the adaptive unit Q_L.

The function of Q_L

Q_L retains the adaptive function. At any instant it receives from Q_S, information regarding the sequence events in the recent past and treats this as an observation. It has the function of discovering all the useful A–α paths and all the useful associations existing in the environment. It has to discover these in the sense that it has to adapt its reactions in such a way that the actions it causes are of an appropriate nature in

relation to the environment. It does so by interacting with the environment and the adaptive modifications which its structure undergoes are the form in which it records its interaction with the environment. This adaptive unit Q_L is thus by its very nature a long term memory unit as distinct from the short term memory unit of Q_S.

We have spoken of Q_L and Q_S separately, to emphasize a difference in function corresponding roughly to category II and category III information. What limitations are imposed on an adaptive device as a consequence of depending on a short term memory to store category III observational sequences? Evidently the device will be unable to respond to associations involving long sequences. There is at least some evidence of separate long and short term memories in the human brain, and this makes it a matter of some interest to understand these limitations and how they are overcome.

Education provides a simple remedy of overcoming the difficulty by the use of written notes and records. It also provides a less obvious method of overcoming the difficulty. This is to use the vast array of concepts made available to us by the language with which we communicate with each other. The sentence "I had a tooth extracted yesterday" is a simple means of describing quite a complex sequence of events. This simplifies describing the experience to someone else and at the same time provides a very simple form in which to store the facts in ones own memory.

Short term memory restrictions are not the only limitations apparent in the human brain. It also appeared to have difficulty with much above second order association. At least second order association often forms the basis for questions in intelligence tests which are presumably intended to select the more intelligent from the less intelligent. Here again the educational background helps to overcome the limitation too, by providing formalised techniques for the acquisition of information and for computation and manipulating of information which go far beyond simple association.

EDUCATION AND THE ASSIMILATION OF INFORMATION

Here we should like to make a general comment on the importance of tuition and education. Through these two, techniques and information which have been gathered by human society over years, can be made available to an individual who could only have gathered a small fraction

of them in a lifetime. These have been discovered partly by accident and partly by years of painstaking work by many groups of people, and they form part of our cultural heritage. Such information, having in part been gathered by the co-operative efforts of many has done much to by-pass the limitation of the individual human brain.

When it comes to teaching this information to the individual it is carefully converted into a form that will not overtax his mental resources. Painstakingly, over a period of many years he is taught a variety of techniques and methods that have been tried and tested time and time again by others. Eventually these enable him to find solutions to problems of great complexity that he could not possibly have arrived at unaided and by chance. By far the greater part of the skill knowledge of the educated adult has been acquired from this source.

These methods become ingrained in the individual until the way he does anything is largely a function of what he has been taught. Whatever the physical structure of his brain, there is now a logical structure superimposed on it that is to a major extent dictated by the educational process. At this stage, if we look at the methods he uses to arrive at results we shall find that they mainly reflect the structure of the cultural information on which he has come to be so dependent. The characteristics of the elementary brain structure will only show up occasionally when the superimposed logical structure fails to provide an immediate answer.

For the human being the educational and cultural background provides an immensely important addition to the three categories of information already referred to. To include this, category II has been subdivided in Table 2 to include this as category IIb.

Table 2. The Three Categories of Information with which an Adaptive Device is Concerned.

I Foreknowledge of
  (a) (i) Goals
    (ii) Intermediate Goals
  (b) General nature of associations
  (c) Detailed information available about associations.
II Experience from which detailed information about the environment may be extracted. This comprises:
  (a) Information gathered by Direct Experience
  (b) Educational Information.
III The immediate observation or sequence of events that initiates an action.

THE INTELLIGENT DEVICE, A COMPOSITE STRUCTURE

We have tried to stress the importance of educational and cultural background (category II b) information to the human being. This knowledge is just as important in any artificial intelligent device which has to tackle similar tasks. However, unless we want to spend a long period "educating" the device we shall want to build most of this knowledge into the device in the form of category I(c) information.

Without consciously thinking of it in these terms this is what at least some designers of "heuristic" computers appear to have done. By taking a "common sense" attitude to the design of a device they have, in effect embodied in it a considerable amount of their own knowledge and understanding of the problem to be solved, and have thus endowed the device with information which it would have taken a long time to acquire by "adaptive" means.

In view of what has been said, it seems to us that in general neither the wholly adaptive device nor the wholly pre-determined is likely to represent the ideal. One might, perhaps say that the true intelligent device is a judicious mixture of the right amount of adaptive facility, representing category I(a) and (b) information with the right amount of pre-determined structure representing category I(c) information.

The study of the statistical nature of the environment which holds the key to this does not appear to have received the attention it deserves. Particular points which need study are:

(i) The methods of extracting the requisite foreknowledge from the environment.

(ii) The optimum use of this foreknowledge in the basic design of an intelligent device.

(iii) The limitations imposed by the time compression effects of using a short term memory and the effect of restricting association to low orders.

(iv) The detailed design of adaptive parts of an intelligent device.

(v) The contribution of educational and cultural background to the solution of problems and the best way of utilising this information in an artificial intelligent device.

NILS J. NILSSON

Artificial Intelligence Group, Stanford Research Institute, Menlo Park, California.

Adaptive Pattern Recognition: A Survey

THE PATTERN-RECOGNITION PROBLEM

The ability to recognize patterns is an important requirement for intelligent organisms or machines. Selfridge has described pattern recognizers as the "eyes and ears for computers." Pattern-recognition processes have also been applied to such problems as weather forecasting,[1,2] medical diagnoses,[3] and the classification of certain types of electronic waveforms. Indeed, Travis[4] has claimed that "... a bottleneck presently preventing more powerful problem-solving machines is perceptual impotence" indicating that pattern recognition is required for the successful execution of higher mental processes.

In this paper we shall survey some of the research directed toward the development of pattern-recognition techniques. We shall concentrate on those techniques called "adaptive"—that is, those which adjust to the specific problem at hand. In order to consider specific examples of patterns rather than "general patterns," we shall use examples of visual or graphical data because of the primary importance of such data. This choice prevents us from discussing several of the specialized techniques that are being investigated to deal with audio and other inputs.

What is pattern recognition? One definition is that it is a process of *classifying* sensory information into mutually exclusive categories.† Thus a visual pattern recognizer serves the purpose of sorting the scenes it sees into such categories as

(a) The scene is a typewritten letter "A", or

(b) The scene is a typewritten letter "B", etc.

A broader definition of pattern recognition is that it is a process that abstracts sufficient sensory information to answer a question about the

† Selfridge[5] has stated that "··· pattern recognition involves classifying configurations of data into classes of equivalent significance so that very many different configurations all belong in the same equivalence class."

environment.* A visual pattern recognizer then might be expected to be able to answer the following sorts of questions about the scene shown in Figure 1:

Fig. 1. A Visual Scene

(a) How many fence posts are there between the larger house and the tree?
(b) Is there a window in the smaller house?

We shall begin our survey by examining attempts to formalize the pattern-recognition process by using a *classification* model. This model uses the simple *classification* definition of pattern recognition. Many of the concepts derived from the classification model are fundamental to any discussion of the more complex *question-answering* pattern recognizers.

PATTERN RECOGNITION AS A CLASSIFICATION PROBLEM

Many of the simple questions one might ask about a scene† reduce to the following question: to which *one* of several possible categories

* Minsky (Ref. 6, p. 423) states: "What is required is clearly (1) a *list* (of *whatever length is necessary*) of the primitive objects in the scene and (2) a statement about the relations among them."

† We shall use the word "scene" from now on to describe a sensory input field of any modality—visual, acoustic, or other.

does the scene belong? Such an approach has been useful, for example, in character and speech recognition.

In formalizing the process of classification of scenes we must have a way to represent scenes numerically. Suppose that we have methods* for representing any scene by a set of real numbers. We shall call the process of converting a scene into a set of numbers "preprocessing." Let the values of these numbers be denoted by the symbols $x_1, x_2, \ldots, x_d$. We shall call such a set of values a *pattern*. It is convenient to represent a pattern by a point $\boldsymbol{X}$ in d-dimensional space. The coordinates of the point are the values $x_1, x_2, \ldots, x_d$. Alternatively, it will sometimes be convenient to represent the pattern by the vector $\boldsymbol{X}$ with components $x_1, x_2, \ldots, x_d$.

Once a scene has been represented by the pattern $\boldsymbol{X}$, we can speak of the problem of classifying the point $\boldsymbol{X}$ and hence the scene. Suppose the scene can belong to one of R categories. Now if all of the scenes belonging to a single category produced exactly the same set of d numbers, classification would be simple. To each category would correspond a unique set of d numbers or vector $\boldsymbol{X}$. Classification of any scene would then entail only a determination of which one of R unique vectors was identical to the vector representation of the scene.

Unfortunately our preprocessing methods for numerical scene representation are not sophisticated enough to produce the same vector for every instance of a scene of a given category. There is always a *scatter* of vectors in d-dimensional space representative of each class of scene. In many pattern-recognition problems of practical interest, this scatter is quite extensive indeed, and the vectors representing different scene categories may or may not intermingle.

Most attempts to formalize the pattern-classification problem begin by proposing techniques to deal with this scatter of pattern points. One fundamental assumption willingly made by all classification philosophies is that even though the pattern points are scattered, pattern points that are "close" together tend to belong to the same category. This assumption does not necessarily imply, however, that *all* of the patterns belonging to a given category are close together. (The distance by which "closeness" is measured is usually Euclidean distance.) Given this assumption, it appears useful to employ the statistical notion of a probability density

* Some of these are discussed in the section beginning on page 109.

function to describe the scatter of pattern points. Thus let us assume that the patterns belonging to any category, say i, are random variables governed by a probability density function $p(\boldsymbol{X}|i)$.*

Once these probability density functions are known, straightforward statistical analysis can be used to derive "optimum" methods for deciding to which category any given pattern $\boldsymbol{X}$ should be assigned.[7] The optimum method is most often defined as that method which minimizes the probability of classifying a pattern in error. It is easy to show[8] that in this case, the optimum method for classifying a pattern $\boldsymbol{X}$ calls for computing the quantities

$$p(\boldsymbol{X}|i)\, p(i) \quad \text{for} \quad i = 1, \ldots, R$$

and assigning $\boldsymbol{X}$ to that category i_0 corresponding to the largest of these quantities. $p(i)$, $i = 1, \ldots, R$ are the *a priori* probabilities that $\boldsymbol{X}$ belongs to category i.

When the scatter of patterns belonging to a given category is gaussian (an ideal situation unfortunately not often seen in practical classification problems), this rule for optimum classification has appealing simplicity. Let us assume that for each $i = 1, \ldots, R$, $p(\boldsymbol{X}/i)$ is gaussian with covariance matrix $\boldsymbol{K}_i$ and mean vector $\boldsymbol{P}_i$. Then the optimum classifier (minimum probability of error) computes the functions[8]

$$g_i(\boldsymbol{X}) = -\tfrac{1}{2}\boldsymbol{X}^t \boldsymbol{K}_i^{-1}\boldsymbol{X} \quad \text{(second-order terms)}$$

$$+ \boldsymbol{X}^t \boldsymbol{K}_i^{-1}\boldsymbol{P}_i \quad \text{(linear terms)}$$

$$-\tfrac{1}{2}\boldsymbol{P}_i^t \boldsymbol{K}_i^{-1}\boldsymbol{P}_i + \log p(i) - \tfrac{1}{2}\log K_i \quad \text{(constant)}$$

for $i = 1, \ldots, R$, where

$\boldsymbol{K}_i^{-1}$ is the inverse of $\boldsymbol{K}_i$,

$\boldsymbol{X}$ is a column vector,

$\boldsymbol{X}^t$ is the row vector form of $\boldsymbol{X}$, and

$|\boldsymbol{K}_i|$ is the determinant of $\boldsymbol{K}_i$.

The $g_i(\boldsymbol{X})$ in this case are quadratic functions. They contain second-order terms such as $a_{ij}x_i x_j$, linear terms such as $b_i x_i$, and constant terms.

* Read "p of $\boldsymbol{X}$ given i."

The optimum classifier assigns X to that category i_0 corresponding to the largest $g_i(X)$. This decision rule is equivalent to partitioning the pattern space with quadric (second-degree) surfaces. These surfaces separate those patterns X which are assigned to one category from patterns assigned to the other categories.

A simple special case is worth considering. If the *a priori* probabilities are all equal and the covariance matrices are all scalar (a constant times the identity matrix) and equal, then the $g_i(X)$ for the optimum decision method are

$$g_i(X) = X \cdot P_i - \tfrac{1}{2} P_i \cdot P_i$$

where $X \cdot P_i$ is the inner or dot product of X and P_i. In this case the pattern space is partitioned by hyperplanes that bisect the line segments joining pairs of mean vectors. This partitioning seems reasonable in this simple instance in which the scatter of patterns is spherically symmetric and of the same extent for each category. A two-dimensional illustration for a four-category classification problem is shown in Fig. 2.

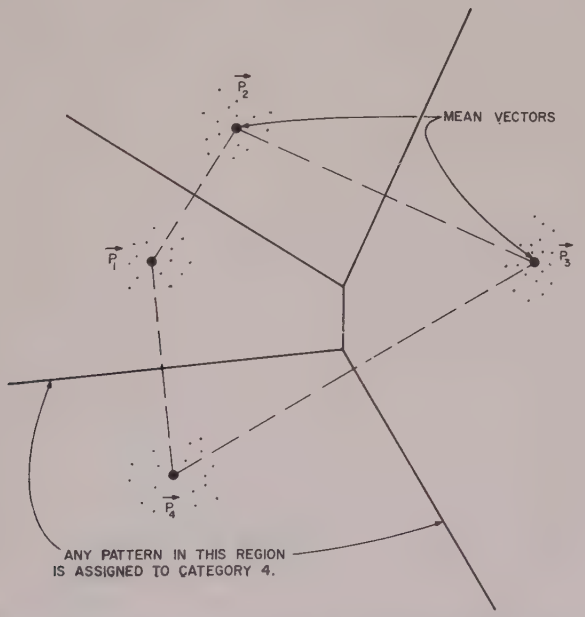


Fig. 2. Four-Category Classification Problem with Spherically Symmetric Gaussian Scatters

If our preprocessing methods for representing scenes numerically produced scatters of known guassian form, our survey of pattern-classification methods would be almost complete. Because scene preprocessing is not yet in such an advanced state, we must explore more powerful classification techniques. Let us discuss some of the inadequacies of the gaussian model just presented and propose some possible remedies.

In the first place, the exact probability densities produced by preprocessors can seldom be deduced *a priori*. We must usually observe a large sample of patterns resulting from many scenes in each category to get an idea of the nature of the scatter. So even if we could know *a priori* that the *form* of the density functions was gaussian, we would probably still have to estimate from pattern samples the mean vector and covariance matrix for each category in any given pattern-recognition problem.* Usually, however, the assumption of a gaussian scatter for each category is quite inappropriate. A more accurate density function would generally have many modes (local maxima) rather than one, and a strange, contorted shape quite different from the familiar gaussian "bell-shaped" distribution. Therefore we need classification techniques which can estimate complex density functions from a large sample of patterns whose categories are known, and then use these densities in the computation of $p(\boldsymbol{X}|i)\,p(i)$ to assign patterns to the appropriate categories. The set of sample patterns on which the density estimates are based is often called the "design set."

A detailed survey of these classification methods begins on page 120 of this paper. It is appropriate to conclude this section with a summary and evaluation of the formalization of the pattern-recognition problem recounted here. The steps of the formalization are as follows:

(a) A scene is represented numerically by a set of numbers $\boldsymbol{X}$. This representation is accomplished by a preprocessor.

(b) On the basis of a large design set of patterns of known classification, probability density functions $p(\boldsymbol{X}|i)$ are estimated to account for the observed scatter.

(c) A pattern $\boldsymbol{X}$ of unknown category can then be classified by assigning it to that category i corresponding to the largest of the quantities $p(\boldsymbol{X}|i)\,p(i)$ for $i = 1, \ldots, R$.

* See Abramson and Braverman[9] and Keehn[10] for discussions of methods for estimating mean vectors and covariance matrices.

This formalization has certain difficulties which should be mentioned. First, it depends on but does not illuminate preprocessing. Excellent preprocessing would result in no pattern scatter and necessitate no elaborate classification method. Poor preprocessing would result in hopelessly scattered and intermingled patterns which the most sophisticated classification techniques could not separate. Unfortunately, the design of the preprocessor is outside of this formalization of the classification problem, and that is a crucial weakness.

Second, this formalization is not broad enough to cover all of the interesting scene-classification methods. For example, the probability-state variable (PSV) approach[11] is excluded, and so are methods such as those used in CYCLOPS,[12] employing inherently sequential testing procedures.

Third, classification of scenes into mutually exclusive categories can be a rather awkward method of answering questions about scenes. To be able to answer a family of questions similar to those on page 104 about the scene in Fig. 1 would require an astronomically large number of mutually exclusive categories and would require an impossibly large number of representative scenes to exemplify all categories. Clearly the estimation of density functions and the principle of classification is ridiculous in such a case. Nevertheless, at some level in the hierarchy of logical processes needed for "question-answering" pattern recognition, a pattern-classification process can be appropriately employed; hence the present survey of contributions that have been made through the study of pattern-classification techniques. Since our formalization excludes the problem of preprocessing, the next section surveys the various *ad hoc* techniques that have been proposed to represent a scene by a set of numbers. Many of these techniques are either motivated by or parallel to biological systems.

PREPROCESSING METHODS

1. Feature Extraction

By "preprocessing" we mean a method of representing a scene by a set of numbers. If the scene is acoustic, one might give it a numerical representation by first converting it to an electrical signal and then finely

sampling this signal to produce the set of numbers. A visual scene might be finely divided into a raster of cells and then the light intensity of each cell converted to an electrical signal. In both cases, an accurate representation of the complete scene could well lead to an unmanageably large set of numbers. Furthermore, the probability-density functions needed to represent the scatter of a pattern set produced by such preprocessing are likely to be inestimably complex. (For example, any one of a large number of small translations of a visual scene over its X, Y frame might produce patterns quite different from the pattern corresponding to the untranslated scene.)

The mass of complex data resulting from such simple preprocessing schemes prompts one to attempt to extract only the significant "features" from the sensory information.* Thus in the recognition of alpha-numeric characters it might be appropriate to detect only the presence or absence of line segments of certain slopes and lengths, intersections, corners, and arcs of certain curvatures. These features would have binary values; 1 for feature present, 0 for feature absent. Other features might have any numerical value indicating the "intensity" of the feature. The collection of feature values for a scene is then used as the numerical representation, X, of the scene. In most pattern-recognition work the researchers identify their feature-extraction method as a major element of the recognition process. Hopefully, the feature extractor will provide some telltale signature for classification of the scene.

Unfortunately there seems to be no general theory to help guide our search for the relevant features in any given recognition problem. The design of feature extractors is largely an empirical matter following different *ad hoc* rules found to be useful in each special situation. One learns little, if anything, about how to design visual feature extractors from a study of successful acoustic feature extractors. This lack of guiding principles makes the study of biological prototypes especially interesting and relevant.

We shall not attempt here to review the relevant physiological research, but we shall survey some important preprocessing techniques that have their parallels in living organisms. Space does not permit a treatment of preprocessing for all sensory modalities, so we will confine our atten-

* For example, Selfridge[5] states: "Pattern recognition is the extraction of the *significant* features from a background of irrelevant detail."

tion to that modality which has received the most attention, namely vision.

2. Scene-Transformation Operations

We shall take as the primary input to a visual preprocessor a rectangular grid of cells each one of which is either black or white.* For our illustrations, it is convenient to represent black cells by the number 1 and white cells by the number 0. The primary object of preprocessing is to reduce this grid of numbers to a manageable set of features relevant to the classification of the scene.

Before extracting features from the grid, it may be expedient to do some preliminary operations on the grid itself. These operations might include such tasks as speck removal, gap filling, thickening, thinning, and edging. Many of these operations are explained in detail in an article by Dineen,[13] and we shall discuss a few of them here.

One simple operation can achieve gap filling, speck removal, thickening, and thinning. This operation we shall call "averaging". In the averaging operation each cell of the grid representation of a scene is made the center

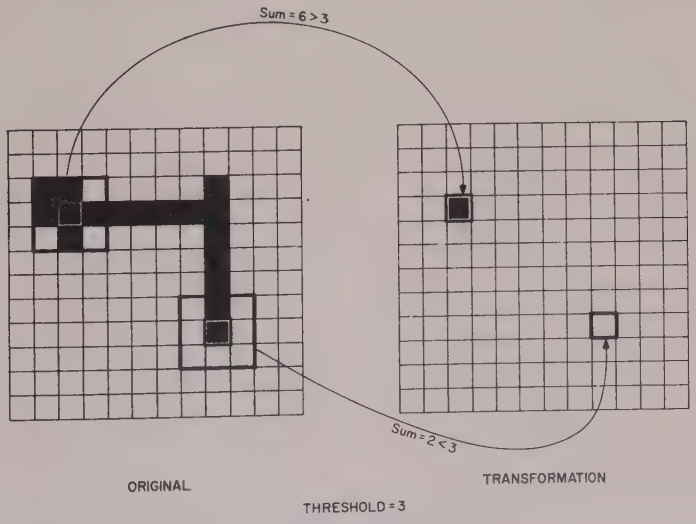
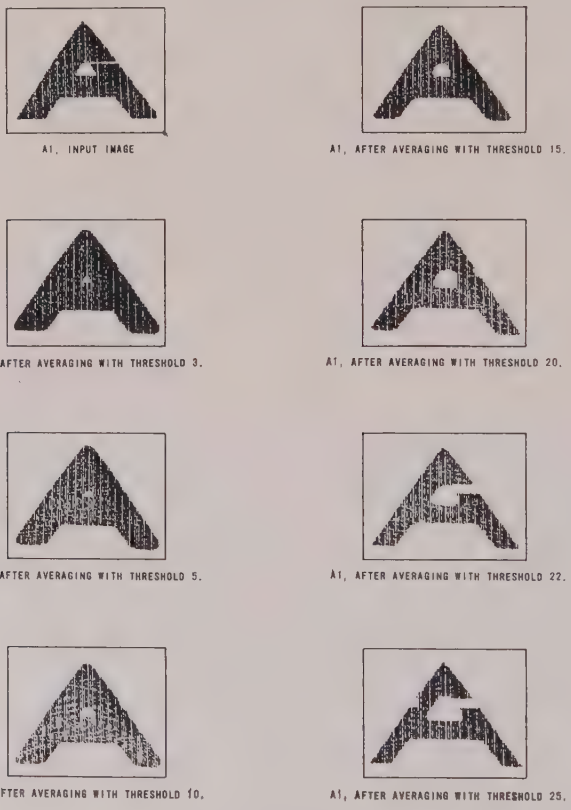


Fig. 3. The Averaging Operation

* Extensions to scenes with gray levels is straightforward.

of a small square window of cells. Two of these windows are shown in Fig. 3. The number of 1's (black cells) within each window are totaled and compared with a threshold, θ. The purpose of these comparisons is to create a new or transformed grid representation of the scene. If the sum of the black cells in the window centered on the (i, j)th cell exceeds the threshold, then the (i, j)th cell in the new matrix is set equal to $+1$. Otherwise it is set equal to 0.

Some examples of this averaging operation, applied to black and white scenes, are shown in Fig. 4 for various thresholds. It can be seen that the averaging operation tends to change any grid element which differs from



By permission, Institute of Electrical and Electronic Engineers, from "Programming Pattern Recognition," by G. P. Dineen, Proceedings Western Joint Computer Conference, 1955.

Fig. 4. Examples of Averaging (90 × 90 image; 5 × 5 window)

a sufficient number of its neighbors. Thus isolated gaps are filled in, and isolated specks are removed. For low thresholds, white lines on a black field are thinned and black lines on a white field are thickened. For high thresholds, black lines on a white field are thinned, and white lines on a black field are thickened. (Note the thickening of the horizontal white line through the right side of the *A* in Fig. 4.) Window size controls the extent of the filling and removal operations. Small windows produce little change in the scene, while larger windows destroy detail. Such operations may be cascaded with different threshold settings and different window sizes for interesting effects. (Note that since the operations are non-linear they are not permutable.)

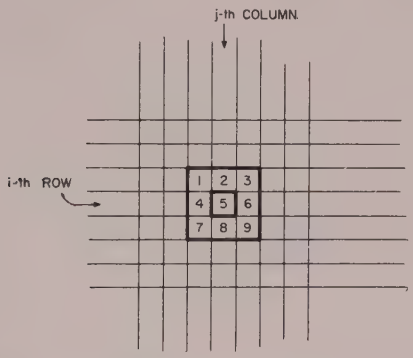


Fig. 5. Window Used for Edging Operation Centered on (i, j)th Cell

Dineen also describes an edging operation. This operation is intended to preserve centers of asymmetry such as edges, corners, junctions, and ends of lines. The edging operation employs a window centered on each black cell (1) of the grid representation of a scene. White cells (0) remain white in the transformed scene. Consider the (i, j)th cell and the window centered on it depicted in Fig. 5. (For simplicity our explanation considers only a 3×3 window. Extensions to larger windows are easily made.) Let the cells in the (i, j)th window be locally numbered 1, 2, ..., 9 as shown. (Window cell 5 is the (i, j)th cell.) Now since we will center the window only on black cells, cell 5 is black. The edging operation proceeds by tallying the sum of the following scores:

(1) Score 1 if cell 1 is black and the diagonally opposite cells – cells 6, 9, and 8— are all white.

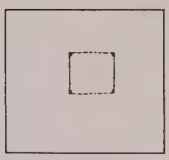
(2) Score 1 if cell 2 is black, and cells 9, 8, and 7 are all white.
(3) Score 1 if cell 3 is black and cells 8, 7, and 4 are all white; and so on around the ring.

If the summed tally is larger than a threshold θ, then the (i, j)th cell in the transformed scene is kept black. Otherwise it is changed to white. The threshold θ is determined by first counting the total number of black cells in the window and then setting the threshold equal to some fraction of this number.

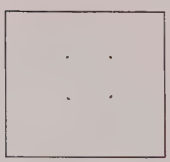
The edging operation sharpens differences; it is like a two-dimensional derivative since we count the changes about the center element. Some examples of the edging operation for various thresholds, expressed as a fraction of the black cells within the window, are shown in Fig. 6.

We shall discuss one more important operation. This one, which we shall call the "isolation" operation[14] selects one connected figure in a scene and erases all other parts of the scene not connected to it. In this operation we assume that a "figure" is made up of 1 or black cells and

SQUARE



EDGING WITH THRESHOLD 4/8.



EDGING WITH THRESHOLD 8/8.

Fig. 6. Example of Edging (90 × 90 image; 7 × 7 window)

the "ground" is made up of 0 or white cells. The isolation operation begins by selecting an arbitrary black cell. A window is centered on this cell and all black cells (including the original one) inside the window are marked "retained." The window is then centered on each "retained" cell in turn and new black cells are marked "retained" by the same criterion. This operation continues until no new "retained" cells are left to become window centers. Then each "retained" cell in the grid has its corresponding cell in the transformed grid set equal to 1. All other cells in the transformed grid are set equal to 0. Thus if the window is 3×3, only the black cells connected (diagonally or adjacently) together and to the original cell are preserved. All other black (figure) cells are converted to white (ground) cells. Larger windows allow the process to "hop across" small gaps in an otherwise connected figure.

The purpose of these preliminary operations, just reviewed, is to "clean up" the scene by removing extraneous information. The dimension of the grid representing the scene has not been changed, however. It is just as large as before. We shall next describe some feature extraction techniques whose intent is to reduce the quantity of numbers needed to describe the scene and to simplify the nature of the probability density functions needed to represent the scatter associated with each category.

3. Features

There are many different varieties of features that one might discuss in connection with visual scenes. Some features, such as "the presence of any edge in the left-hand half of the scene," deal with a specific characteristic that might occur anywhere within a certain area of the scene. If a feature is of this type, we shall call it a "complex" feature. Others, such as "the presence of an edge in the third column of the matrix" deal with a specific characteristic at one specific place. These features we shall call "simple." Some features, such as the two mentioned above, can be defined in terms of more-or-less "local" properties of the scene, while others such as "the average value of all cells in the matrix" are defined only in terms of "global" properties of the scene.

Many local and simple features can be extracted by window techniques similar to those used in the averaging operation. We shall illustrate how

windows can be used to detect an edge. Consider the pair of adjacent windows shown in Figure 7 superimposed on a grid representation of a scene. The scene in this case is a white (0 cells) vertical strip along the left-hand side of the scene with a transition to black (1 cells) between the fourth and fifth columns. The two windows in Figure 7 are labeled

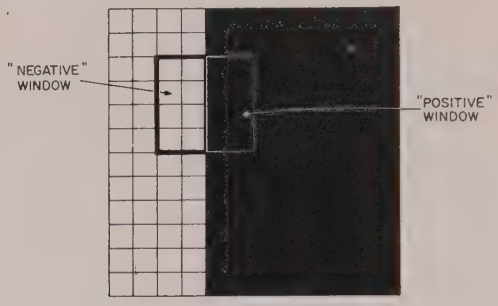


Fig. 7. Example of Simple Edge Detector

"negative" and "positive". The sum of all black cells inside the positive window is added to the negative of the sum of all black cells inside the negative window. If the combined sum exceeds a threshold, θ, we say that the window pair has detected the presence of an edge.

Suppose that in our example of Figure 7, the threshold is 6. The combined sum is 8 so we detect an edge at the window-pair position. Note that this window pair detects only a very specific edge. If the edge in the scene were shifted one column to the right, the combined sum would only be 4, and no edge would be detected by the window pair in its original position. Similarly, if the edge were shifted to the left. If the scene were reversed (0's $\rightarrow$ 1's and 1's $\rightarrow$ 0's), the combined sum would be —8, and no edge would be detected. So our window pair of Fig. 7 detects only white to black (from left to right) edges of a vertical orientation in a specific position.

Other window pairs can be used to detect the presence of other specific features in specific locations (simple features). These window pairs are shown in Figure 8.

Often it is not too important to know more than the general location of a specific feature, but the nature of the window-pair method for feature extraction always gives us a precise location. We can throw away this precise-location information, and thus reduce the dimensionality of the

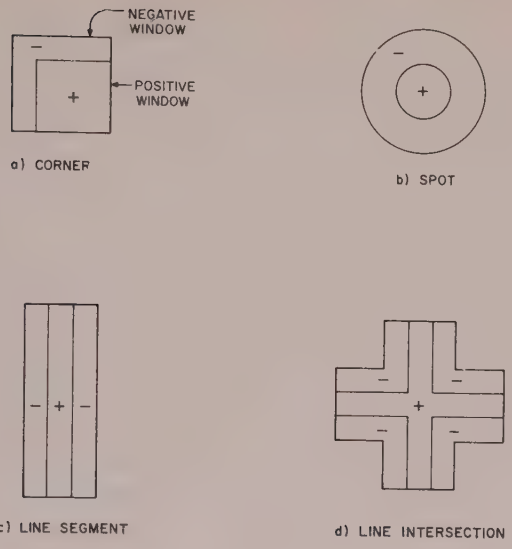


Fig. 8. Window Pairs for Simple Features

numerical representation of a scene, by grouping several simple features together into families. Thus we might group together all corners in the upper half of the scene into one family called, say, "the upper family of corners." Then, if any one member of this family of features is present in the scene, we could say that the family is present in the scene. The presence or absence of the family in the scene is itself a complex feature. Grouping simple features together into complex features is a useful method of reducing the dimensionality of the numerical representation of a scene. But to detect any complex feature requires first that we have the means for detecting any member of its family by, for example, the window-pair method. The detection of complex features may then require a large number of window-pair calculations.

Many of the grid-transformation and feature-detection operations that we have discussed find parallels in the operations performed at the neuron level in living organisms. Bliss and Macurdy[15] discuss some neural models to account for observed grid-transformation phenomena in the sensory mechanisms of animals. A book by Ratliff[16] discusses neural mechanisms that perform edging, averaging and feature detection opera-

tions. Hubel and Wiesel,[17] in a classic set of experiments, discovered (1) neural systems in the lateral geniculate body of the cat that extract "spot" features (see Fig. 8b) and (2) neural systems in the visual cortex of the cat that extract "line segment" (see Fig. 8c) and "edge" features. Furthermore, the simple feature detectors are grouped together into families of translations to detect complex features. Similar sorts of neural mechanisms for the detection of primitive features were found to exist in the eye of the frog. [18]

Other quite different types of features have also been proposed for visual preprocessing. Two interesting features are the total number of intersections of a set of straight lines with figures in the scene and the sum of the lengths of the segments of these lines lying within figures. (Certain theorems of integral geometry have been employed[19,20] to analyze the properties of these features.) Another is the average amount of *detail* in certain subareas of the scene where detail is a measure of the number of changes from white to black or black to white per given length of a row or column of the grid.[21]

Most preprocessing techniques used in successful pattern-classification experiments have used features selected on intuitive grounds as being suited to the particular classification problem at hand. Often experiments were made with a variety of features; the features were then modified or supplemented by new features, more experiments were performed, and in this way the feature complement was upgraded.

Some effort has been devoted to automating this evolutionary process of feature selection. Uhr and Vossler[22] describe a process in which features are first generated (by, for example, a random process), then employed in a classification system and graded on performance, and finally either saved or replaced by new automatically generated features depending on their grade. Kamentsky and Liu[23] have selected sets of features useful for typewritten and printed character recognition by a process which generates a large number of features, computes the "information content" for each of them and selects the most useful of those with the highest information content. Daly[24] *et al.* have described an evolutionary method for feature generation in which large numbers of randomly generated features are used and the best few saved. These are tested alongside another batch of randomly generated features, and the best few are saved, and so on until an adequate number of good features has been determined.

4. Implementation of Feature Detection

Various alternatives exist for implementing in hardware the detection of features such as those we have discussed. Once the scene has been represented by a grid of 1's and 0's and stored in a computer memory, it is a straightforward matter to write computer programs that can calculate whether or not features such as those mentioned above are present in the scene. Sometimes these calculations are quite lengthy, however, and sometimes the scene resolution is such that the number of bits that must be stored in the computer to represent a scene is unmanageably large (1000 lines $\times$ 1000 lines $= 10^6$ bits).

Fortunately, there are many special techniques that can be employed to detect the presence or absence of features in visual scenes without storing the grid representation of the scene in a computer. Kamentsky and Liu[23] have used a shift register loaded from a flying-spot scanner to detect features in a scene. Logic circuits drawing bits from the appropriate cells of the shift register are provided for each *complex* feature being looked for. As the shift register is being loaded by the scanner each logic circuit is sensitive to the precise pattern of bits constituting its feature, regardless of when this pattern of bits occurs and thus regardless of where this feature occurs on the scene. Because the logic circuit reports the presence of a feature if it sees it at any time during a scene loading, it actually reports on the presence or absence of a family of identical but translated features and thus is a complex-feature detector.

Brain[14] and his colleagues have used a parallel optical system which literally makes use of windows to detect the presence of edges in a scene. This system employs an array of about 1000 lenses to form 1000 complete replications of the scene. Each pair of replications of the scene is viewed through a positive and a negative window (transparent rectangular regions in an otherwise opaque mask) by photocells which measure the total light coming through each window. A comparison of the two photocell currents then reveals whether or not an edge existed at this window-pair position. In this way 500 different edges (simple features) can be detected using 500 positive and 500 negative windows. These edges are grouped into 100 edge families. Each edge family contains five edges of identical angular orientation but slightly different translational position. The edge families are detected by ORing together groups of five of the simple edge detectors. Each edge family represents a complex

feature. This set of complex features shows some degree of invariance to small translations of the scene because of the way the families are defined. (Such a technique for obtaining translational invariance was originally proposed by Rosenblatt in his "similarity constrained perceptron." See Chapter 15 of Ref. 25.) It is possible to change the simple features used in this preprocessor from edges to any other simple features desired by simply removing a single photographic plate containing the 1000 rectangular windows (for edges) and substituting a plate with appropriately shaped windows.

Another preprocessing implementation is discussed by Gerdes[21] *et al.*, who are able to detect a large number of interesting features using only simple video circuits. Their equipment includes a flying-spot scanner for reading the scene and special circuits to operate on the video signal from the flying-spot scanner. With these circuits they are able to detect efficiently such features of a scene as brightness variance, detail content, long and short lines and edges, line and edge groupings with specific spacings, isolated objects superimposed on a background of contrasting brightness, and similar features.

The features and implementation methods discussed in this section certainly do not exhaust the subject of preprocessing. Several other techniques, such as two-dimensional optical filtering using Fourier transform methods,[26] exist for visual scenes and, of course, completely different sorts of preprocessing have been employed in acoustical scenes. Our purpose here was to give some examples of features and feature-detection methods and to refer the reader to a few of the specific instances where they have been used. In the next sections of this paper we assume that the features have been selected and concentrate on the problem of how to assign a feature vector, $\boldsymbol{X}$, (and thus a scene) to one of a number of classes.

ADAPTIVE CLASSIFICATION METHODS

1. The Fix and Hodges Method

We shall now assume that the scene to be classified has been represented numerically. In this section we shall be discussing several related classification techniques. These techniques attempt either to estimate directly

the values of probability-density functions from a design set of sample patterns, or to circumvent the explicit estimation of density functions by some roughly equivalent classification procedure.

One attractive method for classifying patterns based on estimates of the values of probability-density functions is the Fix and Hodges method.[27] The Fix and Hodges method uses the patterns in the design set to classify a new pattern, $\boldsymbol{X}$, as follows: select an integer k, and collect the k design patterns closest to $\boldsymbol{X}$ ("closeness" can be measured by Euclidean distance). Suppose that of these k closest patterns, n_1 patterns belong to category 1, n_2 to category 2, ..., and n_R to category R, and $n_1 + n_2 + n_R = k$. Then assign $\boldsymbol{X}$ to that category, i_0, corresponding to the largest of these numbers.

The Fix and Hodges method is clearly an attempt to estimate a set of numbers proportional to $p(\boldsymbol{X}|i)\,p(i)$ for $i = 1, \ldots, R$ around the point $\boldsymbol{X}$. If such a set of numbers is well approximated by the set $n_1, n_2, \ldots, n_R$, then the Fix and Hodges method will lead to nearly optimum (minimum error probability) classifications.

Selection of the integer k is quite important in the application of the Fix and Hodges procedure. If k is too small, the resulting decision rule will be too sensitive to the particular spatial locations of the design patterns. If k is too large, the rule will not be sensitive enough to the actual variations of the unknown probability distributions with $\boldsymbol{X}$. If the design set is large, it has been shown that the Fix and Hodges method leads to the same classifications as would be made if the (unknown) probability distributions were known and used. In general, the value of k should increase without limit with increasing N, if N is the total number of patterns in the design set. The value of k/N, however, should decrease toward zero with increasing N.

2. The Nearest-Neighbor Method

Even when k is small, however, the Fix and Hodges method still leads to classifications comparing favorably with those made by the optimum (minimum probability of error) classifier in the limit of very large design sets. This comparison is still quite favorable even when $k = 1$. When $k = 1$, the Fix and Hodges method places an unknown pattern, $\boldsymbol{X}$, in the same category as that of the closest pattern in the design

set. Such a rule results in what is called the "nearest-neighbor method." The nearest-neighbor method can be shown to be reasonably effective compared with the optimum classifier. Specifically, Cover and Hart[28] have shown that if PE^* is the theoretical minimum probability of error of the optimum classifier using the true (but unknown) probability-density functions, and if PE is the probability of error resulting from the nearest-neighbor method, then PE is bounded by

$$PE^* \leqq PE \leqq PE^*\left(2 - \frac{R}{R-1} PE^*\right)$$

in the limit of an infinitely large design set. Certainly in the limit PE is never greater than twice PE^*. In this sense it may be said that half the classification information in an infinite design set is contained in the nearest neighbor.

The Fix and Hodges and nearest-neighbor methods appear superficially to be reasonable answers to the problem of pattern classification. They suffer one important drawback, however. To classify any pattern, $\boldsymbol{X}$, the distance between $\boldsymbol{X}$ and each of the patterns in the design set must be computed. If these computations are to be performed rapidly, each of the design patterns must be stored in some rapid-access memory. Because the methods work best when the number of design patterns is large, the rapid-access storage requirements are often excessive. This disadvantage has motivated a search for other methods that preserve some of the features of the Fix and Hodges classifier without requiring the individual storage of every design pattern in a rapid-access memory.

3. Nearest-Prototype Methods

Several alternative methods can be discussed at this point. One might decide to use the design patterns differently to estimate the values of the probability-density functions. For example, Sebestyen[29, 30] has proposed a process which synthesizes an estimate of each probability-density function by adding together a number of component gaussian functions. The locations and extents of the component functions are determined from the design patterns by an iterative process which does not require rapid-access storage of all of the design patterns.

Alternatively, one might abandon the idea of obtaining an *explicit* estimate of the values of the probability-density functions. Instead one could attempt to use the design patterns in a way which would lead to the same category decisions for most unknown patterns as those which would be made by the Fix and Hodges or nearest-neighbor rules. Such a procedure might be nearly equivalent to one based on estimating probability-density functions even though it does not explicitly estimate them. These equivalent procedures are considered in more detail in the remainder of this section.

The first of these equivalent procedures can best be explained as a "scaled-down" nearest-neighbor rule. We shall call it the "nearest-prototype method." Rather than storing all of the design patterns, we might store only a few typical patterns called "prototypes." Each prototype selected for storage might actually represent many design patterns that cluster around it in the pattern space. Classification of an unknown pattern $\boldsymbol{X}$ is then accomplished by assigning it to the category of the closest prototype. If there are a sufficient number of prototypes it is reasonable to assume that the prototype closest to $\boldsymbol{X}$ will usually be of the same category as the design pattern closest to $\boldsymbol{X}$. Thus the nearest-prototype method will usually assign patterns to the same category as does the nearest-neighbor method. The problem now is to discover a method for finding a reasonable number of prototypes to represent the design patterns.

One simple method is to provide only one prototype for each category and set that prototype equal to the center of gravity of all of the design patterns in the category. In some classification problems where the scatter of patterns belonging to each category is small, such a single-prototype-per-category method may produce excellent results. For example, in situations where the scenes can be carefully designed to minimize scatter (such as automatic reading of special magnetic characters on bank checks[31]), then one prototype per category may adequately represent all of the patterns belonging to that category. As soon as we encounter situations where a wide variety of scenes might belong to the same category, and our preprocessing methods are not able to respond relatively invariantly to this variety, then it becomes quite important to provide several prototypes per category.

On the other hand, if we are assuming that the design set is large enough to represent truly the inherent scatter in the problem, then the

design patterns could not be so completely intermixed as to require one prototype for each design pattern. If such were the case we certainly could not expect that any other new patterns would be likely to lie close to any of the prototypes, and the category of the nearest prototype would probably be irrelevant anyway. In this case the design set is certainly too small a set to base the design of a classifier on. Thus, in realistic situations, we should be able to find some number of useful prototypes smaller than the number of design patterns and greater than or equal to the number of categories.

Several procedures have been suggested to find adequate prototypes or centers of clusters of patterns. Stark *et al.*,[32] Firschein and Fischler,[33] Ball and Hall,[34] Bonner,[35] and Mattson and Damon[36] are among those that have proposed methods. Ball has reviewed several of these in a survey article.[37] As an example of one of the methods, we shall discuss a promising one proposed by Rosen and Hall.[38]

The Rosen and Hall method begins by arbitrarily selecting one prototype for each pattern category. These prototypes can be conveniently established by selecting one design pattern from each category at random and setting the initial prototypes equal to these patterns. The subsets of design patterns closest to each prototype are then determined. Each pattern in a subset that does not belong to the same category as the prototype is put on an ERROR list. The ERROR list will sometimes have patterns on it belonging to all R categories.

The patterns on the ERROR list are then grouped according to category. If more than N_1 patterns of a given category (say category i) occur on the ERROR list, then a new prototype is established for category i. The new prototype is set equal to that pattern in category i on the ERROR list which is farthest (Euclidian distance) from the original category i prototype. Thus after establishing an ERROR list, we may establish new prototypes equal to some of the patterns on the ERROR list.

We next move each *original* prototype, if it qualifies, to the center of gravity of the subset of patterns closest to it, excluding those patterns on the ERROR list. A prototype does not qualify for motion if the number of patterns in its subset is smaller than N_2. If a prototype does not qualify, it is discarded. None of the prototypes just created from the ERROR list is either moved or discarded during this cycle. The first cycle of the process is now complete.

The second and subsequent cycles of the process are the same as the

first, with the exception that at the beginning of any cycle all the prototypes surviving and all those generated during the previous cycle are regarded as original prototypes. New prototypes are now set up to be those patterns on the ERROR list farthest from any existing like-category prototype. The process either stops naturally when no new prototypes are created and the existing prototypes stabilize, or it can be stopped by the experimenter at his discretion.

The parameters N_1 and N_2 are process parameters that must be set by the experimenter. Typically N_1 should be set to reflect the percentage error realistically expected on the design patterns; its exact value is not too critical. The value of N_2 should depend on the size of the design set and the number of prototypes available. For example, if we had 5000 design patterns, we might reasonably insist that each prototype represent about 100 patterns if we had 50 prototypes available. Perhaps in this case we might let N_2 be somewhat lower than 100, say 50.

In several experiments with real and artificial pattern data,[38] the Rosen and Hall method has located effective prototype patterns. By "effective" we mean that these prototypes could be used to classify new patterns according to the nearest-prototype method, with error rates favorably comparable to any other methods tried.

4. Dot-Product Units

There is yet another related family of classification techniques which employs neither explicit estimates of probability-density functions nor explicitly determined prototype points. Instead it seeks to determine simple hypersurfaces in the space of patterns that adequately separate the design patterns into their proper categories. This family of techniques can best be motivated by an examination of the way in which the distances to prototypes can be calculated and ranked.

Suppose the prototypes are given by the points $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_i, \ldots, \boldsymbol{P}_M$ where there are M prototypes distributed over the R categories of patterns ($M \geqq R$). To classify a new pattern $\boldsymbol{X}$ by the nearest-prototype method, we must calculate the distance to every prototype and find the smallest distance. Finding the smallest squared distance results in the same assignment, so we can instead calculate the quantities

$$|\boldsymbol{X} - \boldsymbol{P}_i|^2 \quad \text{for } i = 1, \ldots, M.$$

These quantities can be written

$$\boldsymbol{X} \cdot \boldsymbol{X} - 2\boldsymbol{X} \cdot \boldsymbol{P}_i + \boldsymbol{P}_i \cdot \boldsymbol{P}_i.$$

Now, to find the smallest of these is equivalent to finding the largest of the following expressions:

$$\boldsymbol{X} \cdot \boldsymbol{P}_i - \tfrac{1}{2}\boldsymbol{P}_i \cdot \boldsymbol{P}_i \quad \text{for} \quad i = 1, \ldots, M,$$

since $\boldsymbol{X} \cdot \boldsymbol{X}$ occurs in all of the expressions and multiplication by $-\frac{1}{2}$ reverses the ordering.

Therefore we can employ very simple calculations in using the nearest-prototype method. The dot product of the pattern vector to be classified with each of the prototypes is computed, and a bias equal to one-half the squared length of each prototype is subtracted from each dot product.

Let us assume that each of these dot product calculations is computed by a device we shall call a "dot product unit" (DPU). The DPU is shown in Figure 9. A DPU may correspond to special-purpose electronic

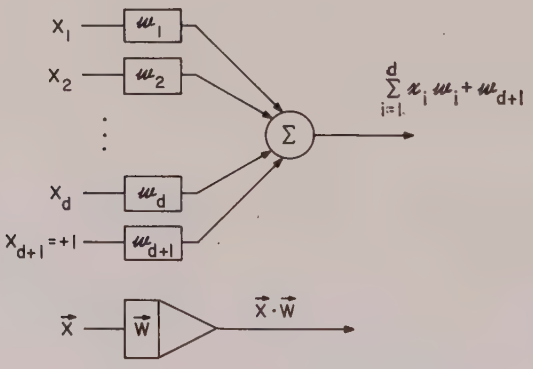


Fig. 9. Dot-Product Unit (DPU)

circuitry (e.g., a resistance adder) or to a digital computer subroutine. For patterns with d components, the DPU has $(d + 1)$ "weights" $w_1, w_2, \ldots, w_d, w_{d+1}$ and computes a weighted sum,

$$S = \sum_{i=1}^{d} w_i x_i + w_{d+1},$$

of the pattern components. The first d weights correspond to the d pattern components, and the extra $(d + 1)$th weight allows a bias term to be added to the sum. It is often convenient to imagine that this $(d + 1)$th weight multiplies a fictitious $(d + 1)$th pattern component which is set permanently equal to some convenient value such as $+1$. If the $(d + 1)$ pattern components are represented by the $(d + 1)$-dimensional vector $\boldsymbol{X}$, and the $(d + 1)$ weights are represented by the $(d + 1)$-dimensional vector $\boldsymbol{W}$, then the output of the DPU can be simply represented by the dot product $\boldsymbol{X} \cdot \boldsymbol{W}$. A classifier which employs DPU's is shown in Figure 10. In general, it may have M DPU'S where $M \geqq R$. For the nearest prototype method a DPU computation is made for every prototype. Each DPU has its first d weights set equal to the components of one of the prototype vectors and its $(d + 1)$th weight set equal to minus one-half the squared length of the prototype vector.

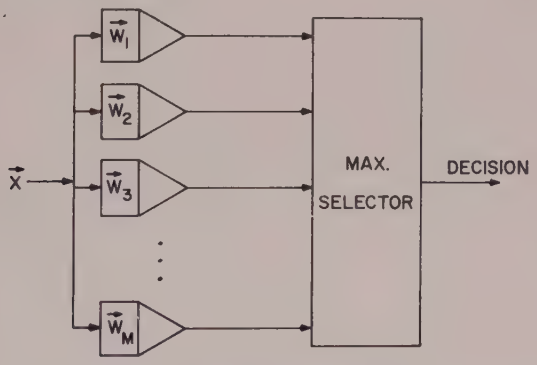


Fig. 10. Classifier Composed of DPU's

The surfaces in the pattern space which separate the patterns assigned to different categories by the nearest-prototype method are composed of segments of hyperplanes, These hyperplane segments are the perpendicular bisectors of line segments connecting prototype points belonging to different categories. Several interesting pattern-classification techniques are based on the principle of finding a good set of hyperplane surfaces while ignoring whether or not these hyperplanes have any relation to imaginary prototype points.

Such surface-oriented methods could employ a structure identical to the one illustrated in Figure 10 using DPU's. For any set of weights

assigned to the DPU's, such a structure implements a certain set of surfaces consisting of segments of hyperplanes. If for each DPU, the $(d + 1)$th weights are set equal to minus one-half the sum of the squares of the first d weights, then we have seen that there exist prototype points such that the surfaces perpendicularly bisect line segments joining them. If this condition on the $(d + 1)$th weights is relaxed in a certain way we might still be able to find prototype points such that the surfaces are perpendicular to but do not bisect line segments joining them. But in general there will not exist any meaningful prototype points for arbitrary $(d + 1)$th weights, and our attention must focus solely on the properties of the resulting hypersurfaces.

We shall call such a machine with DPU's having arbitrary weights a "piecewise linear" (PWL) machine[8] since the surface that separates any pair of categories consists of segments of linear (hyperplane) surfaces. Several of the famous adaptive pattern classifiers including the ADALINE, MADALINE,[39] MINOS,[14] and PERCEPTRON[25] networks are special instances of PWL machines.

The designer of a PWL pattern classifying machine must have methods for setting the weights on each DPU and for deciding how many DPU's to assign to each category.* These decisions should be based on a large design set of patterns. Some successful, simple methods for setting the weights exist, but as yet there are no simple methods for deciding how to assign DPU's to the various categories. The weight-setting methods that we shall discuss here involve the following steps:

(1) Selection of arbitrary initial weights of a PWL machine
(2) Trial tests of the ability of this PWL machine to classify patterns in the design set correctly
(3) Adjustment of the weights in response to errors made by the PWL machine on the design patterns.

For obvious reasons, methods employing these steps are called error-correction training methods. We shall first discuss a training method for an interesting special case of the PWL machine in which there is just one DPU provided for each pattern category. Such a machine we shall call a "linear machine."[8]

* Of course he could first find some prototype points and then set the weights so that the PWL machine implements a nearest-prototype classification method. But here we wish to discuss simpler techniques for weight setting that do not depend on explicit reference to prototypes.

5. Linear Machines

The linear machine is probably the most common of all pattern-classification devices. When its weights are set to implement a nearest-prototype classification method, there can be but one prototype per category. Even so it has been usefully employed in many recognition devices. When its weights are allowed to have arbitrary values its range of usefulness is probably even greater. Highleyman[40] has discussed methods for setting the weights, as have Griffin, King, and Tunis.[41] Steinbuch and Piske[42] have employed essentially the same structure, calling it a "learning matrix."

We shall discuss a particularly simple training method for the linear machine. First the weights of the linear machine are set at arbitrary initial values. A pattern is then selected from the design set and tested on this initial version of the linear machine. If it is correctly classified, another pattern is selected and tested. When a pattern, say $\boldsymbol{X}$, is not correctly classified we shall adjust the weights of *two* of the DPU's. Suppose the $(d + 1)$-dimensional weight vectors prior to adjustment are denoted by $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_R$. Let the integer i be the *actual* category of the erroneously classified pattern, and let the integer $j \neq i$ be the category decision of the machine. Then only the weight vectors $\boldsymbol{W}_i$, and $\boldsymbol{W}_j$ are modified. Let their new values be given by the symbols $\boldsymbol{W}_i'$ and $\boldsymbol{W}_j'$. The prescribed modifications in this case are

$$\boldsymbol{W}_i' = \boldsymbol{W}_i + c\boldsymbol{X}$$

and

$$\boldsymbol{W}_j' = \boldsymbol{W}_j - c\boldsymbol{X}$$

where c is an arbitrary constant held fixed* at the same value during training.

This rule is a straightforward attempt to correct the error in classification of $\boldsymbol{X}$. Obviously $\boldsymbol{W}_i \cdot \boldsymbol{X}$ was smaller than $\boldsymbol{W}_j \cdot \boldsymbol{X}$ when it was required that $\boldsymbol{W}_i \cdot \boldsymbol{X}$ be the largest output over all DPU's. Addition of $c\boldsymbol{X}$ to $\boldsymbol{W}_i$ will certainly increase the output of the ith DPU, and subtraction of $c\boldsymbol{X}$ from $\boldsymbol{W}_j$ will decrease the output of the jth DPU. These ad-

* Actually it is possible to relax the requirement that c be fixed (see Ref. 43). In some experiments better results are obtained if c is decreased slowly during training.

justments may or may not completely correct the error (depending on the value of c relative to the magnitude of the difference between $\boldsymbol{W}_i \cdot \boldsymbol{X}$ and $\boldsymbol{W}_j \cdot \boldsymbol{X}$), but nevertheless they are steps in the right direction.

After a weight-vector adjustment, training continues by testing the linear machine on the design patterns, one at a time. The design patterns can be selected for test in fixed order, cycling through the set over and over, or in any random order that ensures that each pattern will be tested an infinite number of times if training continues for an infinite length of time.

Suppose it is known that there exists some setting of the weights of the linear machine that will correctly classify all of the design patterns. Then the application of this error-correction training rule will, after only a finite number of corrections, produce a linear machine which does indeed classify all of the design patterns correctly. This result and its proof was first stated by Kessler.[8] A subsequent proof by Duda and Fossum[43] covers a somewhat more general version of the rule.

This rule has been applied successfully in many experiments. Of these we might mention the ones conducted by Duda and Fossum[43] and the one reported by Casey *et al.*[44]

6. The Threshold Logic Unit (TLU)

A further interesting specialization of the PWL machine can be made. Consider the case of a linear machine for $R = 2$. Then there are only two DPU's. These compute the quantities

$$S_1 = \boldsymbol{W}_1 \cdot \boldsymbol{X}$$

and

$$S_2 = \boldsymbol{W}_2 \cdot \boldsymbol{X}$$

The linear machine in this case must only decide which is larger, S_1 or S_2. Such a simple comparison can also be made by testing to see if $(S_1 - S_2)$ is positive or negative. Let S denote the difference $(S_1 - S_2)$. S can be computed as follows

$$S = (\boldsymbol{W}_1 - \boldsymbol{W}_2) \cdot \boldsymbol{X}$$

or setting $\boldsymbol{W} = (\boldsymbol{W}_1 - \boldsymbol{W}_2)$

$$S = \boldsymbol{W} \cdot \boldsymbol{X}.$$

Therefore for the $R = 2$ linear machine a single DPU will suffice. This DPU can be followed by a "threshold element" to determine if $\boldsymbol{W} \cdot \boldsymbol{X}$ is positive or negative. Such a structure, called a "threshold logic unit" (TLU) is illustrated in Figure 11. The TLU is identical with the

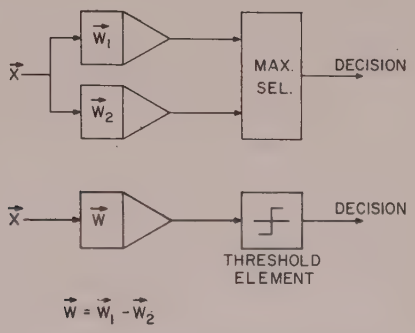


Fig. 11. Threshold Logic Unit (TLU)

pattern-classifying device called ADALINE by Widrow.[39] It is also identical to the final decision element of the simple α-perceptron of Rosenblatt.[25] It divides the space of patterns into two classes by a single hyperplane. When the two-category classification problem is simple enough for such a surface to suffice, the TLU has been employed with excellent results.

The error correction training rule for linear machines has a simple form for the TLU. If $\boldsymbol{X}$ is classified in error, one of two adjustments is made to the weight vector $\boldsymbol{W}$ at that stage. If $\boldsymbol{W} \cdot \boldsymbol{X}$ was erroneously negative (a category 1 pattern was assigned by the TLU to category 2), then the new weight vector is given by

$$\boldsymbol{W}' = \boldsymbol{W} + c\boldsymbol{X}$$

If $\boldsymbol{W} \cdot \boldsymbol{X}$ was erroneously positive (a category 2 pattern was assigned by the TLU to category 1), then the new weight vector is given by

$$\boldsymbol{W}' = \boldsymbol{W} - c\boldsymbol{X}$$

This rule and the proof of its convergence was first stated by Rosenblatt.[25] Its convergence was later also proved by Ridgway[45] and Novikoff.[46] While we have presented it as a special case of the general rule for linear

machines the specific case (as usual) preceded the general by a few years. Actually, an adjustment technique very similar to the TLU rule was proposed by Agmon,[47] and Motzkin and Schoenberg[48] for the iterative solution of a set of linear inequalities several years earlier. The Agmon-Motzkin-Schoenberg rule turned out to be nearly identical to a training process for the TLU proposed later by Widrow and Hoff[49] called a "minimum mean square error" (MMSE) rule. Recently Koford and Groner[50] have shown that to the extent that the MMSE rule accomplishes its objective, the weight vector produced is identical to that prescribed for the optimum classifier when the patterns obey gaussian density functions with the same covariance matrix.

The TLU has served as the basis for another interesting two-category classifier, called the MADALINE or committee machine.[8, 39, 51] A committee machine is a classifier that bases its classification on the majority vote of an odd number of TLU's. Each TLU responds with a $+1$ or -1 to any pattern, according to whether its dot product is positive or negative respectively, and a "vote-taker" sums the outputs of the TLU's. This sum must be either positive or negative, since there are an odd number of TLU's. If the sum is positive, the committee machine responds with an output of $+1$; if the sum is negative, the committee machine responds with an output of -1. These two outputs are then associated with the two pattern categories. The committee machine is illustrated in Figure 12. This organization has been used in at least two pattern-recognition devices.[39, 14]

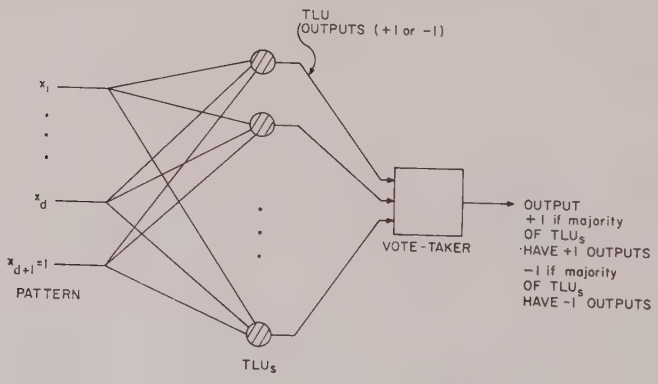


Fig. 12. Committee Machine

Straightforward extensions of the error-correction rule can be made for training of the committee machine.[8] Unfortunately, none of these rules have the provable convergence properties associated with the error-correction rule for a linear machine. Preliminary experiments[14] indicate that the committee-machine method of employing DPU's does not seem to perform classification quite as well as the PWL-machine method using the same number of DPU's. This conclusion must still be regarded as tentative, however, since it is based on limited experience.

It is at this point, while attention is focused on the two-category case and on the TLU, that our subject of pattern classification comes in closest contact with the discipline of switching theory. This contact is made when we consider especially pattern vectors with binary components. Switching theorists have long discussed the problem of which switching functions are linearly separable (see Singleton[52]) and how to find TLU weights for linearly separable functions. They have also considered the problem of how to implement switching functions that are not linearly separable out of networks of cascaded TLU's.[53] Some switching theorists[54, 55] have explicitly discussed the application of their techniques to pattern-classification problems. (The article by Winder[54] contains an extensive bibliography.) The author is not aware, however, of any networks designed by analytical switching-theory techniques whose performance on difficult pattern-recognition problems competes with that of the adaptive classifiers discussed in this paper.

While we are on the subject of two-category classifiers and the TLU, let us make one more digression before returning to the mainstream of our discussion, which is PWL machines and how to train them.

7. Parallel Organizations of Dichotomizers

The classifying systems discussed so far for problems with more than two categories have all involved comparing a set of numbers, such as dot products or distances. An alternative approach to the design of pattern-classifying systems for more than two categories ($R > 2$) has employed a parallel organization of two-category classifiers (dichotomizers). (See, for example, Mattson.[56]) The dichotomizers may be of any type; for example, they may be PWL machines, simple TLU's, or committee machines. A classifier employing some number, say q, of

independent dichotomizers can classify patterns into as many as 2^q different categories since there are 2^q different possible output combinations. This type of dichotomizer uses a code converter to implement a coding scheme whereby the R pattern categories are associated with the 2^q possible dichotomizer output combinations. Such a classifier is illustrated in Figure 13.

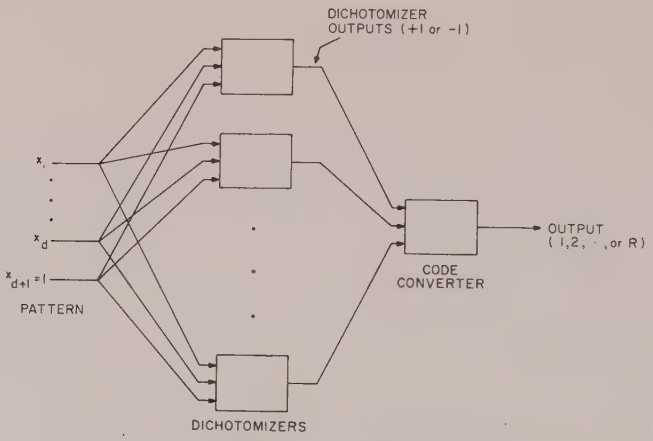


Fig. 13. R-Category Classifier Using Parallel Dichotomizers

It has been found useful to use a value of q such that 2^q is much larger than the number of categories R. Such a redundant use of dichotomizers permits the use of error-correcting codes. For example, suppose $q = 9$ and $R = 32$. There exists a single-error-correction nine-bit code with five information bits. For each of the 32 code words belonging to this code, there corresponds one of the pattern categories. This correspondence defines the desired outputs for each of the nine dichotomizers. When a pattern is presented to this machine the response of five of the dichotomizers can be used to specify the 32 categories after any single errors have been corrected in the complete nine-bit output.

Another type of code employs as many dichotomizers as there are categories ($q = R$). Each dichotomizer is a specialist. The ith dichotomizer then would respond with a $+1$ to all patterns in category i and with a -1 to all other patterns. Here we must interpret illegal response combinations (all R dichotomizers responding with a -1, or more than one dichotomizer responding with a $+1$) as a reject or confused decision.

One of the outstanding research problems connected with the parallel-dichotomizer organization is the question of which coding scheme to use. Very little is yet known, for example, about the trade-off between redundancy in the number of dichotomizers and accuracy of classification. Insufficient experience with the parallel organizations of dichotomizers precludes commenting on their effectiveness. In one experiment,[14] a parallel organization of committee machines as specialists was compared with a PWL machine having fewer DPU's. The PWL machine performed better. In any case the coding problem faced in the use of parallel organizations of dichotomizers is avoided in the PWL machine, and that itself is an important factor.

8. Piecewise-Linear Machines

Let us return now to our discussion of PWL machines and of simple methods for training them. We have seen that in the special case of a linear machine an effective error-correction training rule exists. A similar procedure has been suggested by Nilsson[8] for training a PWL machine. This is an error-correction method which adjusts weight vectors but leaves their assignments to given categories fixed. In applications of this method, then, some arbitrary number of weight vectors must be assigned to each category.* Design patterns are presented to the PWL machine one at a time for trial. After presenting a pattern that the machine classifies correctly, we make no changes in any of the weight vectors. Suppose, however, that a pattern $\boldsymbol{X}$ belonging to category i causes an incorrect response. Such would be the case if the DPU with the largest output were assigned to some category $j \neq i$. The adjustment method first subtracts $c\boldsymbol{X}$ from the weight vector used by this DPU with the largest output. Of those DPU's assigned to category i, we next determine which of this subset has the largest output for $\boldsymbol{X}$. The corresponding weight vector is adjusted by the addition of $c\boldsymbol{X}$. (The factor c is again a constant.)

* No one has yet proposed effective error-correction methods for changing the category assignment of weight vectors during training. Such a rule would be helpful to make efficient use of a limited number of weight vectors by allowing transferral from categories having a surplus assigned to them to categories having a shortage.

It is known that the mere existence of a solution (i.e., the existence of a PWL machine that correctly classifies all of the design patterns) is not sufficient to guarantee convergence using this rule. Cyclic presentation of patterns, for example, can lead to cyclic variations in the weights. Nevertheless, the rule has been applied successfully in a large number of experiments. It has been used by Duda and Fossum[43] on some artificially generated patterns, and by Brain *et al*[14] on some preprocessed, hand-printed characters. In the experiments performed to date it has always done better than committee-machine organizations, but it was outperformed by the Rosen-Hall prototype-generating methods[38] on the artificially generated patterns used by Duda and Fossum. Perhaps this latter difference in performance can be partially explained by the fact that these artificially generated patterns consisted of true clusters of patterns—a perfect problem for the Rosen-Hall method.

The rule appears to become progressively less effective as the number of weight vectors assigned to each category is increased. It appears that more than three weight vectors per category cannot be effectively trained using the version of the rule just described.

An important aspect of all error-correction rules, whether or not convergence on properly separable design sets can be proved, is their behavior when the design set cannot possibly be 100 per cent correctly classified by the machine being trained. Obviously, if we assume that some errors must always exist, these error-correction training rules never stabilize since they always prescribe change in response to error. It has been found empirically that these methods still work quite well for theoretical minimum error rates on the design set of 10 per cent or less.

Some insensitivity to the disruptive effects of continuing adjustments can be purchased at the expense of larger weight values by the use of a "margin" requirement. By a margin requirement we mean that an error in training occurs if a DPU required to have the largest output does not in fact have an output exceeding that of any DPU assigned to another category by some fixed amount, called the "margin". Thus, under some circumstances the machine is adjusted even when an appropriate DPU has the largest output. The use of the margin requirement leads to more adjustments at the beginning of the training process but in fewer adjustments after training has progressed to some extent (when adjustments might be most disruptive).

The idea of a margin can also be employed after the machine has been

trained and is being used to classify new patterns. A margin can be set such that patterns which fail to satisfy the margin requirement are simply *rejected* rather than classified. This strategem usually results in a substantial reduction of the percentage error made on the classified patterns at the expense of rejecting a small percentage of patterns whose classification is uncertain.

Duda and Singleton have suggested another technique[57, 58] to smooth out the variations in performance of a classifier caused by weight adjustments during training. This technique involves keeping track of the average of each weight vector during training. When training is halted the average weight vectors are substituted for their erratic counterparts, usually with greatly enhanced results in classifying new patterns. The technique is particularly effective on design sets for which high error rates must be tolerated or for design sets for which the same pattern may occur repeatedly as an example of several categories rather than just one.

So far we have discussed training methods that depend on a design set of patterns whose categories are known. Some researchers have considered the problem of dividing a set of uncategorized patterns into their "natural" categories. This is a taxonomic problem and can only be discussed meaningfully when some *a priori* criterion for class membership is imposed. For example, we may insist that each of the invented categories contain only patterns which are close (Euclidean distance) together. In this case the problem becomes identical with finding "clusters" of patterns. Such a process has often been called "learning without a teacher."

The concept of learning without a teacher has been employed in an interesting manner by Nagy and Shelton.[39] They consider the problem of making adjustments in a partially trained classifier while that classifier is attempting to categorize new patterns whose actual categories are never revealed. Their experiments have been conducted with a nearest-prototype classifier with a possibility of several prototypes per category.* The prototypes have been selected so that the classifier performs very well on alphanumeric characters of one set of fonts. They then test the classifier on characters belonging to a different set of fonts and performance deteriorates as expected, but not so much that no credibility can

* The prototypes in each category all lie on a subcube in the original pattern space. Such a configuration essentially allows for "don't care" pattern components.

be attached to the decisions of the classifier. Since these decisions are often (though not always) correct, Nagy and Shelton reason that they can be used to adjust the prototypes used by the classifier and thus hopefully improve its performance by a bootstrapping process.

Their adjustment process is as follows: If a pattern is assigned to category i (that is, the pattern is closest to the ith prototype), then the ith prototype is moved a small distance toward the pattern. In this way the prototype eventually becomes more nearly matched to the new fonts than it was previously. This technique substantially improved the performance on characters belonging to the new font set. It seems that it would also be useful to "track" slowly changing patterns. Of course, important questions must be asked about the stability of the process and about how accurately the classifier must be able to classify new patterns before bootstrapping can be effective.

9. Polynomial Machines

We have discussed in this section a wide variety of classification techniques. They have all used hyperplanes or surfaces composed of sections of hyperplanes as boundary surfaces to separate patterns belonging to different categories. Techniques using more complex surfaces, such as surfaces described by polynomial equations, have also been employed in pattern-classification research. Indeed, second-degree or quadric surfaces are optimum for general gaussian probability-density functions.) It has been shown[8, 60] that the use of polynomial surfaces can be considered theoretically as a simple extension of the use of linear surfaces. Error-correction training techniques can be used with "polynomial machines" with the result that in a finite number of steps the machine will make no errors on the training set if the set is suitably separable by a polynomial surface within the power of the machine to implement.[61] To use polynomial surfaces, however, one must calculate a large number of product terms such as $x_i x_j x_k$. To determine which technique is preferable in any given situation, the expense of making these computations as a step toward more complex surfaces must be compared with the expense of adding a few more DPU's to a PWL machine.

10. Experimental Results

Evidence is accumulating that the nearest-prototype methods and other manifestations of PWL machines represent an economical and generally adequate solution to the pattern-classification problem. They have been employed in a variety of experiments with excellent results. Casey *et al.*[44] report fewer than 20 errors in more than 6000 typewritten alphanumeric characters distributed over nine fonts. Brain and Munson[14] have achieved under 1 per cent error on complex hand-printed symbols. Gerdes *et al.*[21] have been able to classify sections of complex aerial photographs into different categories (such as urban areas, tank farms, airfields, bridges, industry, etc.) with accuracies of 90 percent and better.* Moreover, it appears that this quality of performance is not going to be substantially improved solely by using more sophisticated classifiers. Whatever performance gap exists between present abilities and future desires will probably have to be closed mainly by improved preprocessing. But now we are getting into the subject of the next section which will discuss some of the outstanding unsolved problems. So, at the risk of closing the book prematurely on the subject of pattern classification, let us broaden our perspective by briefly considering some larger problems.

SOME OUTSTANDING PROBLEMS IN PATTERN RECOGNITION

That aspect of pattern recognition which involves assigning a scene to one of many mutually exclusive categories has been divided into two parts: preprocessing and classification. So far, research results on classification techniques have outstripped those on preprocessing; it would be appropriate now to restore the balance. If it turns out that preprocessing research, unlike classification research, cannot be guided by general principles, then improved preprocessing techniques will come only after long and arduous empirical investigations. Such investigations might begin by settling on a classification technique, such as the nearest-neighbor rule, and then gradually collecting sets of features that work

* Gerdes' classifier uses essentially a nearest-prototype method, except that distances to each prototype are calculated using a metric peculiar to the prototype. This simple modification results in piecewise quadric surfaces rather than piecewise linear surfaces. See also Sebestyen.[29]

well for each problem type. Perhaps automated evolutionary methods, such as those of Uhr and Vossler,[22] can be used to mechanize the emergence of the best specific features within larger sets selected on intuitive grounds.

One possibility, of course, is to shrink the role of the preprocessor and expand that of the classifier. The scenes with which this giant classifier must deal must now be represented numerically (without elaborate preprocessing) in such precise fashion that no information about the category of the scene is lost in the translation to numerical form. In many instances this will mean (for visual scenes) sampling over a very fine grid. Now, however, nearest-prototype classification methods correspond to a simple type of template-matching. For many problems, the number of templates needed might be astronomical. Furthermore, the number of training patterns needed to determine the best prototypes (templates) will be even larger. Clearly, for most problems, the preprocessor must be retained and the problem of its design must be faced squarely if not altogether confidently. Perhaps we can hope to find some more clues to the riddle of preprocessor design by studying successful biological pattern recognizers.

Now let us grant for a moment that soon preprocessors and classifiers will be working harmoniously and effectively together. Are such systems for scene classification relevant at all to the larger problems of pattern recognition involving scene analysis, or will entirely new techniques be needed? We recall that by scene analysis we mean the ability to answer specific questions about a scene. We assume that the number of questions that might be asked, and the variety of possible answers, are so large as to rule out solving the problem by a simple classification of the scene into one of a great many categories. Instead, scene analysis will probably require a hierarchy of logical operations that may call upon pattern-classification subroutines at various stages.

Consider, for example, one of the questions asked in connection with Figure 1: How many fence posts are there between the larger house and the tree? We will assume here that such a question is stated in a form "understandable" by a computer system. (This assumption might require either that the computer be able to translate "natural language" into some unambiguous internal representation or that the question be phrased in some formal language.) A hypothetical system for answering such a question might operate as follows:

(a) Since the words "larger house" are used, the computer would generate the subtask of locating the two houses in the scene. This

task of locating an object of a certain category might be accomplished by a technique to be discussed in a moment.

(b) Since the words "the tree" are used, the computer would generate the subtask of locating the tree.

(c) The area between the now-located larger house and the tree would next be searched to locate objects classified as fence posts and a simple counting operation employed to answer the original question.

At least for this question, we see that a crucial operation is that of locating an object of a given category. Since many questions that might be asked of a scene involve this basic operation, let us see how we might implement it.

One straightforward method is to scan the scene with a moving window or frame of adjustable size. The scan can be specified so that the frame searches for an object, centers on it, and then expands or contracts so that the object just fills the frame. Then we can classify the subscene within the frame into one of many categories corresponding to the kinds of objects that are possible. The classification of the subscene can be accomplished by methods such as those discussed in the last section. Any preprocessing used by the classifier is understood to operate only on that portion of the scene currently within the movable frame.

One difficulty is presented by the possibility of having several fragments of objects within the frame during an attempt to classify the subscene. Consider first the simple situation in which none of the objects in the scene touch each other or overlap. If each of the basic objects recognizable by the classifier is connected (not disjoint), then the simple isolation operation described above in the section on preprocessing can be used to eliminate all but one object in the subscene.

The steps of the process for this simple case, then, are the following:

(a) Locate an object by scanning the frame

(b) Adjust the frame position so that the contents of the frame are centered within the frame

(c) Isolate the object by deleting fragments in the frame not connected to the object just located

(d) Repeat Steps (b) and (c) until no further centering or isolation is needed

(e) Expand or contract the frame until the object just fills it

(f) Repeat Steps (b) thru (e) until stable

(g) Classify subscene
(h) Resume scanning.

Such a sequence of operations (except for frame expansion and contraction) have been implemented at SRI[14] as an example of how a master computer program using simple pattern-classifying subroutines might be constructed to search for objects in complex scenes.

Scenes with overlapping (instead of isolated) objects present greater difficulty. Single examples of scenes with overlapping objects have been analyzed using the so-called CYCLOPS[12] pattern-recognition process. More complex scenes with overlapping objects have not yet been successfully dealt with, and present a challenging problem for future research.

Let us now summarize some of the main points of this paper. We decided that scene classification (classifying a scene into one of many categories) is a useful model for many (but not all) of the tasks that we normally associate with pattern recognition. Scene classification in turn can be broken down into two components: preprocessing and classification. While several useful preprocessing techniques are known (at least for visual scenes), there exists no theory guiding the synthesis of preprocessors. Several of the visual-preprocessing methods such as averaging, edging, and simple-feature detection find parallels in biological systems, and it is hoped that future biological research may provide further clues. In contrast to the preprocessing situation, useful conceptual guidelines do exist for designing classifiers, and we discussed some adaptive methods involving the use of prototype patterns or piecewise-linear machines. We concluded that the adaptive-classification problem seems to be well understood, and adequate methods for classifying patterns are available. Further improvements in performance will probably have to be the result of advances in preprocessing.

Finally, we discussed how preprocessors and classifiers might be used as subroutines in a higher program directing a search for objects in a scene. Much research remains to be done to design these higher programs and to incorporate them into scene-analysis programs capable of performing pattern recognition in its broadest form.

Even after satisfactory scene analysis can be performed on scenes composed of simple non-overlapping objects, much further research will be needed before machines will be able to deal effectively with the complex "real" scenes of the world such as landscapes and cityscapes. Rosenblatt[62] has given some attention to the types of picture processing

needed to deal with such real scenes. He suggests that generalized boundaries (such as rapid transitions in texture, color, depth, brightness, etc.) will be important features for a real-scene preprocessor. To deal with these scenes effectively, however, it appears that simple classification techniques must be merged with other aspects of artificial-intelligence research, such as optimum search procedures, question-answering systems, and representational methods. Perhaps at this interface lie some of the most important research problems in pattern recognition.

REFERENCES

1. Duda, R. O., and J. W. Machanik (1963). "An Adaptive Prediction Technique and its Application to Weather Forecasting," *1963 WESCON* (August).
2. Hu, M. C. J. (1964). "Application of the Adaline System to Weather Forecasting," *Stanford Electronics Labs Report 64-066*, TR 6775-1, Stanford University, Stanford, Calif. (June).
3. Okajima, M., *et al.* (1963). "Computer Pattern Recognition Techniques: Some Results with Real Electroa(rdiographic Data," *IEEE Trans. on Bio-Med. Electronics*, Vol. BME-10, No.3 (July).
4. Travis, L. (1964). "Experiments with a Theorem-Utilizing Program," *AFIPS Conference Proceedings*, 1964 Spring Joint Computer Conference, Vol. 25, Spartan Books, Washington, D. C.
5. Selfridge, O. G. (1955). "Pattern Recognition and Modern Computers," *Proc. 1955 Western Joint Computer Conference*, pp. 91–93 (March).
6. Minsky, M. (1961). "Steps Toward Artificial Intelligence," *Proc. IRE*, Vol. 49, No. 1, p. 14 (January). Also in *Computers and Thought*, Feigenbaum and Feldman (eds.), McGraw-Hill Book Co., Inc., New York. (1963).
7. Chow, C. K. (1957). "An Optimum Character Recognition System Using Decision Functions," *IRE Trans. on Electronic Computers*, Vol. EC-6, pp. 247–253 (December)
8. Nilsson, N. J. (1965). *Learning Machines: Foundations of Trainable Pattern Classifying Systems*, McGraw-Hill Book Co., New York.
9. Abramson, N., and D. Braverman 1962). "Learning to Recognize Patterns in a Random Environment," *IRE Trans. on Info. Theory*, Vol. IT-8, No. 5, pp.558–563 (September).
10. Keehn, D. G. (1965). "A Note on Learning for Gaussian Properties," *IEEE Trans. on Info. Theory*, Vol. IT-11, No. 1, pp. 126–132 (January).
11. Lee, R. J., *et al.* (1963). "Theory of Probability State Variable Systems," Adaptronics, Inc. *Tech. Documentary Report No. ASD-TDR-63-664*, Vol. I (December) prepared under contract AF 33(657)-7100.
12. Marill, T., *et al.* (1963). "CYCLOPS-1: A Second-Generation Recognition System," *Proc. Fall Joint Computer Conference.*
13. Dineen, G. P. (1955). "Programming Pattern Recognition," *Proc. 1955 Western Joint Computer Conference*, pp. 94–100 (March).

14. Brain, A. E., and J. H. Munson (1966). "Graphical Data Processing Research Study and Experimental Investigation," Stanford Research Institute Report No. 22, Final Report (April) prepared for U. S. Army Electronics Command under contract DA 36-039 AMC-03247 (E).
15. Bliss, J. C., and W. B. Macurdy (1961). "Linear Models for Contrast Phenomena," *Journal of the Optical Society of America*, Vol. 51, No. 12, pp. 1373–1379 (December).
16. Ratliff, F. (1965). *Mach Bands: Quantitative Studies on Neural Networks in the Retina*, Holden-Day, Inc., San Francisco.
17. Hubel, D. H., and T. N. Wiesel (1962). "Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex," Journal of Physiology, Vol. 160, pp. 106–123. Also (1966) in *Pattern Recognition*, L. Uhr (ed.), Wiley, New York.
18. Lettvin, J. Y., *et al.* (1959). "What the Frog's Eye Tells the Frog's Brain," Proc. of the IRE, pp. 1940–1951 (November).
19. Novikoff, A. B. J. (1962). "Integral Geometry as a Tool in Pattern Recognition," Trans. of the Illinois Symposium, *Principles of Self-Organization*, pp. 347–368, Von Foerster and Zopf (eds.), Pergamon Press, New York.
20. Ball, G. H. (1962). "An Application of Integral Geometry to Pattern Recognition," Stanford Research Institute Final Report (February) prepared for Office of Naval Research, Washington, D. C. under contract Nonr 3438(00).
21. Gerdes, J. W., *et al.* (1966). "Automatic Target Recognition Device (ATRD)," Rome Air Development Center, Griffiss Air Force Base, New York, *Tech. Report No. RADC-TR-65-438* (January).
22. Uhr, L., and C. Vossler. "A Pattern Recognition Program that Generates, Evaluates, and Adjusts its Own Operators," *1961 Proc. of the Western Joint Computer Conference*, pp. 555–569. Also (1963) in *Computers and Thought*, Feigenbaum and Feldman (eds.), McGraw-Hill Book Co., Inc., New York; and (1966) *Pattern Recognition*, Uhr (ed.), John Wiley & Sons, Inc., New York.
23. Kamentsky, L. A., and C. N. Liu (1963). "Computer-Automated Design of Multifont Print Recognition Logic, "*IBM Journal of Research and Development*, Vol. 7, No. 1, pp. 2–13 (January).
24. Daly, J. A., R. D. Joseph and D. M. Ramsey (1963). "An Iterative Design Technique for Pattern Classification Logic," *1963 WESCON Paper No. 1.3* (August).
25. Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, D. C.
26. Van der Lugt, A. (1964). "Signal Detection by Complex Spatial Filtering," *IEEE Trans. on Info. Theory*, Vol. IT-10, No. 2, pp. 139–145 (April).
27. Fix, E., and J. L. Hodges, Jr. (1951). "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," University of California Report 4 (February) on *Project* 21-49-004, prepared under contract AF 41(128)-37 for USAF School of Aviation Medicine, Randolph Field, Texas. (This report may be obtained through the Defense Document center for Scientific and Technical Information, Cameron Station, Alexandria, Va.; refer to ATI 110633.)
28. Cover, T. M., and P. E. Hart (1967). "Nearest Neighbor Pattern Classification," *IEEE Trans. on Info. Theory*, Vol. IT-13, No. 1, pp. 21–27 (January).

29. Sebestyen, G. (1962). "Pattern Recognition by an Adaptive Process of Sample Set Construction," *IRE Trans. on Info. Theory*, Vol. 178, No. 5, pp. 582–591 (September).
30. Sebestyen, G. (1962). *Decision-Making Processes in Pattern Recognition*, The Macmillan Co., New York.
31. Eldredge, K. R., *et al.* (1956). "Automatic Input for Business Data Processing Systems", *Proc. Eastern Joint Computer Conference*, pp. 69–73 (December).
32. Stark, L. M., *et al.* (1962). "Computer Pattern Recognition Techniques: Electrocardiographic Diagnosis," *Comm. of the ACM*, Vol. 5, pp. 527–532 (October).
33. Firschein, O., and M. Fischler (1963). "Automatic Subclass Determination for Pattern-Recognition Applications," *IEEE Trans. on Electronic Computers*, Vol. EC-12, No. 2, pp. 137–141 (April).
34. Ball, G. H., and D. J. Hall (1966). "ISODATA, An Iterative Method of Multi-Variate Data Analysis and Pattern Classification," *Proc. of International Communications Conference*, Philadelphia (June).
35. Bonner, R. E. (1964). "On Some Clustering Techniques," *IBM J. of Research and Development* (January).
36. Mattson, R. L., and J. E. Damman (1965). "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems," *IBM J. of Research and Development*, pp. 294–302 (July).
37. Ball, G. H. (1965). "Data Analysis in the Social Sciences: What About the Details?" *Proc. Fall Point Computer Conference*, Vol. 27, Part I, Spartan Books, Washington, D. C., pp. 533–559.
38. Rosen, C. A., and D. J. Hall (1966). "A Pattern Recognition Experiment with Near-Optimum Results," *IEEE Trans. on Electronic Computers*, Vol. EC-15, No. 4 pp. 666–667 (January).
39. Widrow, B. (1962). "Generalization and Information Storage in Networks of Adaline 'Neurons'," *Self-Organizing System-1962*, pp. 435–461, Yovits, Jacobi and Goldstein (eds.), Spartan Books, Washington, D. C.
40. Highleyman, W. H. (1962). "Linear Decision Functions, with Application to Pattern Recognition," *Proc. IRE*, Vol. 50, No. 6, pp. 1501–1514 (June).
41. Griffin, J., J. King and C. Tunis (1963). "A Pattern Identification System Using Linear Decision Functions," *IBM Systems Journal* (September-December).
42. Steinbuch, K., and V. A. W. Piske (1963). "Learning Matrices and Their Applications," *IEEE Trans. on Electronic Computers*, Vol. EC-12, No. 5, pp. 846–862 (December).
43. Duda, R. O., and H. Fossum (1966). "Pattern Classification by Iteratively Determined Linear and Piecewise Linear Discriminant Functions," *IEEE Trans. on Electronic Computers*, Vol. EC-15, No. 2, pp. 220–232 (April).
44. Casey, R. G., (ed.) (1965). "An Experimental Comparison of Several Design Algorithms Used in Pattern Recognition," *IBM Research Report No. RC* 1500 (November 10).
45. Ridgway, W. C. (1962). "An Adaptive Logic System with Generalizing Properties," Stanford Electronis Labs Technical Report 1556-1 (April) prepared under Air Force Contract AF 33(616)-7726.

46. Novikoff, A. B. J. (1963) "On Convergence Proofs for Perceptrons," *Symposium on Math. Theory of Automata*, Brooklyn, New York, pp. 615–622, Polytechnic Press.
47. Agmon, S. (1954). "The Relaxation Method for Linear Inequalities," *Canadian J. of Math.*, Vol. 6, No. 3, pp. 382–392.
48. Motzkin, T. S., and I. J. Schoenberg (1954). "The Relaxation Method for Linear Inequalities," *Canadian J. of Math.*, Vol. 6, No. 3, pp. 393–404.
49. Widrow, B., and M. E. Hoff (1960). "Adaptive Switching Circuits," Stanford Electronics Labs Technical Report 1553-1 (June 30).
50. Koford, J. E., and G. F. Groner (1966). "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier," *IEEE Trans. on Info. Theory*, Vol. IT-12, No. 1, pp. 42–50 (January).
51. Ablow, C. M., and D. J. Kaylor (1965). "A Committee Solution of the Pattern Recognition Problem," *IEEE Trans. on Info. Theory*, Vol. IT-11, No. 3, pp. 453–455 (July).
52. Singleton, R. C. (1962). "A Test for Linear Separability as Applied to Self-Organizing Machines," *Self-Organizing Systems-1962*, pp. 503–524, Yovits, Jacobi and Goldstein (eds.), Spartan Books, Washington, D. C.
53. Minnick, R. C. (1961). "Linear-Input Logic," *IRE Trans. on Electronic Computers*, Vol. 10, No. 1, pp. 6–16 (March).
54. Winder, R. O. (1963). "Threshold Logic in Artificial Intelligence," *IEEE* Publication S-142 entitled Artificial Intelligence, pp. 107–128 (January).
55. Akers, S. B., Jr., and B. H. Rutter (1963). "The Use of Threshold Logic in Pattern Recognition," *1963 WESCON Paper No. 1.2* (August).
56. Mattson, R. L. (1960). "An Approach to Pattern Recognition Using Linear Threshold Devices," *Lockheed Missiles and Space Division, LMSD-*702680, Sunnyvale, Calif. (September).
57. Duda, R. O., and R. C. Singleton (1964). "Training a Threshold Logic Unit with Imperfectly Classified Patterns," *1964 WESCON*, Los Angeles, Calif. (August 25–28).
58. Duda, R. O. (1966). "Training a Linear Machine on Mislabeled Patterns," Stanford Research Institute Technical Report (May) prepared under cotract Nonr 3438(00).
59. Nagy, G., and G. L. Shelton, Jr. (1966). "Self-Corrective Character Recognition System," *Proc. 1966 International Symposium on Info. Theory*, Los Angeles, Calif. (January).
60. Cover, T. (1965). "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Trans. on Electronic Computers*, Vol. EC-14, No. 3, pp. 326–334 (June).
61. Koford, J. (1962). "Adaptive Network Organization," *Stanford Electronics Labs Quarterly Research Review* **3**, p. III-6.
62. Rosenblatt, F. (1965). Personal Communication (June).

ROGER L. BARRON

Adaptronics, Inc.,

McLean, Virginia

*Self-organizing and Learning Control Systems**

INTRODUCTION AND HISTORICAL DATA

The areas of control and pattern recognition stand out as today's most prominent lines of endeavor in applications of Bionics concepts. These are, of course, cognate subjects: the learning control system deals with patterns in signals from its sensory devices; the pattern recognition system is trained by means of a parameter search algorithm closely resembling adaptive feedback control processes. But important differences exist: pattern recognition systems usually deal with a multitude of sensory signals; by comparison, learning control systems generally work with a limited number of input signals (or use pattern processing to transform many inputs into a few). Furthermore, pattern recognition systems are usually trained off-line; the learning controller learns while doing and operates under ground rules which have to impose much larger penalties for mistakes.

This paper discusses the state of the art in learning control systems and sets forth some of the basic principles which are emerging from analyses and syntheses of these systems. Several recent applications are described. Finally, present learning control system development trends are extrapolated in attempting a five-year forecast.

It is appropriate that an inventory of progress in learning control systems applications be made at this time. Ten years ago, rapid strides in aeronautical design, flight propulsion, and flight vehicle structures seriously threatened to outpace development of techniques for augmentation of pilot control capabilities.[1] The Air Force Flight Dynamics

* Support of Avionics and Flight Dynamics Laboratories, RTD, AFSC, Wright-Patterson AFB, Ohio, and Goddard Space Flight Center, NASA, Greenbelt, Maryland is gratefully acknowledged.

Laboratory, in response to this challenge, initiated development of adaptive control logics that exhibited low sensitivity to changes in control-loop parameters and which reduced the requirement for *a priori* knowledge of cause and effect relationships. Credit is due former Captain R. Rath, former Lieutenant P. C. Gregory, M. A. Ostgaard, and others at Wright Field for this achievement: as Truxal[2] has noted, the appearance of adaptive techniques resulted in significant advances in a technology which (in 1956) had found itself on a plateau in the development of new theory.

Several contributions of considerable long-range importance appeared at about the time the above turning point in flight control arrived. Box[3, 4, 5, 6] and others established the method of steepest descent as a useful tool for experimental exploration of performance response surfaces. Brooks [7, 8] introduced the vital concept of random search. Draper, Li, and H. Laning, Jr.[9] conceived automatic optimalizing control. Ashby[10] formulated a mathematical treatment for self-organization in living systems. To him we also owe the important Law of Requisite Variety. Von Neumann[11] placed the subject of probabilistic logics on firm mathematical ground, drawing partly on the classical work of McCulloch and Pitts.[12] Lee[13] devised a probabilistic neuron model having application to control system networks.

By 1959, the situation was appropriate for creation of the bionics discipline, which would act both as a catalyst and as a source of coherence in advancing the different aspects of physical learning systems, including self-organizing and learning control systems. In these efforts, C. W. Gwinn, former Lt. Colonel L. M. Butsch, and D. R. Moore of the Air Force Avionics Laboratory are particularly deserving of record for their contributions to Engineering Bionics, and attention is also due the close technical association which exists between this group and M. A. Ostgaard, Captain R. P. Johannes, P. E. Blatt, and others of the Air Force Flight Dynamics Laboratory.

Numerous contributions to the learning control systems field date from the beginning of the Bionics program. Sklansky[14] has documented much of the work published through 1964, listing a bibliography of over 130 items. (Incidentally, Sklansky discusses learning control systems from the point of view of pattern recognition.) In retrospect, however, there appears to have been a tendency, particularly until about 1964, for investigators to cling to deterministic approaches in the synthesis of

learning control devices. This has been very evident (and costly) in the area of parameter space search, the *sine qua non* of mechanistic learning. This author believes the Bionics program has made a very important contribution in stimulating the application of random search strategies, which circumvent many of the limitations of conventional "hill climbing" techniques.[15]

Also, many investigators of learning control systems have held that control system adaptation must begin with explicit identifications of plant parameters, with the numerical results of these identifications used in pre-structured control laws. If one makes this assumption, the "learning loop" is relegated to a role of "trainer" for a "system identifier" in the "adaptive loop."[14] The author believes confusion on this point has resulted from the prevalent concept that learning actions in control systems must, necessarily, proceed slowly and occur as a consequence of repeated input stimuli. It is certainly true that on-line identifications must be made, but of system performance rather than plant characteristics. And if the performance evaluation process is made sufficiently rapid, as it often can be, there is little reason why accomplishment of substantial learning should require longer than small fractions of the plant response time. Although preflight training has a place in the "system identifier" approach, the author believes that we should be careful not to overstate the requirements for such training.

PROPERTIES AND DEFINITIONS

Learning control systems are adaptive systems which act on-line to improve their performance using, if possible, no other information than that contained in histories of controller input, output, and environmental feedback signals, plus a specified criterion for the evaluation of performance trends. Self-organizing control systems are learning control systems in which internal re-structuring, as in, for example, the variation of probability distributions, takes place as the controller acquires information. In popular usage, self-organizing controllers are also often taken to be those learning controllers using small memory retention intervals (of the order of 0.1 per cent or less of the plant response time).

Adaptive systems theory, which underlies self-organizing and learning control, has been approached in many ways. For example, the theory

has been variously equated to discriminant analysis and to statistical decision theory. Since living adaptive systems often involve networks of nerve cells, it is sometimes assumed that a network of artificial neurons having properties resembling to some extent those of living neurons, is an adaptive system. Also, that which is called automata theory is essentially the theory of a certain class of formal languages. Although all of these various approaches have some utility in the analysis of adaptive systems, none is particularly well-suited for control applications. A better approach for our purposes is to define the necessary elements of adaptive systems in general and to indicate the design choices for these elements, i.e., to take a synthetic approach, rather than an analytic approach.

Parenthetically we should note that terminology for adaptive systems is diverse and often confusing. At times various authors will use the terms adaptive systems, self-organizing systems, automata, cognitive systems, intelligent machines, Bionic systems, and learning systems as essentially equivalent. Other authors have drawn fine distinctions among several of these terms, depending on their purposes. In this paper we assume that all these terms are more or less interchangeable, but indicate some of the distinctions that have evolved in control applications.

There are two essential properties a system must have to be classified as adaptive[16]:

1. The system must be tightly coupled to an environment, so tightly coupled that the system behavior is almost meaningless if not referred to environmental behavior or conditions.
2. The system behavior must be a function of the prior behavior of both the system and the environment.

It can be noted that no linearity assumption is made for adaptive systems; in general, adaptive systems are strongly nonlinear and some classes of adaptive systems can be devised that "deliberately" experiment with the environment.

From the two defining properties of adaptive systems, some of the features of such systems can be distinguished and labeled. The coupling of the system is generally abstracted for control applications as a pair of variables, input and output, although other methods of abstraction and representation are possible. The internal condition of the system is abstracted as the internal state variable of the system. In these terms,

we can write two equations which express the key properties of adaptive systems:

$$y = f(x, s) \tag{1}$$

$$s_{k+1} = g(x_k, s_k) \tag{2}$$

where

x = system imput vector
y = system output vector
s = system internal state vector
subscript "k" refers to the kth instant of time
subscript "$k + 1$" refers to the $(k + 1)$st instant of time

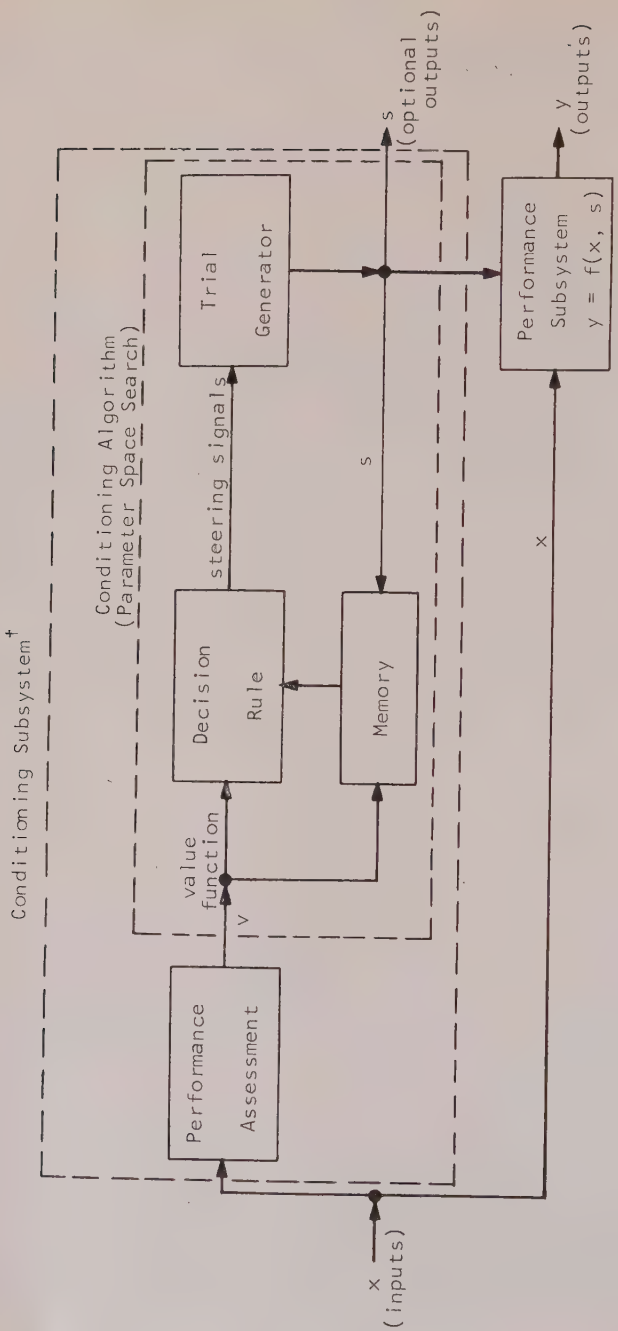
Equation 2 refers to systems in which the internal state changes at discrete times. It is not necessary for adaptive systems to change internal state at discrete times, but some cases are more conveniently discussed in discrete terms.

A mechanization of Equation 1 is termed a *performance subsystem*, as shown in Figure 1. Performance subsystems fall into one or more of five basic categories[17], consisting of subsystems that:

1. Provide algebraic transformations of input signals.
2. Provide some type of compensation or phase-gain adjustment on input signals.
3. Provide some type of discrete or Boolean functions on the input signals.
4. Process digitally coded signals using arithmetic registers and either fixed or stored programs.
5. Generate signals coded in analog or digital form.

For purposes of control, one can usually neglect categories 1 and 3, above. Of the remaining categories, the most important are the subsystems that generate signals. This may appear anomalous in view of the fact that controllers are usually designed to provide compensation of some type, rather than to generate signals. It is a consequence of the adaptive systems approach, however.

A mechanization of Equation 2 is termed a *conditioning subsystem* (see figure). It is clear that conditioning must always be such that system behavior is brought into accordance with a preassigned or desired criterion of performance. The *conditioning algorithm* is a parameter space search procedure that is used for changing the internal state of the adaptive system. More broadly, the entire process of selecting a change



†Conditioning Subsystem implements learning strategy, $s_{k+1} = g(x_k, s_k)$

Fig. 1. General Configuration of Adaptive System

of system internal state or probability of internal state is often called a *learning strategy*. Adaptive systems in which the "next state" equation (Equation 2) is a deterministic function of present state and input are called *deterministic state adaptive systems*, while those in which the probability of the next state is adjusted on the basis of current input and probability are termed *probability state variable* (PSV) systems. PSV systems have characteristics which make them particularly well suited for on-line control system applications (see discussion in sections which follow).

The preassigned *performance criterion* in the conditioning subsytem may be abstract or even ill-defined (such as "surval") or may be more precisely defined, as in the case of control, where the criteron may be expressed in terms of stability or in terms of always driving the error to zero as rapidly as is possible within the requirement for stability. Many criteria of system performance are possible.[18] All must be stated unambiguously so that the *performance assessment* functions within the conditioning subsystem can be implemented. Statements of criteria often take the form of a *score function*, which is a continuously computed index of the closeness of actual performance to desired performance or of trends in the performance. In general, score functions involve input and output variables, and it is sometimes appropriate to have a score function that is the difference between a performance function that measures the quality of the performance and a penalty function that measures the cost of obtaining improved performance.

Considerable latitude exists in the selection of performance criteria. For example, it is possible to add such constraints on controller performance as minimization of expended resources or minimization of load factors (in addition to primary constraints such as stability). These constraints are sometimes added via a modification of the input error function(s), retaining the same basic means of performance assessment.

At this point we should distinguish again between two varieties of adaptive systems: those with long-term memory and those for which memory retention intervals are on the order of 0.1 per cent of the plant time constant. The Air Force has used the term *self-organizing* to apply to those systems with short-term memory and *learning* to apply to systems with long-term memory. On the basis of this usage, a "learning" system takes a comparatively long time to adjust to new conditions, whereas a "self-organizing system" is one which adjusts to new conditions within

a small fraction of the plant time constant. In "learning" control systems, the primary emphasis is on use of memory functions. In "self-organizing" control systems greater attention is given to statistical decision theory (what is the true performance of the plant at this instant?), prediction theory (what is the performance trend?), and rapid trial generation (what must be done to improve the performance trend?).

Before taking up details of recent applications work, we shall discuss further the central topics of parameter space search and performance assessment.

PARAMETER SPACE SEARCH

Control system conditioning algorithms are autonomous parameter space search processes involved in system adaptations of their internal states or internal state probabilities. These strategies are members of the larger family of system optimization techniques[15] and belong to that branch of optimization which deals with problems having incompletely defined environment mathematical models.

In some adaptive and learning control systems[14], the purpose of parameter search is to identify (and/or up-date and refine the estimation processes used to identify) approximate numerical values of plant parameters, subject to certain assumptions about uniqueness of the identifications. This time-honored approach permits use of the slower search algorithms by placing reliance on pre-structured control laws to maintain stable response until adaptation has occurred. In comparatively recent times, the advent of random search techniques has sparked a revolution in learning system design theory[7, 8, 19, 20, 21, 22, 23]. With random search techniques it is possible to do without explicit identification of plant parameters (see discussion of a prototype self-organizing controller elsewhere in this paper). When working with high-speed random techniques, design emphasis shifts to realization of means for accurate real-time assessment of the system over-all score function. In other words, one identifies the control system performance in lieu of explicit determination of plant parameters. The random search processes are used in conjunction with a continuously measured score, and the outputs of the parameter search usually become the inputs to the plant.

Brooks[7] has established that the probability that at least one experiment in a sequence of k random trials will be in an acceptable (best) fraction f of an experimental region in n-space is

$$p(f) = 1 - (1 - f)^k \tag{3}$$

Clearly, as $k \to \infty$, $p(f) \to 1$. Furthermore, to the extent that f is independent of the dimensionality, n, $p(f)$ approaches unity at a rate which is independent of the dimensionality. Of course, from intuition, we expect some connection between n and f. Rastrigin[20] has shown that the number of steps (analogous to k in Equation 3) that is required for convergence on a hypothetical linear response surface is proportional to n for a representative gradient method and proportional to only $n^{\frac{1}{2}}$ for the random search method (considering steps of fixed size with uniform distribution of random step directions in n-space). Rastrigin shows that the random search procedure has the advantage for $n = 12$ and all larger values of n. (Rastrigin did not consider techniques which can accelerate the random search.[7, 15, 19, 21, 22, 23, 24])

Regarding the complexity of systems performing systematic searches (as opposed to searches by the random method), Bellman[25] has noted that memory requirements increase as the square of the dimensionality ("The curse of dimensionality"—Bellman.[25]) By way of contrast, memory requirements for random searches increase roughly in proportion to the first power of n.

Capabilities for multidimensional search are important, but (as noted in the Introduction) learning control systems usually deal with relatively few parameters. This brings us to one of the critical considerations in learning systems theory: modality of performance response surfaces. From Equation 3, we see that $p(f)$ is formally independent of surface modality. It follows that the basic random search is equally effective on unimodal and multimodal performance response surfaces.[15, 19] In general, deterministic search strategies cannot handle multimodal surfaces except by somehow isolating all of the peaks so that each may be explored individually. The complexities of these isolation processes have thus far ruled out steepest descent (gradient) methods for on-line applications where the response surface is multimodal.

In time-varying situations, there may be rapid variations in the response surface being searched. This imposes a severe requirement on the speed

of search convergence. Several methods (described later) have been devised for accelerating random searches so that search goals are attained very rapidly. The idea is to accomplish accelerated searching without sacrificing too much of the generality of random techniques. In particular, excessive degradation of the multimodal capability of random searches must be avoided in accelerating the search, since this is required in a great many problems.

The basic random search has uniform probability distributions associated with each of the parameters in the search. Each new experiment is selected completely at random from the set of all possible experiments. Suppose, however, we give the successive trials nonuniform probability distributions. Further, let us control the migration of the "center point" of these distributions. Matyas[21] has formulated a mathematical description for accelerated random searches which use such techniques. His formulation employs an n-dimensional normal random vector, having variable mean value and variable correlation matrix, which is added to the state vector. The state vector, in turn, migrates in accordance with a test procedure which establishes if a new trial is an improvement over all preceding trials ("success") or is inferior to the best trial thus far conducted ("failure").

Matyas considers the adaptation of the random vector mean value and correlation matrix.[21] He shows that adaptation of the mean value increases the probability of a successful step and also serves as a control over step size. Furthermore, he shows that adaptation of the correlation matrix (which Matyas treats in diagonal form) corresponds to control over the standard deviations of all the components of the random vector and hence the degree of approximation of a random trial to the ideal trial. Finally, Matyas notes that his "adaptive random optimization method" is "easily realized on a computer"and has a "simpler logic" than "other methods for the optimization of several parameters."[21]

In this country, Lee and Snyder[26] and Barron *et al.*[27, 28, 29, 30] have synthesized simple PSV self-organizing control systems involving adaptation of the probability distribution of random step directions in the multidimensional control vector space, u_i ($i = 1, 2, \ldots, n$). Lee and Snyder dealt with steps of randomly varying size. Barron has used fixed step sizes for ease of digital logic implementation. In both cases, convergence to optimum performance occurs in approximately 10–50 trials after disturbance of the system (or the failure of a control logic component).

The above results have been obtained with n in the range of 1 to 6. Little evidence of sensitivity to problem dimensionality has been observed within this range. It has been speculated (but never verified) that performance of the simple PSV controller would be adversely affected by large problem dimensionality ($n > 10$).

In the PSV self-organizing control system approach, the duration of one trial corresponds to the time needed for performance assessment. In the work of Lee and Snyder this was approximately 2–10 times the plant response interval. More recent work has used predictive performance assessment techniques,[27, 28, 29, 30] and consideration is now being given to statistical detection procedures (see elsewhere in this paper). It appears that the practical lower bound on the performance assessment interval is in the neighborhood of 10^{-4} times the plant response interval for essentially linear plants of low order. In nonlinear plants, or in very high-order linear plants, the time required to assess performance is generally greater. When the assessment interval reaches a value in excess of about 10^{-3} times the plant response interval, consideration must be given to partially pre-structured control laws and/or greater use of memory in the learning control system.

The use made of memory in parameter space search generally determines both the capabilities and limitations of the search technique. We have seen (and more data will be given later in this paper) that systems with probability-state memory (viz., the memory inherent in biased direction-of-step probability distributions) are very useful for high-speed learning of the type found in "self-organizing" controllers. The addition of memory functions permits further control over migration of the system state vector. Using memory, several meaningful questions can be asked, for example:

1. Is present performance better than was obtained with the previous trial?
2. Is present performance better than that indicated by a weighted-sum of previous trials?
3. Is present performance better than that obtained with all previous trials?

Question 3 above was the basis for system design studies by Snyder *et al.*[22] and Moddes *et al.*[23] involving training of simulated 360-parameter networks suitable for ballistic re-entry trajectory predictions, the ultimate object being the final-value path control of intercept vehicles. These

prediction networks dealt with patterns in radar metric data. Snyder and Moddes demonstrated repeated success in training the networks to produce close to their best performance after only 200–300 random trials, starting from arbitrary initial parameter combinations.

Question 2 has been suggested by Barron[15] as a means of incorporating a "forgetting memory" in the learning system. Such a memory is important in control and regulator systems because experiences in the remote past usually have less pertinence to present actions than do relatively recent experiences. The harmonic sequence appears to have value as a basis for applying weights to previous trials. It is interesting that "exponential forgetting" (produced by use of the harmonic sequence) is also found in the animal kingdom, as noted by von Foerster.[31]

Question 1 above suggests a strategy that would require little memory but could conceivably improve PSV system performance during those phases of operation when performance trends are obscure.

This discussion of parameter space search indicates the importance of performance evaluation processes. The following section treats the theory of performance assessment within the framework of predictive techniques for self-organizing control systems.

PERFORMANCE ASSESSMENT

A self-organizing control system is one which acts to improve its performance through experience. The goal of the self-organizing controller (SOC) may be expressed mathematically in terms of the minimization of an integral performance index of the general form

$$I = \sum_{k=0}^{k=k_f} \int_{t_k}^{t_k+\Delta t} F\,dt \tag{4}$$

where $F \geqq 0$; the discrete times, t_k, are clocking points at which the controller state changes may occur; and Δt is the interval between two successive clock occurrences.

Assuming the controlled plant to be continuous in the time domain and describable in terms of an Nth-order characteristic equation, one may approximate Equation 4 as follows:

$$I = \sum_{k=0}^{k=k_f} \left[F_{t_k}\Delta t + \left(\frac{d}{dt}F\right)_{t_k} \frac{\overline{\Delta t}^{2}}{2} + \cdots + \left(\frac{d^N}{dt^N}F\right)_{t_k} \frac{\overline{\Delta t}^{N+1}}{(N+1)!} \right] \tag{5}$$

At each point t_k, F and all its derivatives below $d^N F/dt^N$ are predetermined (they are the system state variables). $d^N F/dt^N$ is the lowest-ordered derivative which may be altered by the SOC acting at a point.

It follows that the goal (min I) produces a necessary condition

$$\left(\frac{d^N F}{dt^N}\right)_{t_k} < 0 \quad (k = 1, 2, \ldots, k_f) \tag{6}$$

which system actuator elements can always produce in restraint free cases. Furthermore, given sets of alternative values for $d^N F/dt^N$ at each point t_k, the time-optimal control policy is that which always chooses the most negative value of this derivative. In practice, however, it is sometimes difficult to realize this latter (sufficient) condition because its use requires *a priori* knowledge of plant polarity.

For the special case of $N = 2$, Condition 6 leads to a requirement on d^2F/dt^2. A form of the integrand F which has been investigated very extensively is $F \equiv | e_p |$, where e_p is a predicted error function. The differential order of e_p is of importance in determining the shape of transient responses. In general, to define this shape completely, one would have to deal with

$$e_p = e_{t_k} + \left(\frac{d}{dt}e\right)_{t_k} T + \cdots + \left(\frac{d^{N-1}}{dt^{N-1}}e\right)_{t_k} \frac{T^{N-1}}{(N-1)!} \tag{7}$$

where T is a positive constant. Usually, $T \gg \Delta t$, because this produces a well-damped response of the system during transients.

For a system to have a stable equilibrium state, some scalar function, V, of its state variables must be positive-definite while $\dot{V}$ must be negative-definite. The trivial solution, $V = 0$, is then asymptotically stable in the sense that the state vector will converge toward the trivial solution as $t \to \infty$. For example, if $N = 2$ the state variables are e and $\dot{e}$, and $V = f(e, \dot{e})$. The trivial solution, and hence the equilibrium state for which asymptotic stability is sought, then corresponds to e and $\dot{e}$ becoming zero simultaneously. A geometric interpretation of the above would consist of a plot of the surface defined by $V = f(e, \dot{e})$. Because of the conditions imposed on V and $\dot{V}$, this surface is in the form of a cup or cone which is always above the $e - \dot{e}$ plane, touching it only at the origin which, in this case, represents the desired equilibrium state $e = \dot{e} = 0$.

In synthesis of a self-organizing control system for the case $N = 2$, one approach might be to select V as a positive-definite function of e

and $\dot{e}$ and mechanize $\dot{V}$ in the performance assessment unit. If the value function were generated on the basis of sgn $\dot{V}$ (Lyapunov stability theory), one might expect that the system would seek stable behavior. In practice, however, it is found that the use of $\dot{V}$ does not go far enough (except for the case in which $N = 1$), because $\dot{V}$ says nothing about the desired quality of control, i.e., no requirements are imposed on the manner in which the equilibrium state is to be approached. An improvement is obtained by basing the performance assessment on $\ddot{V}$ (for $N = 2$), requiring that $\ddot{V}$ be negative. The value function (v) is then positive for $\ddot{V} < 0$ and negative for $\ddot{V} > 0$. This last condition achieves two objectives: First, if $\dot{V}$ is positive, the learning process is directed to make $\dot{V}$ negative, so as to assure stable convergence. Second, because V may be viewed as a measure of the "distance" from the desired state (Kalman and Bertram[32]), the requirement $\ddot{V} < 0$ leads to optimizing the rate with which this distance is reduced, i.e., for $N = 2$ the system strives to accelerate convergence toward the equilibrium state.

Recalling that accelerated "motion" to the phase plane origin is desirable only during the initial phase of response, the function V for $N = 2$ may be chosen to be

$$V = |e_p| = |e + T\dot{e}| \tag{8}$$

Equation 8 does not comply with the general requirements for a Lyapunov function, since it is not positive-definite, i.e., $e + T\dot{e} = 0$ is a line in the phase plane passing through the origin. This, however, is often a desirable form, since anywhere on the line the solution of $e + T\dot{e} = 0$ yields

$$e = e(0) \exp(-t/T) \tag{9}$$

where $e(0)$ is the initial attitude error and t denotes time. Thus, once on the line given by $e + T\dot{e} = 0$, system error convergence tends to take place exponentially and T may be viewed as the nominal time constant for the terminal region of response.*

Note that for $N = 2$ the condition $\ddot{V} < 0$, with V given by Equation 8, is equivalent to Condition 6, with F taken as the absolute value of e_p from Equation 7. Both of these conditions call for accelerated system acquisition of the switching line and constrained (exponential) convergence along the switching line to the origin. This is illustrated in Figure 2,

* It is recommended that the function $V = |e| + T|\dot{e}|$ also be investigated.

where the surface defined by Equation 8 now takes the form of a trough touching the phase plane along the $e = -T\dot{e}$ line. Although this surface seems to be unimodal if viewed formally in terms of the function $V = V(e, \dot{e})$, it is, in general, multimodal in terms of the controller independent variables in the vector space u_i $(i = 1, 2, \ldots, n)$.

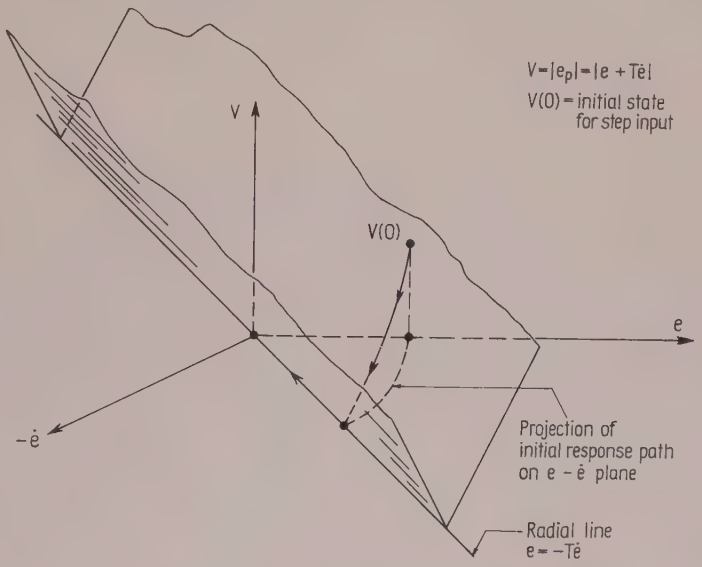


Fig. 2. State Space Representation of SOC Performance

Condition 6 will now be employed to derive two types of performance assessment (PA) criteria for the case $N = 2$.

"Type 1" PA Criterion

For the case $N = 2$ and taking $F = |e_p|$, Condition 6 requires that

$$\frac{d^2}{dt^2}|e_p| < 0 \tag{10}$$

where, from Equation 7

$$e_p = e + T\dot{e} \tag{11}$$

Introducing a binary value function, v,* Equation 10 yields

$$\operatorname{sgn} v = -\operatorname{sgn} \frac{d^2}{dt^2} |e_p| \tag{12}$$

where $v = +1$ is defined as a positive reinforcement (reward) and $v = -1$ as a negative reinforcement (punishment). Equation 12 is referred to as the "Type 1" PA criterion.

One also readily obtains

$$\operatorname{sgn} v = -\operatorname{sgn} e_p \cdot \operatorname{sgn} \frac{d^2}{dt^2} e_p \tag{13}$$

as an alternative (but functionally identical) Type 1 criterion.

Operation of the Type 1 criterion is best visualized with reference to Equation 13. The sign of predicted error (viz., sgn e_p) is coordinated with the sign of the acceleration of predicted error (viz., sgn d^2e_p/dt^2). If the latter has the opposite sign from the former, a positive reinforcement pulse is obtained and *vice versa* for the case in which the two quantities have the same sign. It is thus immaterial whether the plant polarity $(\partial \dot{e}_p/\partial u)$ happens to be positive or negative.

"Type 2" PA Criterion

In some applications the following restrictive assumption is admissible

$$\operatorname{sgn} \frac{d^2}{dt^2} e_p = -\operatorname{sgn} u_p \tag{14}$$

whence Equation 13 becomes

$$\operatorname{sgn} v = \operatorname{sgn} e_p \cdot \operatorname{sgn} u_p \tag{15}$$

where (as before)

$$e_p = e + T\dot{e}$$

and now

$$u_p = u + k \operatorname{sgn} \Delta u \tag{16}$$

The parameter k in Equation 16 is a positive constant set sufficiently large that sgn u_p = sgn Δu, except when $\Delta u = 0$ (which event can occur at $u = u_{\max}$ and at $u = u_{\min}$).

* In PSV systems, it is necessary to convert score function information into a probability bias, and very little of the information in the score function is needed for this purpose. In general, all that is necessary is the algebraic sign of the score function. All such quantized versions of the score function are called *value functions*.

The obvious merit of Equation 15, which is termed the "Type 2" PA criterion, is the absence of a d^2e_p/dt^2 term, making operation of the SOC much less sensitive to environmental and sensor noise and the order of the plant. The drawback of the Type 2 criterion is its restriction to applications in which the polarity of $\partial\ddot{e}_p/\partial u$ is known (this polarity is taken as negative in Equation 14).

The state space surface shown in Figure 2 pertains to the Type 2 as well as to the Type 1 criterion.

Multivariable PA

The assessment of performance in a self-organizing or learning control system refers primarily to total system performance and not to the performance of individual elements or degrees of freedom within the system. This suggests that control system performance should be expressed mathematically in terms of an ensemble of sensor signals combined within a single error time function. The purpose of this error function is to place proper relative weight on the different response variables, including derivatives of the more important variables. By selection of the weighting factors, the designer determines such things as transient response shapes and the trade-offs that apply between the variables.

One general multivariable error function which is suggested for learning control systems investigation is:

$$E = \sum_{i=1}^{i=M} k_i \left[|e_i| + T_i |\dot{e}_i| + \cdots + \frac{T_i^{N_i-1}}{(N_i-1)!} \left| \frac{d^{N_i-1} e_i}{dt^{N_i-1}} \right| \right] \qquad (17)$$

where:

E = multivariable predicted error function
e_i = ith error component (command minus response)
k_i = ith weighting factor (positive constant)
T_i = ith interval of prediction (positive constant), usually selected empirically to obtain desired response shape in terminal stages of transients (value of T_i not critical).
N_i = differential order of ith response variable

The value of M is usually a function of the number of degrees of freedom of the controlled plant.

ON MECHANIZATION OF SELF-ORGANIZING CONTROL SYSTEMS

One of the objects in synthesis of self-organizing and learning control systems is to realize designs which require a minimum of *a priori* information to produce good system performance. The Type 1 performance assessment criterion provides important advantages over the Type 2 criterion, because the derivation of the Type 1 criterion contains no assumptions as to the polarity of the controlled elements.

Control-loop sign reversals occur frequently in advanced applications. For example, the coupled lateral/directional dynamics of high-performance aircraft exhibit effects such as adverse yaw, which can produce aircraft heading changes opposite from the directions implied by rolling maneuvers. A further example is the problem of spacecraft magnetic actuation, often used for the purpose of removing stored angular momentum. Here, the relative orientation of the geomagnetic field as seen by the vehicle can reverse periodically during the orbit of an Earth-pointing satellite. Of course, many reversal effects are computable, but at what point does their computation become uneconomical?

Reversals of control-loop sign are generally characteristic of body orientation problems whenever the assessment of performance is conducted relative to space-fixed coordinates while actuation (torque generation) occurs relative to bodyfixed axes. As the body rotates, the angular accelerations seen in the inertial reference frame show a nonlinear dependence on angular accelerations existing in body coordinates. Airplane stunt pilots soon learn that, when flying inverted, they should push the stick forward to accomplish effects (relative to the ground) which normally call for rearward stick forces.

Since most of us have done some amount of computer programming, we tend to look on the sign of a variable with a certain degree of disdain. The sign, we have found, is worth only one bit in a digital data word, only one amplifier in the analog program. Surely sign information is the last thing to be concerned about!

As it happens, sign information is also the first thing to be concerned about. It is the key to understanding of the mechanization, capabilities, and limitations of various self-organizing control systems.

Figure 3 presents a block diagram of the Type 1 criterion. This is actually an idealized PA module because, in practice, the signal $\ddot{e}_p$,

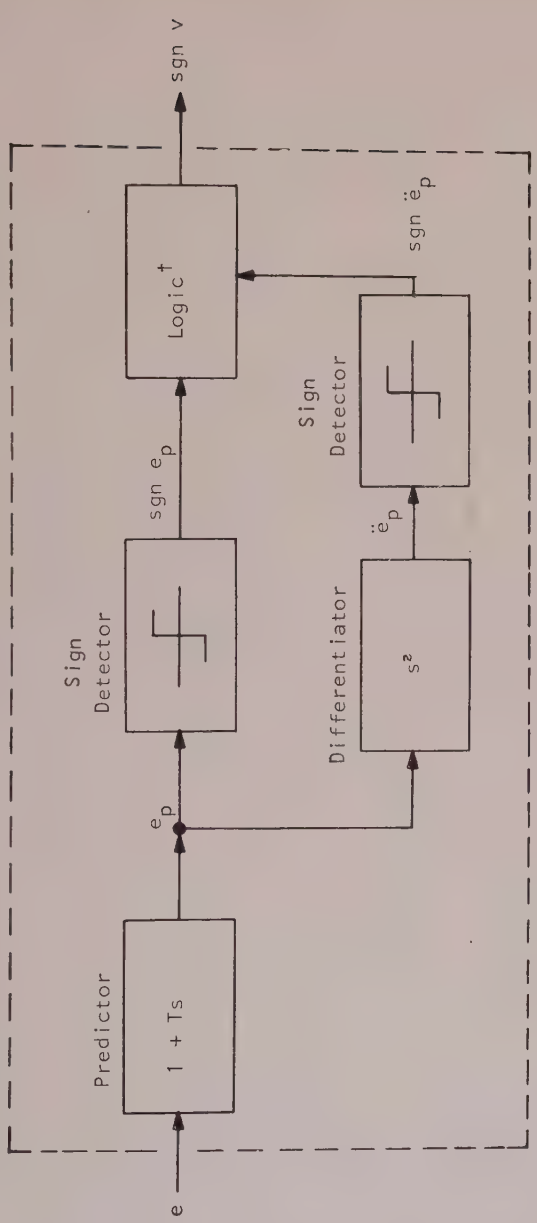


Fig. 3. PA Module for Self-Organizing Controllers (Type 1 Criterion).*

* Patent Pendig

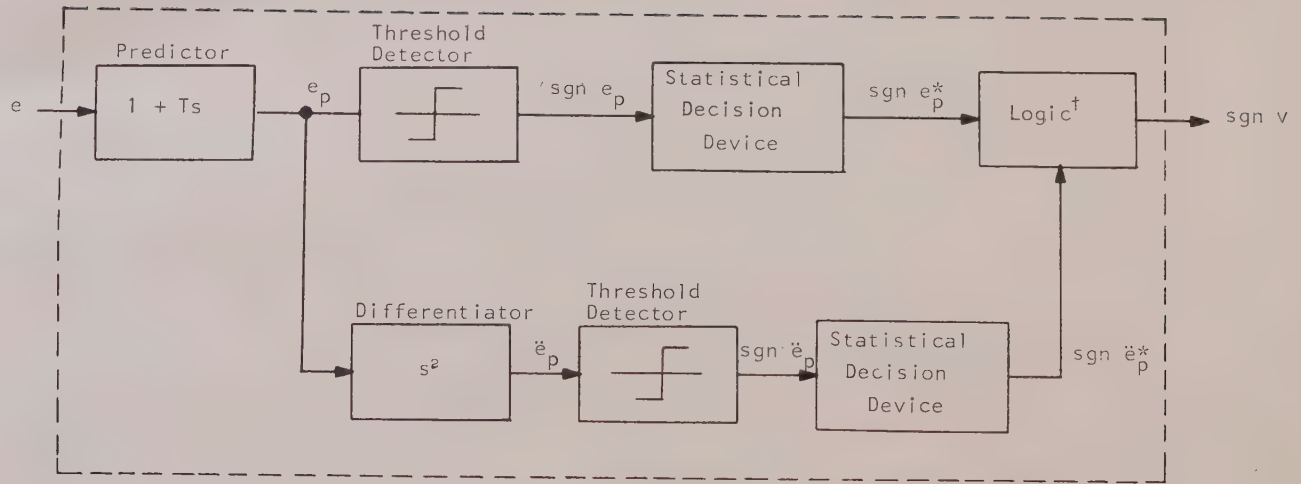


Fig. 4. PA Module Employing Statistical Decision Devices (Type 1 Criterion).

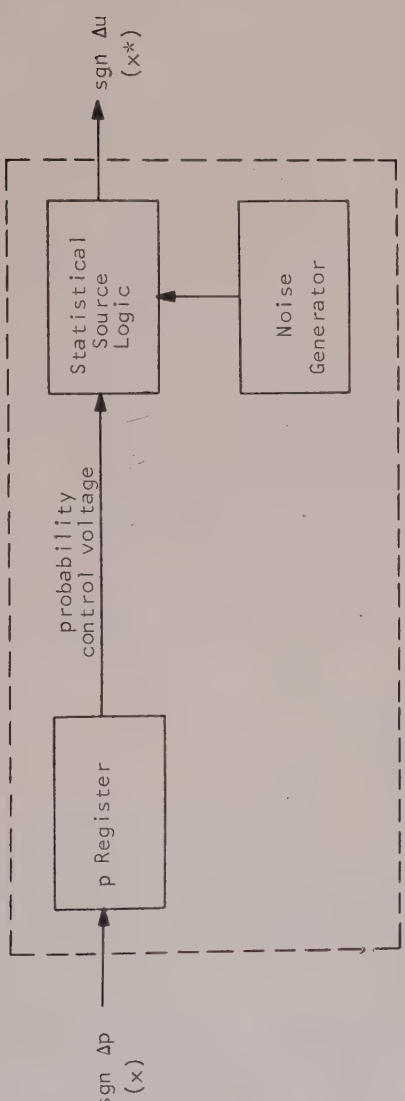


Fig. 5. A Statistical Decision Device.

when found by the operations shown here, is much too noisy for effective determination of the value function. Figure 4 illustrates how *statistical decision devices* might be used to improve operation of the Type 1 PA module by decreasing the frequencies of "false alarms" in determination of sgn $\ddot{e}_p$ and sgn e_p.

Figure 5 illustrates the functions contained in one form of statistical decision device. The binary (± 1) inputs, drive a bi-directional, reversible binary counter, called the *p Register*. The contents of this register are converted into a multi-level d.c. voltage used to control a form of comparator termed the *statistical source*.[23] The statistical source is a random pulse generator with biasable statistics, as shown in Figure 6. The noise

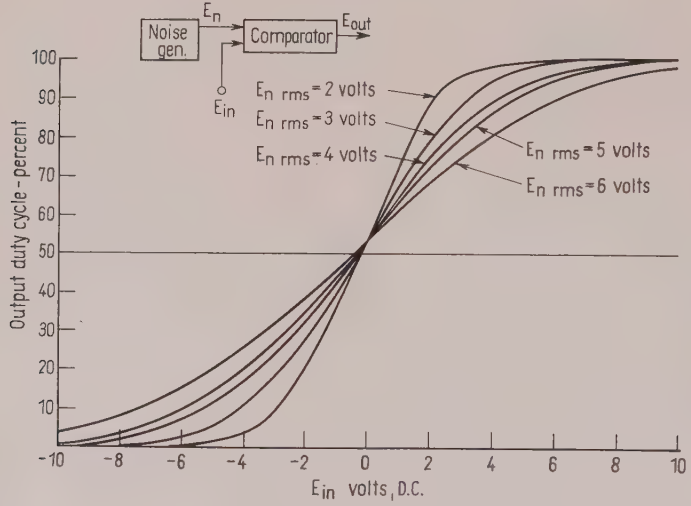


Fig. 6. Transfer Function of Typical Statistical Source (Output Duty Cycle vs. Input Control Voltage).

generator provides a signal, E_n, which is tested against the probability control voltage from the p Register to determine whether the statistical source output will be positive or negative at any instant of comparison. We see that the p Register contains information as to the *probability state* of the decision device, and the output, $x*$, is a function of the input, x, and the probability state.

The role of the noise generator* is interesting philosophically. Ashby's Law of Requisite Variety[10] tells us that if the environment of a system contains variety, the system must also have available variety, because "only variety can nullify variety".[1] In the statistical detection problem, the variety of the environment expresses itself in the form of noise (false-alarms) on the input channel. We are, clearly, dealing here with a game-theoretic problem if we regard the signal x as being the move of an adversary who seeks to mask his intentions via a mixed strategy of moves. In this context, whether or not the control system environment is intelligent need not be known: our decision device seeks only to discern the presence of any bias in the $x(t)$ signal.

The familiar "threshold learning process" described by Sklansky[14] and others uses a reinforcement process to move a decision boundary (threshold) as a function of the value of the transmitted signal. The problem with this process remains one of value assessment.

Application of the statistical decision device shown in Figure 5 may be visualized in terms of the problem of reducing false-alarm frequency in a signal, sgn e_p, which is taken from a sign detector that is fed by a prediction circuit. With reference to Figure 7, we see that the *a priori*

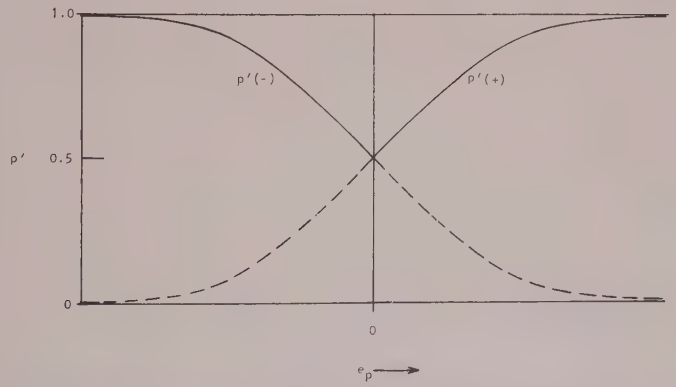


Fig. 7. Single-Sample Probability (p) of Measuring e_p with Correct Sign in Presence of Unbiased Noise.

single-sample probability, p', of measuring sgn e_p correctly is a function of the magnitude of e_p. At $e_p = 0$, $p' = 0.5$, while $p'(+)$ and $p'(-)$

* A coherent signal source can be used in some instances.

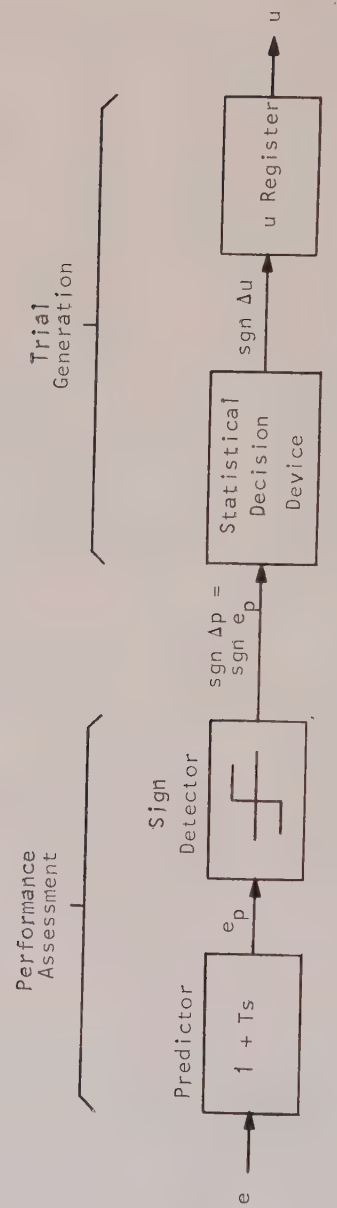


Fig. 8. Basic Functions in Elementary Self-Organizing Controller.

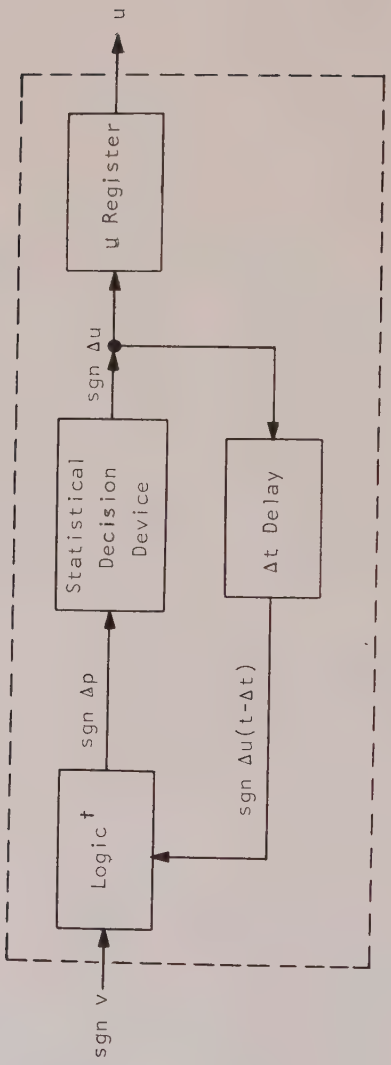
tend to unity as $e_p \to +\infty$ and $e_p \to -\infty$, respectively. The goal of the decision device is to achieve something close to unity probability that its output, sgn e_p^*, will always be correct in terms of the true sign of e_p, regardless of the magnitude of e_p. This goal must obviously elude all real systems in the immediate vicinity of the $e_p = 0$ condition. But for finite e_p values we can usually obtain improvement relative to single-sample probabilites. The amount of this improvement depends on the sample frequency and the size of the e_p derivatives.

Figure 8 is a block diagram of an elementary self-organizing controller which employs the statistical decision device. A lead circuit is used to obtain e_p. A zero-crossing detector then provides sgn e_p. (Up to this point we have a relay controller with lead.) The statistical decision device is inserted between the zero-crossing detector and an up-down counter, called the u Register, which serves to integrate the pulses from the decision circuit. The u Register can be identical to the p Register, described above, but in practice the u Register is constructed with more levels than the p Register. Although the controller shown in Figure 8 qualifies under the definition as a self-organizing controller, it clearly has very restricted adaptation capabilities. This simple controller concentrates its entire attention on the task of sensing true polarity of e_p and driving the plant so as to null e_p. Plant polarity is assumed to be known *a priori.*

A considerable amount of laboratory data has been obtained over the past year with prototype SOC equipment connected so as to realize the controller shown in Figure 8. These data (see next section) establish the general validity of a probability state variable approach to statistical detection in its application to generation of control signals. In brief, the elementary system of Figure 8 has proven to be a versatile, high-performance controller.

Ultimately, it becomes necessary to mechanize the Type 1 PA criterion (Figures 3 and 4), or something equivalent, so as to secure many of the fundamental objectives of self-organizing control, particularly in the area of multi-degree-of-freedom applications, in which no generally-valid assumptions can be made about the signs of the transfer functions $\partial \ddot{e}_p / \partial u_i$ $(i = 1, 2, \ldots, n)$. With this in mind, it is useful to fabricate general-purpose PSV modules (Figure 9) for use as building-block elements in studies of self-organizing controller applications.

The heart of the general-purpose PSV module is a statistical decision



†sgn Δp = sgn v · sgn Δu(t-Δt)

Fig. 9. General-Purpose PSV Module for Self-Organizing Controllers.*

* Patent Pending.

element, of the type described above, which is contained within a sgn Δu feedback path, as illustrated in Figure 9. The input logic of the PSV module accepts a binary value signal from an external means of performance assessment. This value signal controls the sign of the sgn Δu loop via gating which implements the function

$$\text{sgn } \Delta p = \text{sgn } v \cdot \text{sgn } \Delta u(t - \Delta t) \tag{18}$$

where sgn Δp is the polarity of an increment to the p Register (this register is part of the statistical decision device), sgn v is the binary value signal, and sgn Δu $(t - \Delta t)$ is the polarity of the step taken one clock interval earlier. The sgn Δu pulses from the decision device are integrated in the u Register (output register).

J. Gatlin of the Goddard Space Flight Center, NASA, has pointed out that sgn $v = +1$ in Equation 18 produces positive feedback (regeneration) in the sgn Δu loop, while sgn $v = -1$ results in negative feedback (degeneration). A condition of regeneration eventually takes the system to a fully-biased probability state, with the polarity of this bias depending on the detected sense of the correlation between sgn v and sgn Δu $(t - \Delta t)$ over a sequence of samples. Conversely, a condition of degeneration causes the system to revert to essentially no net bias (50 : 50 duty cycle).

One of the purposes of the sgn Δu feedback is to permit an arbitrary number of PSV modules to be driven by a single performance assessment circuit. The assessment of performance can thus relate to the over-all behavior of the self-organizing system, at the same time as the u_i signals (PSV module outputs) relate to appropriate actions for the various actuators. This concept is termed *distributed-actuation* control.

RESULTS OF APPLICATIONS

The elementary self-organizing controller illustrated in Figure 8 may be realized using the general-purpose PSV module of Figure 9 in conjunction with a Type 2 PA element. The block diagram presented in Figure 10 shows the signal-flow paths for such a system. This system is equivalent to the one shown in Figure 8. Usually, several controllers of the type in Figure 8 (or Figure 10) are connected in parallel, with their

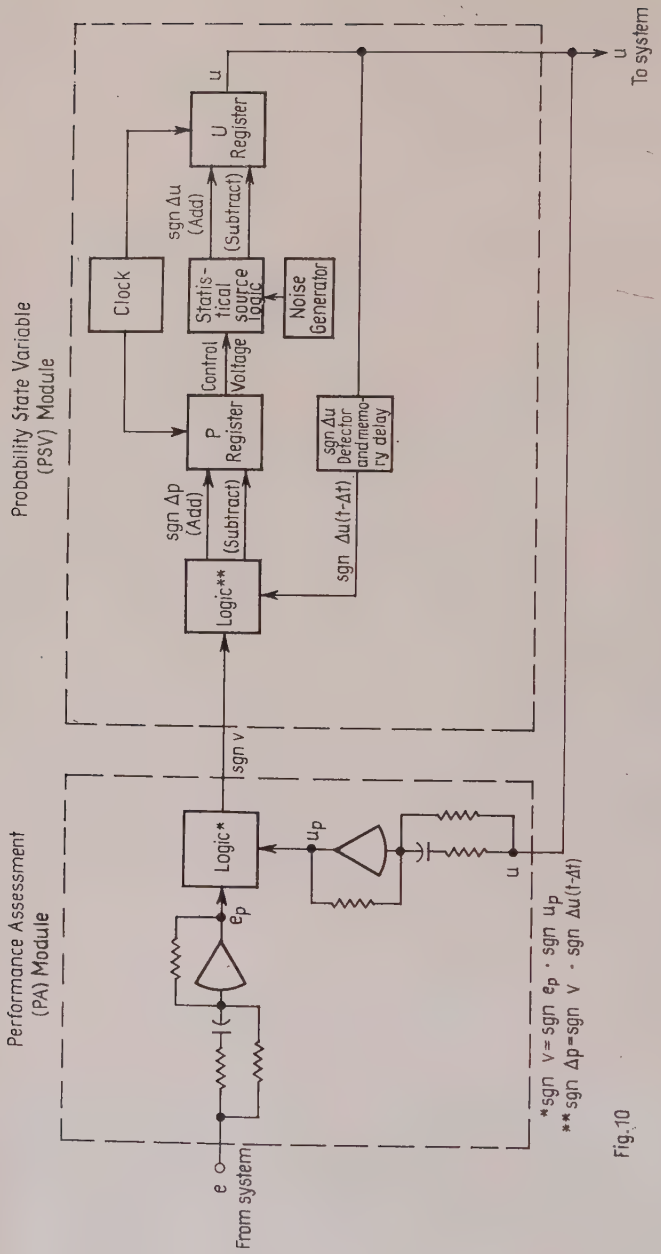


Fig. 10. Mark III Self-Organizing Controller Block Diagram

Fig.10

respective outputs summed algebraically.* This use of a parallel network provides several benefits:

(i) greater dynamic range in the resultant u signal

(ii) greater statistical variety in terms of the microscopic behavior of the resultant u signal

(iii) reliability of operation, because failure of a single path need not be catastrophic vis-à-vis the behavior of the over-all system.

The Mark III self-organizing controller is a laboratory prototype SOC fabricated by Adaptronics, Inc. under Air Force sponsorship. General-purpose PSV modules (Figures 9 and 10) are employed in the Mark III system. These PSV modules can accept a binary value signal input from either Type 1 or Type 2 performance assessment circuits; however, the Type 2 PA has received much greater attention in investigations conducted thus far. Figures 11 and 12 are photographs of Mark III equipment.

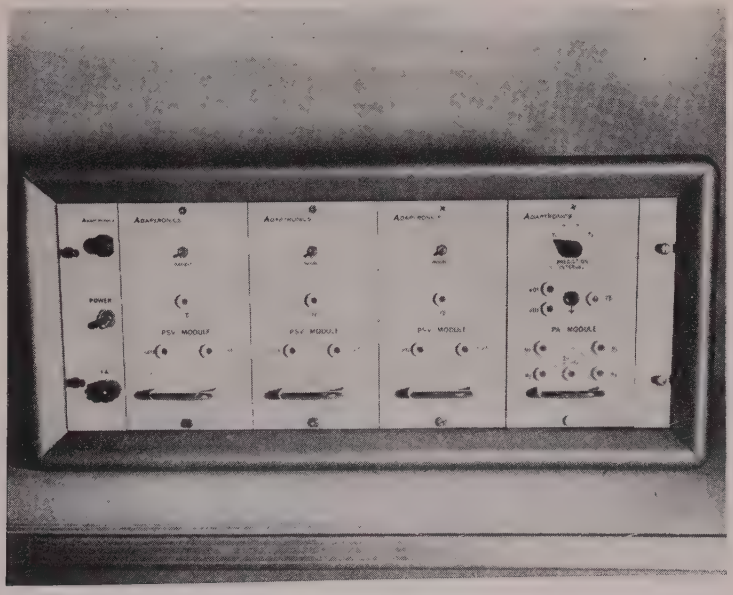


Fig. 11. Mark III Self-Organizing Controller

* In some cases, the outputs are also added to the output of a conventional controller.

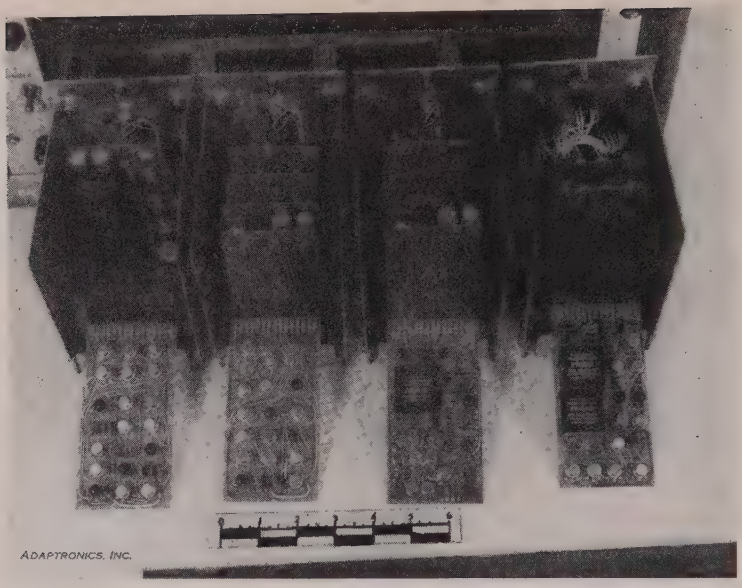


Fig. 12. Mark III Self-Organizing Controller Exploded View

Numerous laboratory experiments, using simulated vehicle dynamics, have been performed with the Mark III system relative to several areas of potential SOC application:

1. pitch-rate and normal-acceleration control of a high-performance aircraft
2. single-axis and multiple-axes control of orbiting satellites
3. throttle control for aircraft landing approaches
4. control of large, flexible space launch vehicles

Work in the first of the above areas is presently the farthest advanced, and the remainder of this discussion is largely devoted to a summary of results obtained in investigations of SOC control augmentation for a representative aircraft.

Figure 13 presents the block diagram of an SOC control augmentation system for handling short-period longitudinal dynamics of the F-101 B aircraft.[30] Representative servo-actuator and sensor dynamics are included. An input prefilter (model) is used to shape the closed-loop response (which would otherwise be excessively fast from the standpoint of pilot opinion). The SOC consists of a network having, nominally,

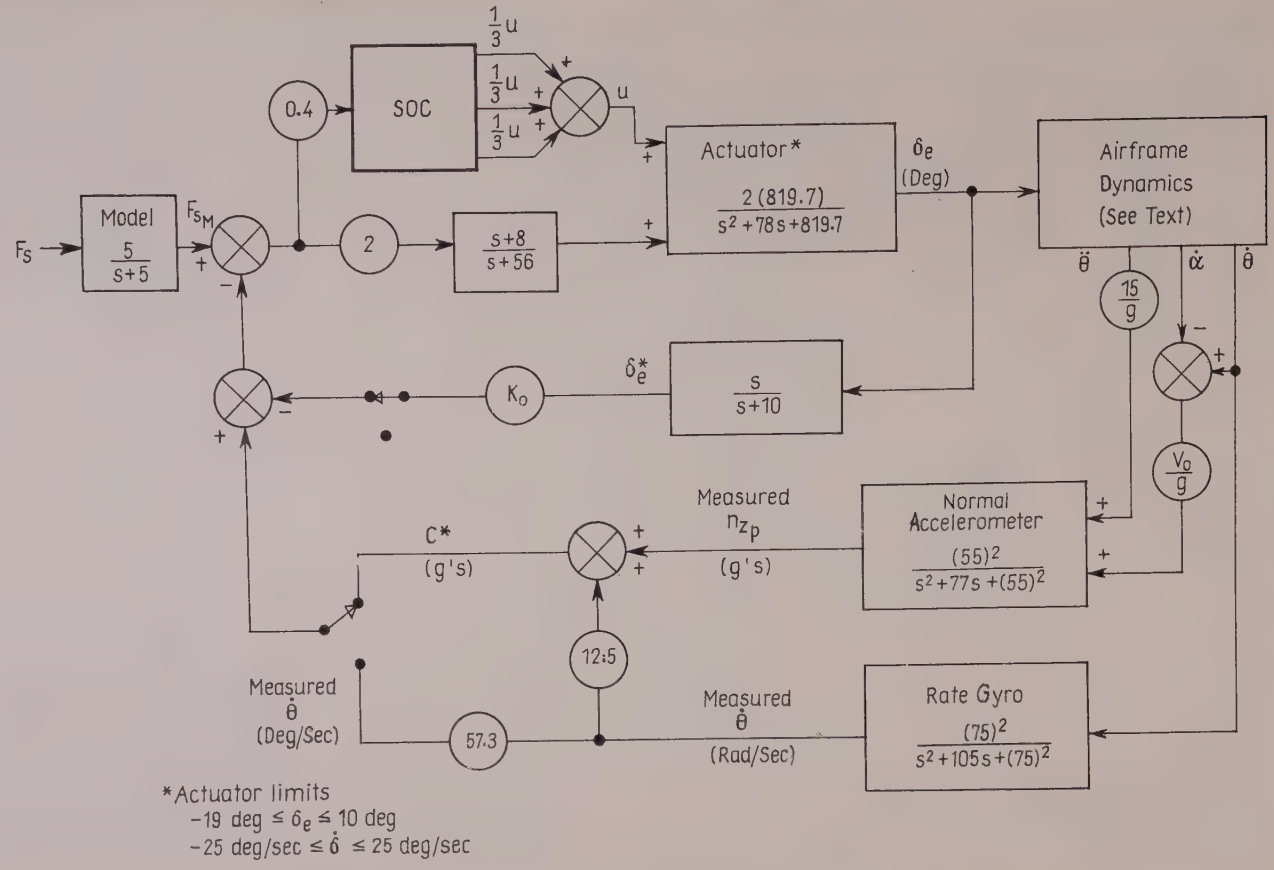


Fig. 13. F-101 B SOC System Block Diagram

three each Type-2 PA/PSV control devices, with algebraic summation of PSV module outputs used to generate the actuator excitation signals. In some of this work a conventional forward-loop controller was inserted in parallel with the SOC; such a controller is represented by the optional $(s + 8)/(s + 56)$ transfer function shown in Figure 13. The δ_e feedback path indicated in the figure is also optional: this circuit was studied as a means for compensation of rate-gyro dynamics.[30]

In dealing with piloted-vehicle applications it is essential to consider the requirements of handling qualities. The blended, pitch-rate and normal-acceleration feedback used in this system produces the following signal

$$C^* \equiv n_{z_p} + 12.5\,(\dot{\theta}/57.3) \tag{19}$$

where $\dot{\theta}$ has dimensions of degrees/second, and n_{z_p} and C^* have dimensions of "g's". The subscript "p", when used in conjunction with n, indicates that the accelerometer is taken to be located at the pilot's cockpit station. The quality of control should be judged by comparing the C^* response with the filtered stick force signal (F_{s_M}). The technique of using C^* feedback as a means of providing suitable short-period handling characteristics for fighter aircraft is discussed by Malcolm and Tobie[33] within the framework of conventional control system design. Figure 14 indicates the C^* time response of the SOC input model to a unit-step command and compares this model response to various performance envelopes established in Reference 33.

Figures 15–33 summarize the results obtained with the self-organizing controller in the F-101B pitch-rate and normal-acceleration control augmentation work.* Because Reference 30 contains a detailed description of these and additional results, we will not present full discussion in this paper.

Figures 15–19 relate to a low-"q" (dynamic pressure) flight condition, Mach 0.4 at 20,000 feet. These figures show in sequence the free airframe response, the response of the closed-loop SOC system to a $2g$ command, a detailed plot of error and error-rate time responses, the response as seen in the phase-plane, and the time history of C^* (plotted on the coordinates of Figure 14). Figures 20–24 present in the same sequence the results obtained at high-q condition, Mach 1.4 at 20,000 feet.

* Prediction intervals, T_i $(i = 1, 2, \cdots)$, of 0.1 sec were used in all PA units in obtaining these results.

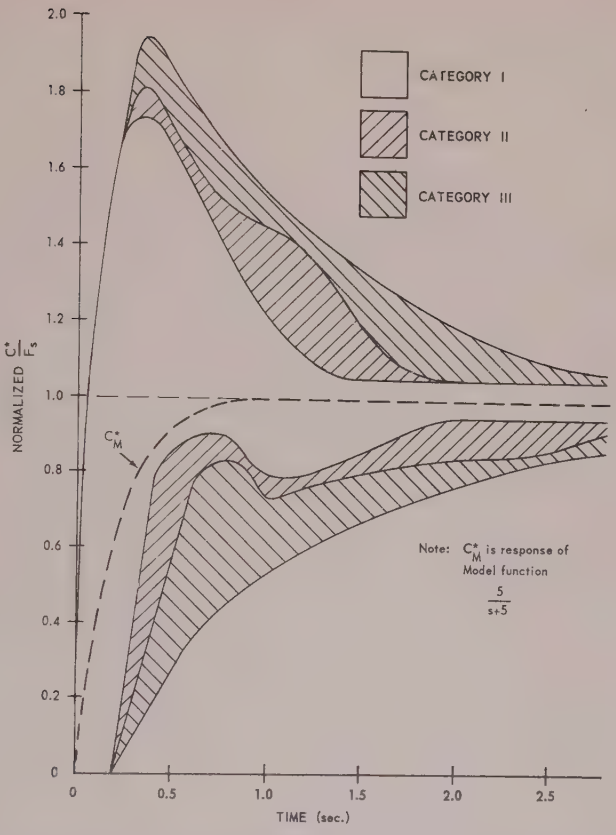


Fig. 14. Various $\frac{C^*}{F_s}$ Step-Response Envelopes (from Reference 33).

It is evident that the SOC compensates for the short-period dynamics of the airframe, as well as the dynamics of actuators and sensors, satisfactorily for both flight conditions, and that the C^* response adheres closely to the model response in both cases. In all, some 12 flight conditions, representative of the entire performance envelope of the F-101 B, have been investigated,[30] with the conclusion that SOC response is essentially uniform under all flight conditions.

Figures 25 and 26 were obtained from a hybrid simulation of airframe dynamics, which permitted a programmed acceleration from Mach 0.4 to Mach 1.6 at 10,000 feet. In Figure 25, the SOC has a constant

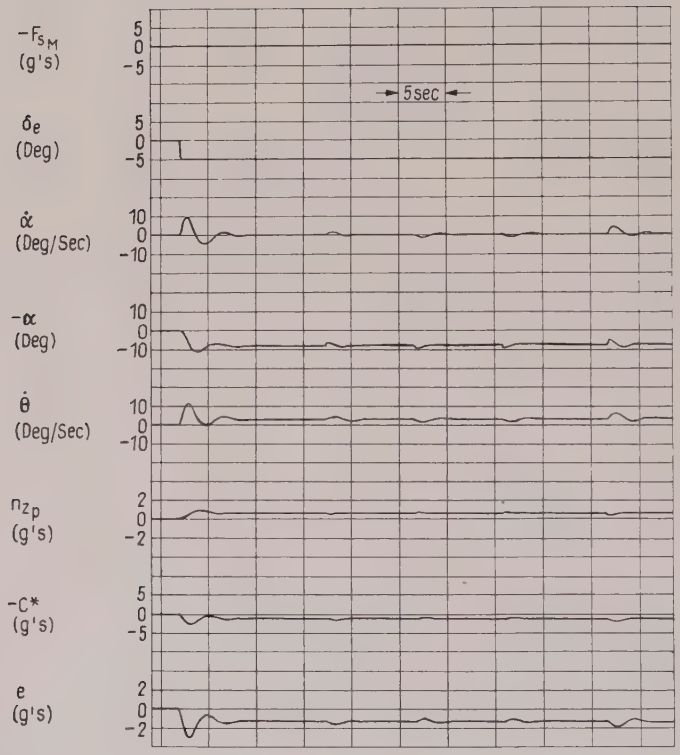


Fig. 15. Free Airframe, F-101 B, 20,000 ft. Mach 0.4.

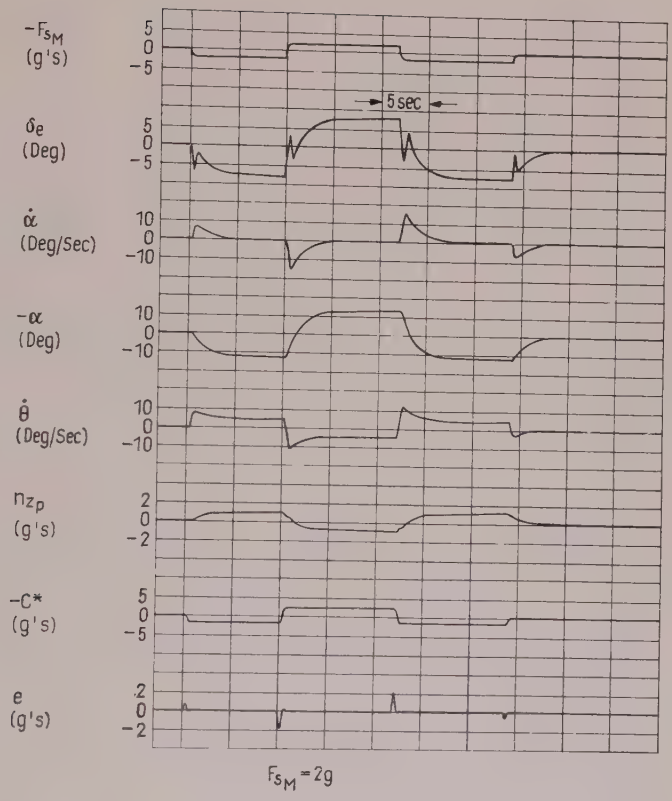


Fig. 16. SOC Configuration A, F-101 B, 20,000 ft. Mach 0.4.

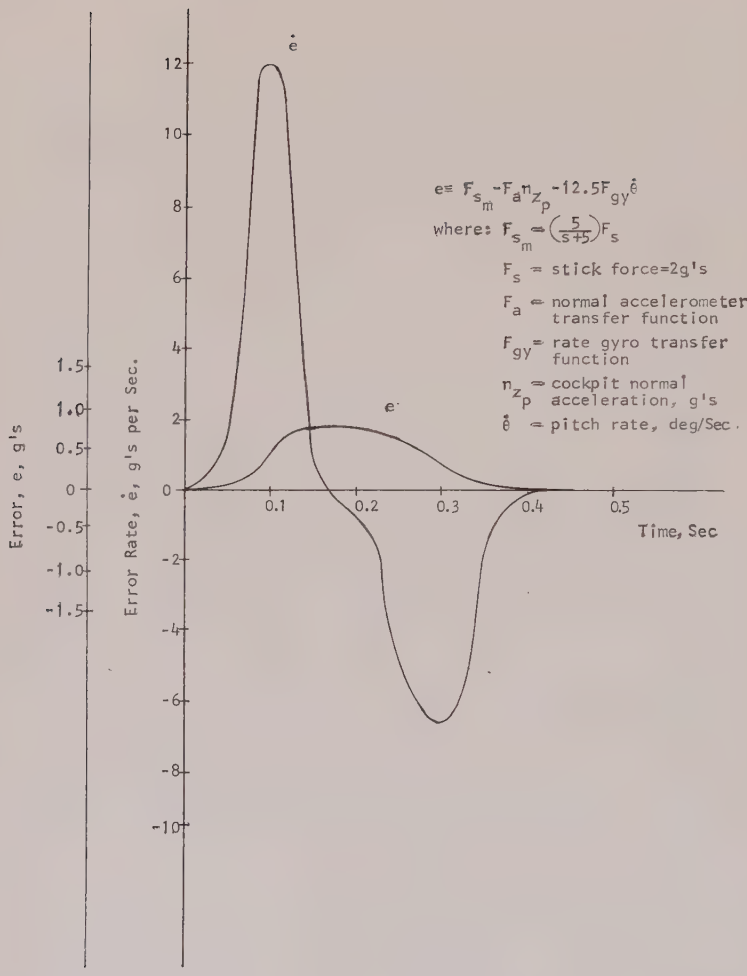


Fig. 17. Error and Error-Rate Time Response, Mark III SOC Configuration A, F-101 B, 20,000 ft. Mach 0.4.

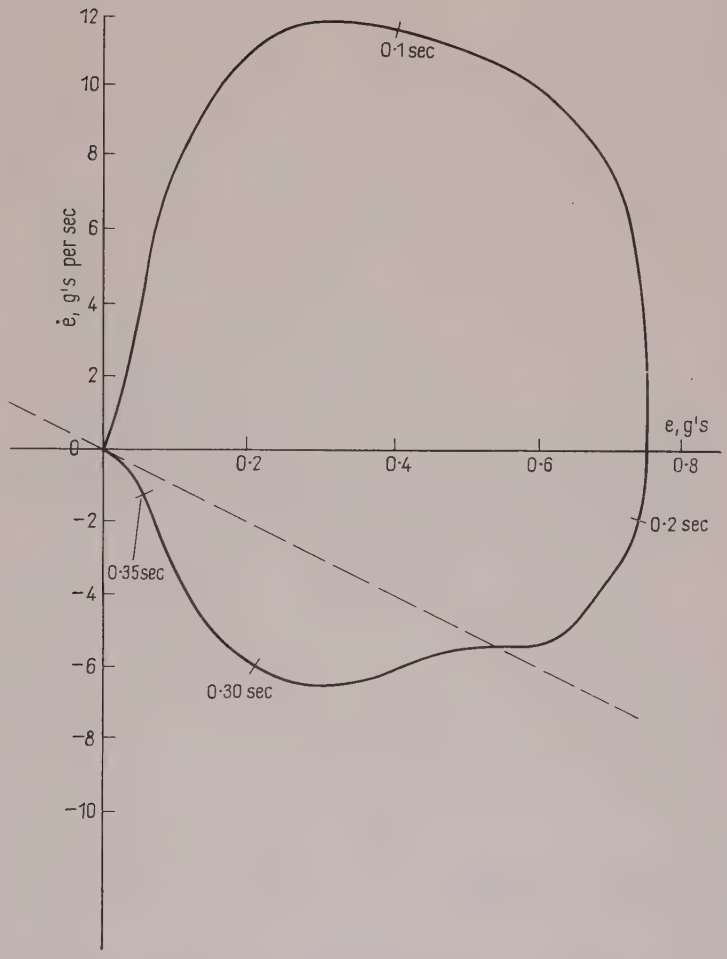


Fig. 18. Phase Plane Diagram, Mark III SOC Configuration A, F-101 B, 20,000 ft. Mach 0.4

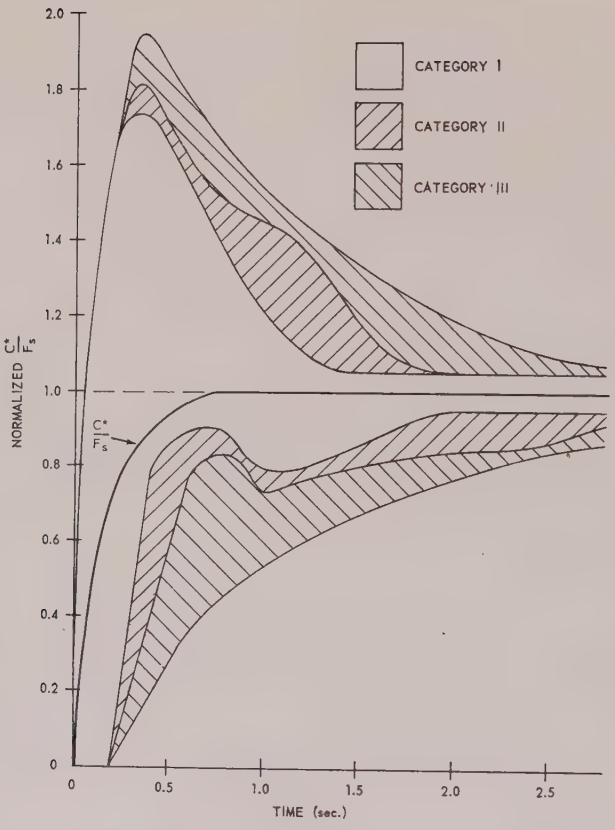


Fig. 19. $\frac{C^*}{F_s}$ Step-Response, 0.4 g Command, SOC Configuration A, F-101 B, 20,000 ft. Mach 0.4.

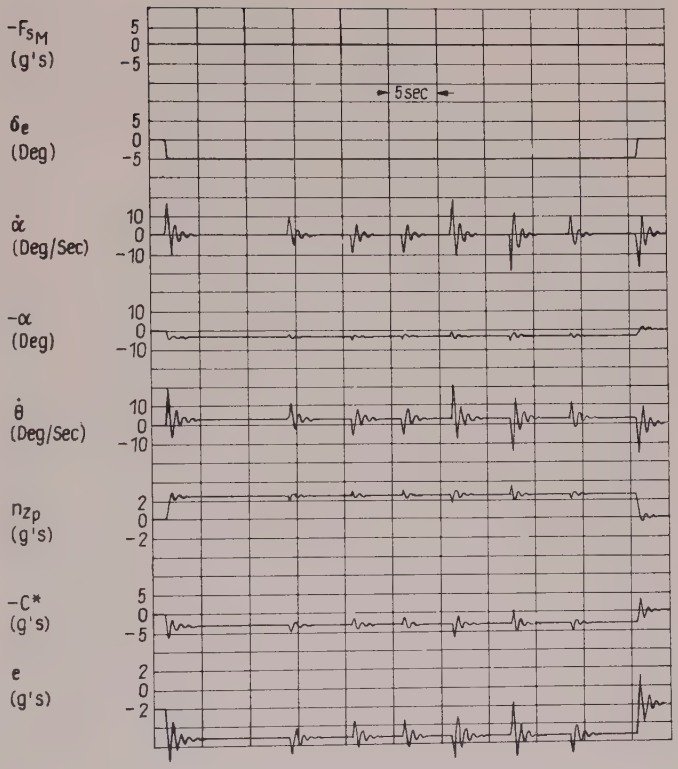


Fig. 20. Free Airframe, F-101 B, 20,000 ft. Mach 1.4.

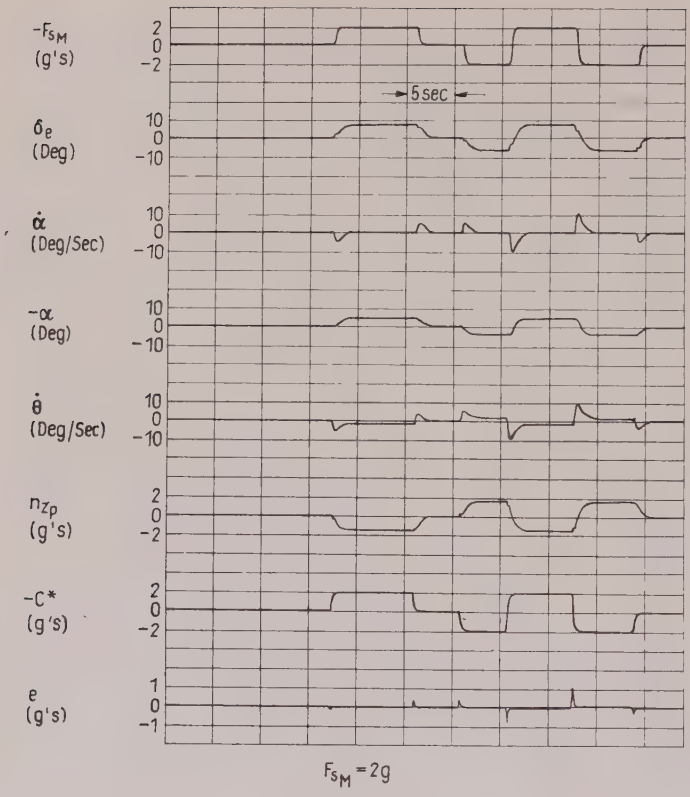


Fig. 21. SOC Configuration A, F-101 B, 20,000 ft. Mach 1.4.

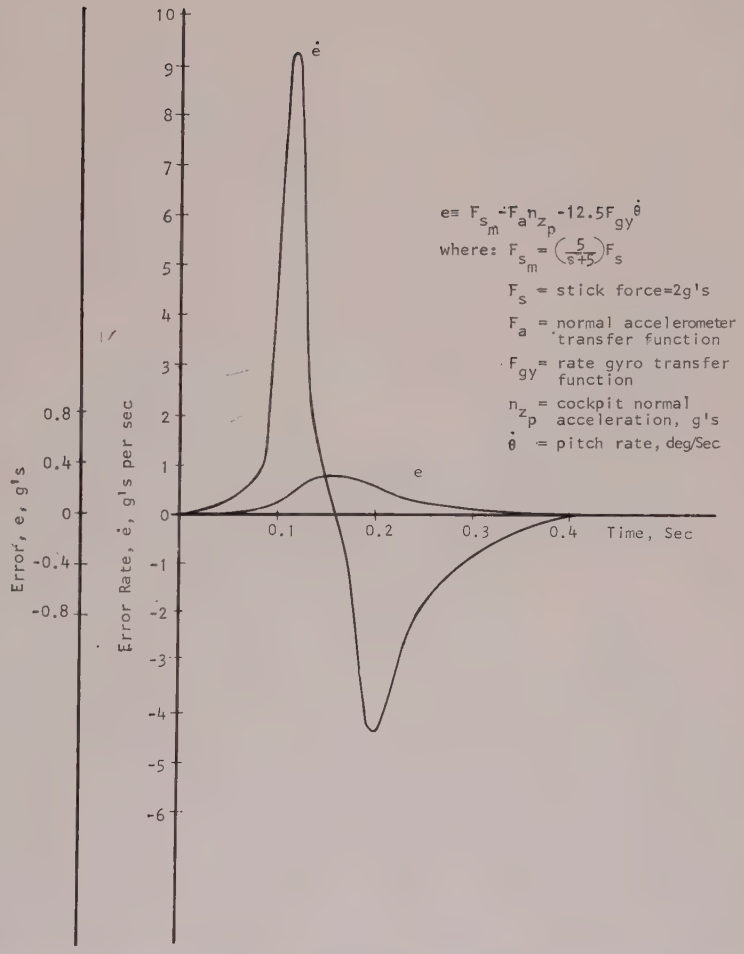


Fig. 22. Error and Error-Rate Time Response, Mark III SOC Configuration A, F-101 B, 20,000 ft. Mach 1.4.

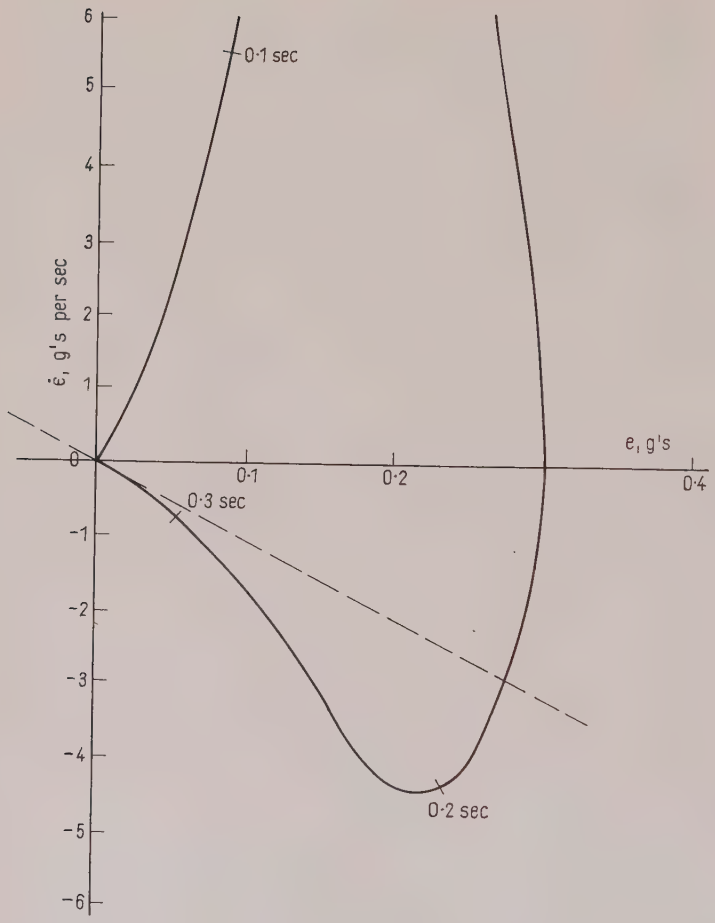


Fig. 23. Phase Plane Diagram, Mark III SOC Configuration A, F-101B, 20,000 ft. Mach 1.4

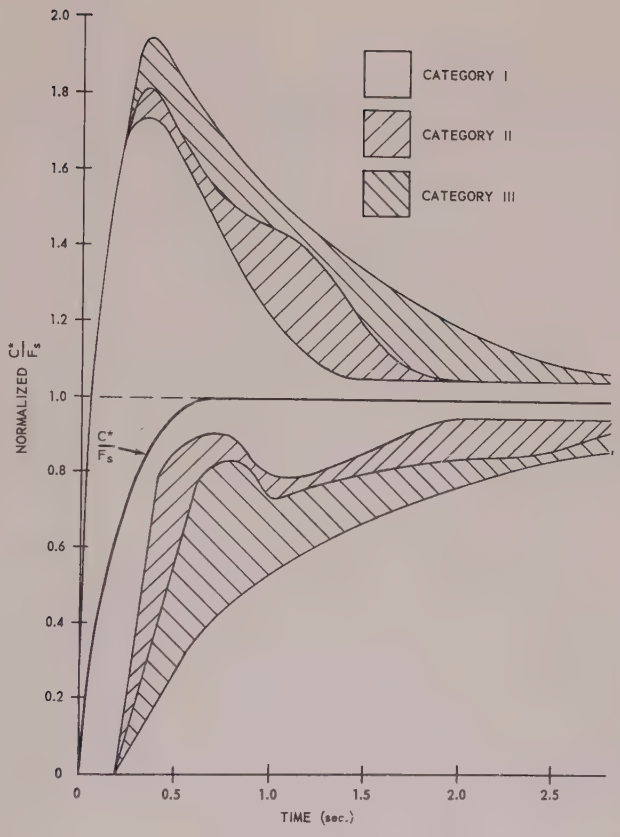


Fig. 24. $\frac{C^*}{F_s}$ Step-Response, 0.4 g Command, SOC Configuration A, F-101 B, 20,000ft. Mach 1.4.

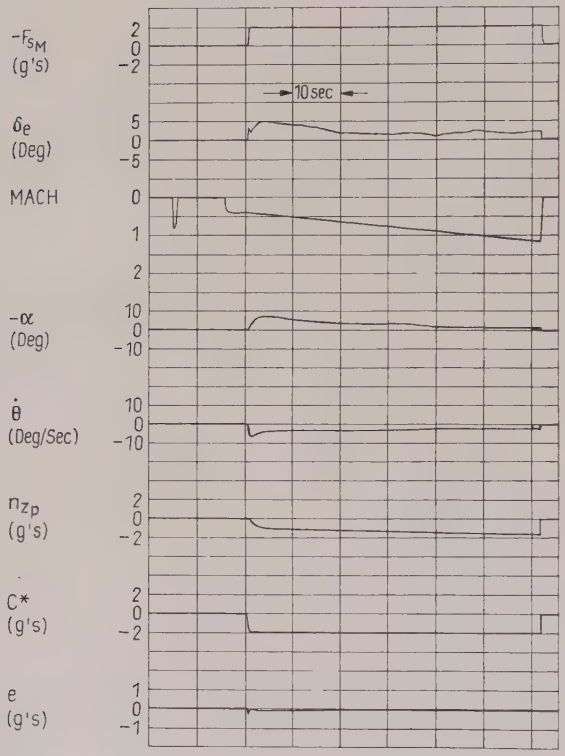


Fig. 25. SOC Configuration F, F-101 B
10,000 ft. Mach 0.4 to 1.6

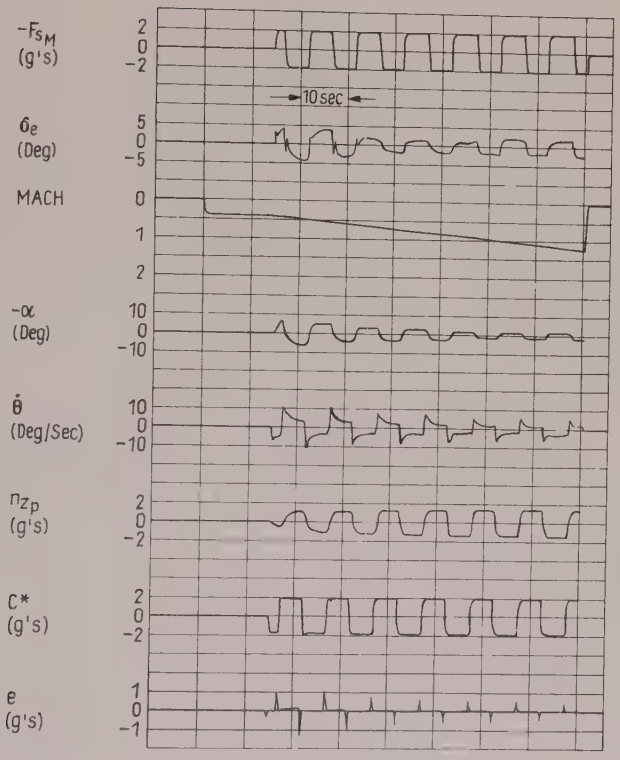


Fig. 26. SOC Configuration F, F-101 B
10,000 ft. Mach 0.4 to 1.6

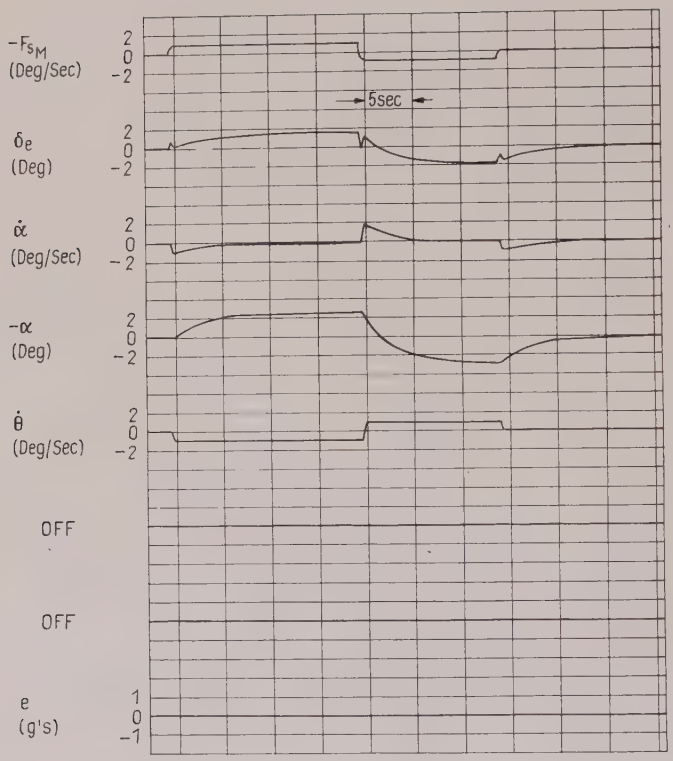


Fig. 27. SOC Configuration D, F-101 B
20,000 ft. Mach 0.4

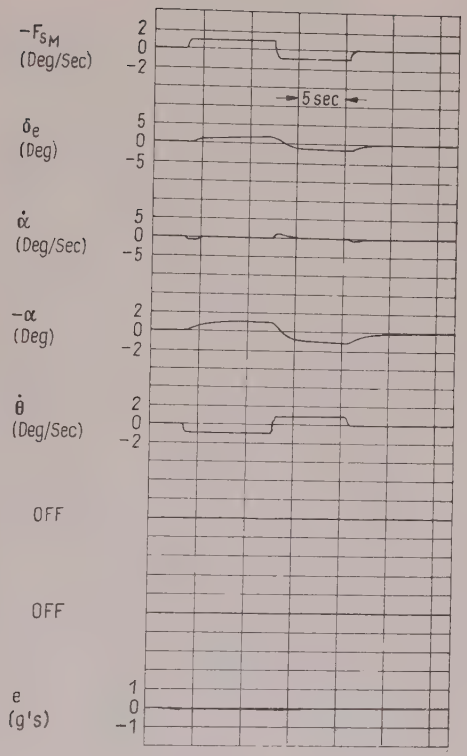
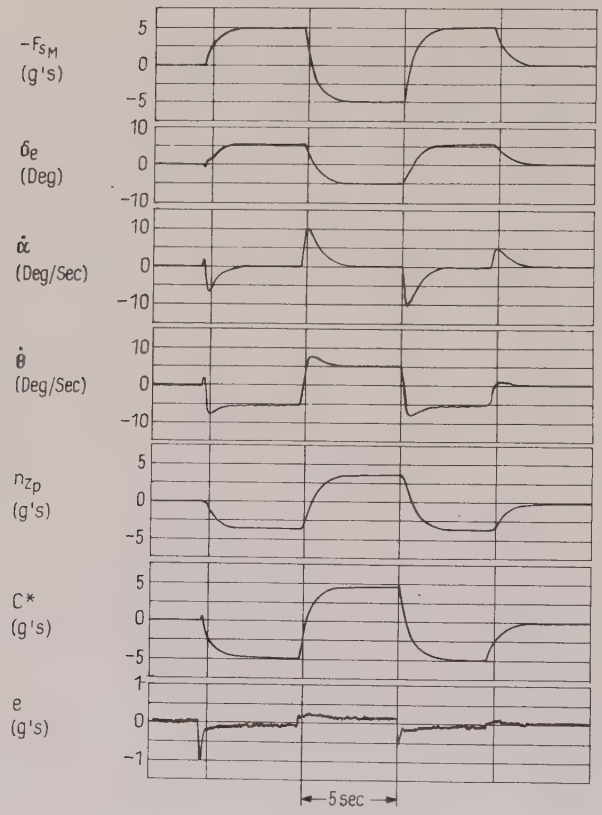


Fig. 28. SOC Configuration D, F-101 B
20,000 ft. Mach 1.4



Sensor Noise Effects

Fig. 29. SOC Configuration A, F-101 B
10,000 ft. Mach 1.2

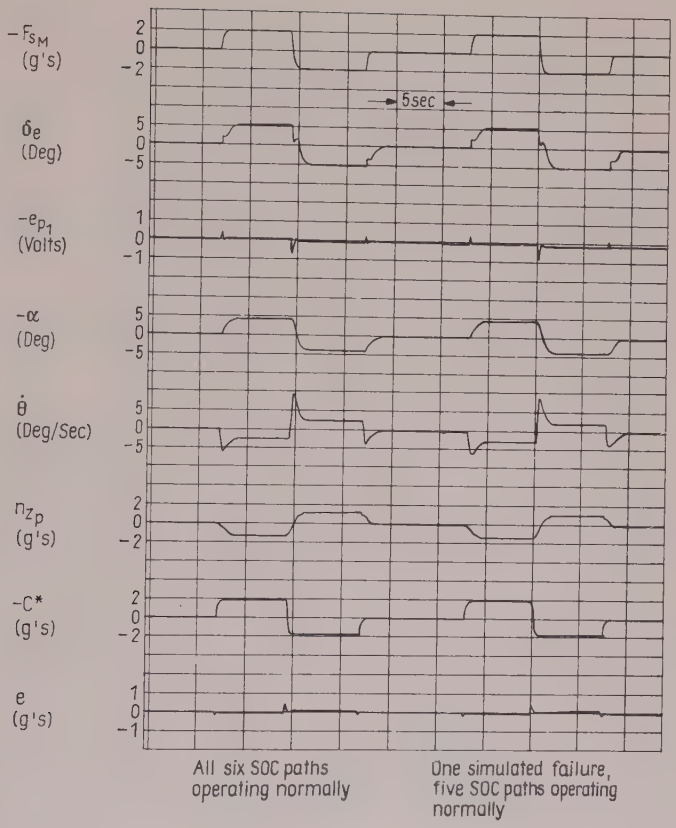


Fig. 30. SOC Configuration G, F-101 B
35,000 ft. Mach 1.0

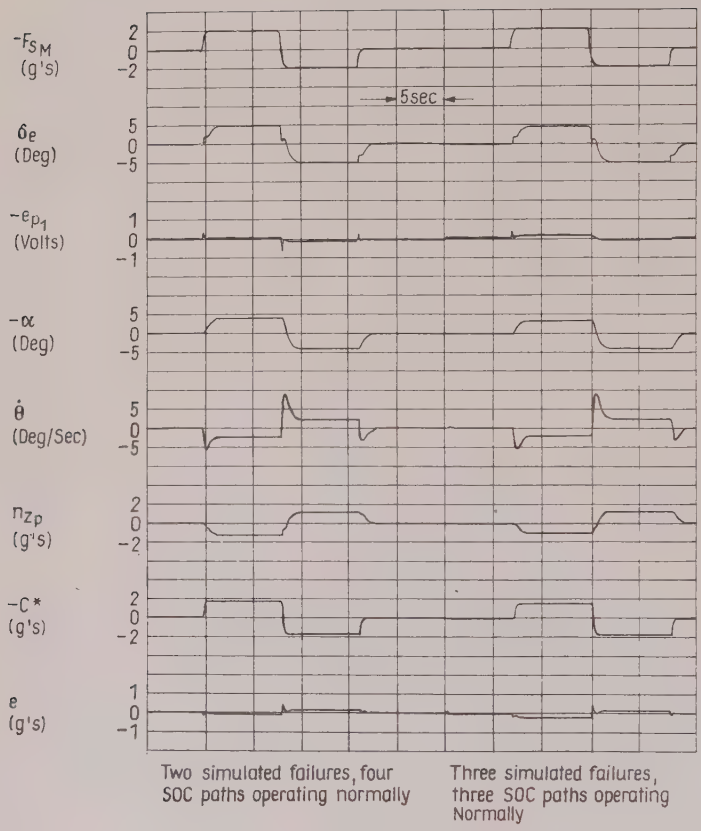


Fig. 31. SOC Configuration G, F-101 B
35,000 ft. Mach 1.0

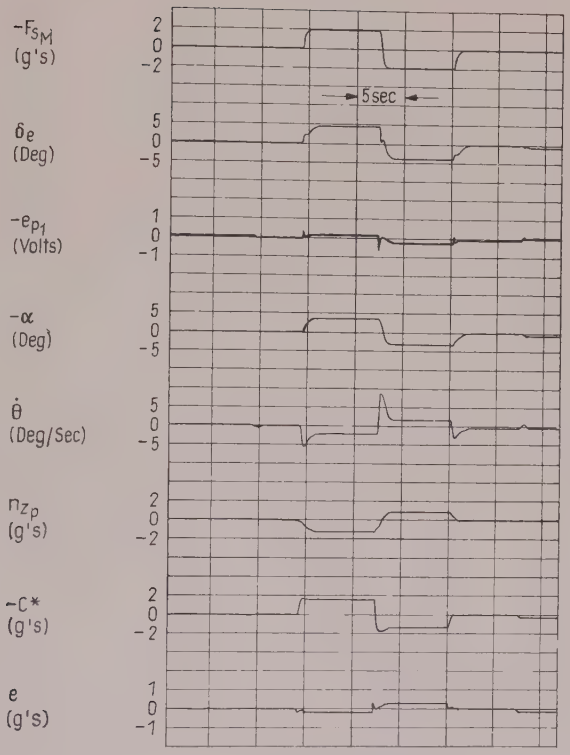


Fig. 32. SOC Configuration G, F-101 B
35,000 ft. Mach 1.0

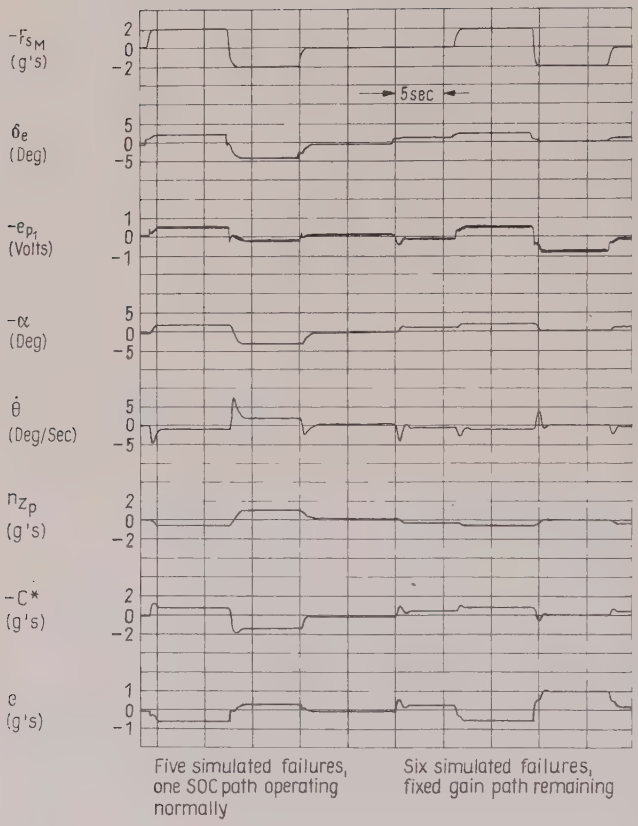


Fig. 33. SOC Configuration G, F-101 B
35,000 ft. Mach 1.0

2g input, and in Figure 26, the input was switched between plus and minus 2g's. These figures are additional evidence of the well-behaved response characteristics of the SOC under varying flight conditions.

The self-organizing controller is capable of pitch-rate control as an alternative to C^* control, and typical performance of the SOC in a pitch-rate application is shown in Figures 27 and 28.

Figure 29 was obtained at the peak-q flight condition (Mach 1.2 at 10,000 feet) with 10 millivolts peak-to-peak high-frequency noise added to the error signal in simulation of sensor noise effects. This noise amplitude corresponds to approximately 5 per cent of the system input signal, which was 2 volts during steady-state. We see that performance of the SOC is not materially degraded by presence of the noise.

Figures 30–33 show the ability of parallel PA/PSV networks to compensate for the failure of control logic elements. In Figure 30, the system consisting of six PA/PSV paths (the output being added to the output of a conventional controller), is initially operating normally. On the right-hand side of the figure, the system closed-loop responses are shown during and after the simulated failure of one of the PA/PSV paths. Because the clock rate of the SOC logic is of the order of 33 microseconds, and because the system can re-organize its probability state in approximately 5–10 clock intervals, no instability resulted from the simulated failure. In Figures 31–33, additional PA/PSV paths were failed as indicated in the figures. We see in Figure 33 that performance was degraded in the course of these simulated failures; however, at no point did the system go out of control. In fact, at the end of the experiment, closed-loop performance of the control system was more or less comparable to that obtainable using the conventional controller working alone.

The nomenclature "SOC Configuration *A*", etc. is defined in reference 30.

The second area of SOC application, viz., attitude control and stabilization of orbiting satellites, has been the subject of a considerable amount of inquiry.[27, 28, 29, 34] References 27, 28, and 29 describe the outcome of experiments with two predecessor SOC implementations (Marks I and II) connected in single-axis, momentum-wheel control loops. This earlier work was conducted using a form of the Type 1 PA criterion and demonstrated the positive potential of self-organization techniques for spacecraft applications. More recently, Fisher and Birchard[34] have applied the Mark III SOC to multiple-axes satellite control with considerable success.

Additional areas of Mark III SOC applications work have included aircraft throttle control[35] and the stabilization of large, flexible space launch vehicles (the latter being investigated by the George C. Marshall Space Flight Center, NASA).

Gouge[36] has formulated reliability prediction techniques for SOC networks of PA and PSV elements. He shows that a network having four each Type 2 PA elements and four each general-purpose PSV modules has a theoretical probability of failure of 1.11×10^{-6} for two-hour airborne missions. This failure probability is achieved principally via the self-organization capabilities of the network. Still lower failure probabilities could result if component-redundancy techniques were employed.

LOOKING AHEAD

Based on present trends in self-organizing and learning control systems R&D, the author forecasts the following:

1967—Improved mathematical and statistical models for self-organizing and learning control systems will be available, greatly facilitating analyses of applications. The role of memory in learning controllers will be more rigorously defined. Decision-theoretic techniques will be directed toward detection problems in high-speed, on-line performance assessments. Theoretical and simulation studies of advanced multi-degree-of-freedom SOC applications will be performed. Flight testing of a self-organizing controller will be conducted to validate its potential performance and reliability pay-offs.

1968, 1969—Extensive design and simulation studies will be made of learning control systems using long-term memory. The non-aerospace applications of self-organizing and learning control systems will develop swiftly. Flight tests of multi-degree-of-freedom self-organizing controllers will be conducted in such areas as (i) lateral/directional control augmentation for high-performance aircraft and (ii) use of SOC distributed-actuation control in spacecraft control-moment-gyro and magnetic-torquer systems.

1970, 1971—Techniques for adaptation of performance assessment criteria will be developed to permit on-line learning of individual pilot preferences in manned systems. In other words, the control will adapt itself to the man, so as to have a consistently good "feel". Operational

experience will be gained with the more elementary applications of self-organizing and learning control systems. Advanced systems will move from test and evaluation stages into serious contention for operational roles. Aerospace applications will continue to pace R&D in this field, but widespread applications to manufacturing processes appear likely. Adaptive control systems will have become a mature technology.

ACKNOWLEDGEMENTS

The author acknowledges the many important contributions of his co-workers at Adaptronics, Inc. Particularly, Messrs. L. O. Gilstrap, Jr., S. Schalkowsky, R. M. McKechnie, III, R. F. Snyder, J. M. Davies, H. J. Cook, C. W. Armstrong, R. E. J. Moddes, N. E. Wilson, and S. Levine, Miss J. B. Hauptli, and Mrs. M. C. Collins have made this work possible. The contributions of Mr. R. J. Lee in earlier phases of the effort are also acknowledged. Mr. P. E. Blatt, Air Force Flight Dynamics Laboratory, has provided valuable guidance and suggestions throughout the development of the present techniques. Messrs. C. W. Gwinn, D. R. Moore, and A. C. Speake, Air Force Avionics Laboratory, have been extremely helpful in the formulation of advanced concepts, providing both support and encouragement in this endeavor. Mr. J. Gatlin, Goddard Space Flight Center, and Mr. L. Griner, Marshall Space Flight Center, NASA, have participated actively in important aspects of the effort. Dr. E. S. Gwathmey, Mr. R. J. Book, and Mr. F. Hayes of Automated Specialties Division, Teledyne, Inc., have stimulated the development of aircraft throttle control applications.

REFERENCES

1. Ostgaard, M. A., and Butsch, L. M., "Adaptive and self-organizing flight control systems," *Aerospace Engineering*, Vol. 21, No. 9, p. 80 et seq., Sept. 1962.
2. Mishkin, E., and Braun, L. Jr., *Adaptive Control*, McGraw-Hill Book Company, Inc., 1961, Chapter 1.
3. Box, G. E. P., and Wilson, K. B., "On the experimental attainment of optimum conditions," *J. Royal Stat. Soc.*, Series B, Vol. 13, 1951, pp. 1–45.
4. Box, G. E. P., "The exploration and exploitation of response surfaces: some general considerations and examples," *Biometrics*, Vol. 10, 1954, pp. 16–60.

5. Box, G. E. P., "Evolutionary operation: a method for increasing industrial productivity," *Appl. Stat.*, Vol. 6, 1957, pp. 88–101.
6. Box, G. E. P., and Hunter, J. S., "Multifactor experimental designs for exploring response surfaces," *Ann. Math. Stat.*, Vol. 28, 1957, pp. 195–241.
7. Brooks, S. H., "A discussion of random methods for seeking maxima," *Operations Research*, Vol. 6, 1958, pp. 244–251.
8. Brooks, S. H., "A comparison of maximum-seeking methods," *Operations Research*, Vol. 7, 1959, pp. 430–457.
9. Draper, C. S., and Li, Y. T., "Principles of optimalizing control systems and an application to internal combustion engine," *ASME Publications*, 1951.
10. Ashby, W. R., *Design for a Brain*, J. Wiley and Sons, Inc., 1952.
11. von Neumann, J., "Probabilistic logics," *Automata Studies*, Princeton Univ. Press, 1956, pp. 43–98.
12. McCulloch, W. S., and Pitts, W., "A logical calculus of the ideas immanent in nervous activity," *Bull. of Math. Biophys.*, Vol. 5, 1943, pp. 115–133.
13. Lee, R. J., *Self-Programming Information and Control Equipment*, Melpar, Inc., 1959.
14. Sklansky, J., "Learning systems for automatic control," *IEEE Trans. on Automatic Control*, Vol. AC-11, 1966, pp. 6–19.
15. Barron, R. L., "Parameter Space Search Techniques for Learning Automata," Adaptronics, Inc. paper delivered at 1966 Bionics Symposium, Dayton, Ohio, 3–6 May 1966.
16. Moddes, R. E. J., and Gilstrap, L. O. Jr., "Research on Optical Modulation and Learning Automata," in *Optical and Electro-Optical Information Processing*, James T. Tippett *et al.*, Eds., M. I. T. Press, 1965, pp. 491–522.
17. Gilstrap, L. O., Jr., "Pattern Recognition by Transformational Automata," Adaptronics, Inc. paper delivered at 1966 Bionics Symposium, Dayton, Ohio, 3–6 May 1966.
18. Gilstrap, L. O., Jr., "Parameterization and Performance Criteria in Transformational Automata," Adaptronics, Inc. paper delivered at 1966 Bionics Symposium, Dayton, Ohio, 3–6 May 1966.
19. Wilde, D. J., *Optimum Seeking Methods*, Prentice-Hall, Inc., 1964, pp. 62–64.
20. Rastrigin, L. A., "The convergence of the random search method in the extremal control of a many-parameter system," translated from *Avtomatika i Telemekhanika*, Vol. 24, 1963, pp. 1467–1473.
21. Matyas, J., "Random optimization," translated from *Avtomatika i Telemekhanika*, Vol. 26, 1965, pp. 246–253.
22. Snyder, R. F. *et al.*, *Advanced Computer Concepts for Intercept Prediction*, Adaptronics, Inc. Final Technical Report under Contract DA-36-034-AMC-0099Z, Nike-X Project Office, U.S. Army Materiel Command, 1964.
23. Moddes, R. E. J. *et al.*, *Study of Neurotron Networks in Learning Automata*, Adaptronics, Inc., AFAL-TR-65-9, AF Avionics Laboratory, RTD, AD 455 688, 1965.
24. Sabroff, A. *et al.*, *Investigation of the Acquisition Problem in Satellite Attitude Control*, TRW/Space Tech. Lab., AFFDL-TR-65-115, AF Flight Dynamics Laboratory, RTD, 1965, pp. 156–159.

25. Bellman, R., *Adaptive Control Processes: A Guided Tour*, Princeton Univ. Press, 1961, pp. 94–95.
26. Lee, R. J., and Snyder, R. F., *Functional Capability of Neuromime Networks for Use in Attitude Stabilization Systems*, Adaptronics, Inc., ASD-TDR-63-549, AD 429 116, 1963.
27. Barron, R. L *et al.*, *Self-Organizing Spacecraft Attitude Control*, Adaptronics, Inc., AFFDL-TR-65-141, AF Flight Dynamics and Avionics Laboratories RTD, AD 475 167, 1965.
28. Barron, R. L. *et al.*, "Self-organizing adaptive control of aerospace vehicles," *Proc. 17th Ann. NAECON*, Dayton, Ohio, May 10–12, 1965, pp. 468–474.
29. Barron, R. L. *et al.*, "Self-organizing adaptive systems for space vehicle attitude control," *Proc. AIAA/ION Guidance and Control Conference*, Minneapolis, Minn., August 16–18, 1965, pp. 163–170.
30. Barron, R. L. *et al.*, *Self-Organizing Control of Aircraft Pitch Rate and Normal Acceleration*, Adaptronics, Inc., AFFDL-TR-66-41, AF Flight Dynamics and Avionics Laboratories, RTD, 1966.
31. von Foerster, H., "Quantum mechanical theory of memory," *Trans. of the Sixth Conf. on Cybernetics*, Josiah Macy, Jr. Foundation, N. Y., 1950, pp. 112–145.
32. Kalman, R. E., and Bertram, J. E., "Control system analysis and design via the second method of Lyapunov I", *Trans. ASME, J. Basic Eng.*, Vol. 82, 1960, pp. 371–393.
33. Malcolm, L. G., and Tobie, H. N., *New Short Period Handling Quality Criterion for Fighter Aircraft*, Boeing Company Document D 6-17841 T/N, 1965.
34. Fisher, J. (Capt., USAF), and Birchard, C. (Capt. USAF), Master's Thesis submitted to AF Institute of Technology, Wright-Patterson Air Force Base, Ohio, 1966.
35. Barron, R. L. *et al.*, *Application of Self-Organization Techniques to the Auto-Throttle Controller*, Adaptronics, Inc. Final Technical Report under Purchase Order No. 01118, Teledyne, Inc., Automated Specialties Division, Charlottesville, Virginia, 1966.
36. Gouge, J. R., "Reliability Prediction for Networks of Probability State Variable Devices," Bionics Symposium, 1966. This volume pp. 655–672

KEITH E. JUSTICE*

and JUDY M. GERVINSKI

Melpar, Inc.

Falls Church, Virginia

Electronic Simulation of the Dynamics of Evolving Biological Systems†

ABSTRACT

A bionic investigation and modeling of organic evolution is described. The project was undertaken in order to provide a deeper understanding of the adaptive processes involved in organic evolution. Of particular interest was a comparison of self-organizing processes in evolutionary systems and analogous processes in Trainable Logical Networks.

The biological prototype for the model is the feral house mouse (*Mus musculus*) as it exists in semi-isolated populations in the southwestern United States. Special emphasis is given to a balanced lethal genetic system known to exist in the species. Using Montè Carlo techniques, the model simulates, for each individual, such events as the probability of survival, migration, mating, reproduction, mutation, genetic segregation, and natural selection. Implementation of the model on a digital computer is described.

Results of experiments performed with the model show that the model behaves in a manner highly analogous to both the biological prototype and to certain aspects of Trainable Logical Networks. Implications of the work for future developments in machine intelligence are discussed.

INTRODUCTION

Primitive learning in biological organisms frequently involves an organizational process called "trial and error" learning. In this learning process, more or less random reactions initially result from an external

* Now at the University of California, Irvine, California.

† This work was conducted by Melpar, Inc., under Contract AF 33(615)-2456 sponsored by the Air Force Avionics Laboratory, Research and Technology Division, Air Force Systems Command, United States Air Force.

(or internal) stimulus. Those reactions which tend to move the organism towards a desirable state are "reinforced." That is, they will be more apt to occur in the future if the same or a similar stimulus is presented. Those reactions which move the organism away from the desirable state tend to be "extinguished." That is, they will be less apt to occur in the future under the same stimulus conditions.

The existence and success of this organizational process in individual living organisms has prompted the construction of several electronic models of the process to be used in investigations of machine intelligence.[1,2] These devices have been given a variety of names: ARTRON (ARTificial neuRON), Probability State Variable Network, Self-Organizing Binary Logical Network (SOBLN), Trainable Logical Network (TLN), etc. In general, they are characterized by the ability to form logical transformations from digital (usually binary) inputs to digital outputs, the connectives being formed in a random manner when the device is in an unorganized state. A goal circuit evaluates the results of the output with respect to the system goals and biases the probability state of the device towards those connectives which tend to move the system toward the goals, and away from those connectives which move the system away from the goals. Thus these devices show random search, reinforcement, and extinction in a manner analogous to primitive learning in individual organisms.

Another biological learning system which utilizes the trial-and-error process is organic evolution through natural selection. A cursory examination of the process of organic evolution reveals the same basic elements and functions noted above for trial-and-error learning in the individual: genetic mutation and recombination (random reactions), survival and reproduction of desirable genotypes (reinforcements), death and failure to reproduce undesirable genotypes (extinction).

Thus we can recognize three broad categories of trial-and-error learning systems: Psychogenic, as exemplified by primitive learning in individual organisms; Bionic, as exemplified by the ARTRON and subsequent elaborations of the Trainable Logical Network; and Evolutionary, as exemplified by organic evolution.

While all three categories have been subjected separately to considerable experimentation, analysis, and theoretical investigation, and although individual learning has received considerable attention by engineers, there appears to be a paucity of investigations of the process of organic

evolution from the bionics standpoint. Such investigations would seem particularly appropriate since the learning capabilities of the individual are derived ultimately from the much broader adaptive resources of the evolving genetic system.

These facts have prompted an investigation in depth of the theoretical principles as well as of concrete examples of the process of organic evolution with the purpose of gleaning therefrom facts and principles that may advance the technology of machine intelligence. This work, which began in June 1965, has used what might be called the "classical bionics approach," that is:

1. Description of the biological prototype,
2. Derivation of a logical or mathematical model,
3. Implementation of the model,
4. Experimentation with and analysis of the model, and
5. Derivation of principles of engineering significance.

During the first eight months of effort on the project, progress has been made in these five areas. The results are reported below.

BIOLOGICAL PROTOTYPE AND DIGITAL IMPLEMENTATION

The biological prototype chosen was that of the feral house mouse (*Mus musculus*) as it exists in semi-isolated populations in the southwestern United States.[3,4,5] This choice was based on the fact that considerable information is available on the population dynamics and genetics of the species. Further, it is known that homozygous lethal genes exist in high frequencies in populations of house mice in the U.S.[6] The gene is maintained at high frequencies, despite lethality in the embryonic stage, by compensating selection in favor of the gene in the gametic stage of life. That is, spermatozoa which carry the gene are overwhelmingly favored over normal sperm in the female reproductive tract, due to some as yet poorly understood physiological mechanism. Since this provides an opportunity for investigation of natural selection operating on both the gametic (germinal) and the zygotic (somatic) stages, it was chosen as the genetic prototype for the first modeling.

During any increment of time in a natural population of feral house

mice, each member of the population is subject to the probability of one or more "vital events" of importance to the model. These are:

A. Survival
B. Migration
C. Mating and Reproduction
D. Mutation (during reproduction only)

The probability of each of these events occurring in a given interval of time is a function of a number of factors, some associated with the individual, some with the population, and some with the external environment. The basic scheme of the model is to "exercise" in a Monte Carlo manner the probabilities of the above events for each individual during each increment of time.

A schematic representation of this process is shown in Figure 1, which also served as the basis for the digital computer implementation. In each increment of time (Δt) each individual in a population is exercised through the indicated logic. The most appropriate value for the basic increment of time appears to the gestation period, which is equal to 20 days in this species. The model will accommodate up to eight populations with as many as 200 individuals in each population.

The first decision which is made is whether the individual is to survive or not during the next Δt. This probability of survival can be expressed as a product.

$$P\{S\} = P\{S_a\} \cdot P\{S_s\} \cdot P\{S_g\} \cdot P\{S_d\} \cdot P\{S_c\}$$

where, $P\{S_a\}$, the age specific survival factor, is an empirical function of age; $P\{S_s\}$ is a function of sex; $P\{S_g\}$ is a specified function of genotype; $P\{S_d\}$ is the survival factor due to population density (equal to $1 - N/K_1$ where K_1, the carrying capacity, is a random variable with specified mean and variance, and N is the population size); and $P\{S_c\}$ is a scalar factor which can be used to control the overall survival rate of the population.

The computer program computes $P\{S\}$ for each individual and a decision is made as to whether the individual lives or dies. If it dies, it is deleted from the population of which it is a member and its storage cell becomes available for assignment to a newly arrived member of the population. Also on the death of an individual, related "interaction" variables such as population density and gene frequency are updated.

Fig. 1. Main Loop of Simulation Program

Fig. 1. Main Loop of Simulation Program

If the individual does not die, it is then exercised according to the probability of migration.

The probability that an individual will migrate during an increment of time can be described as follows:

$$P\{M\} = P\{M_a\} \cdot P\{M_s\} \cdot P\{M_d\} \cdot P\{M_c\} \cdot P\{M_r\}$$

where, $P\{M_a\}$ is an empirical function of age, $P\{M_S\}$ is a function of sex,

$$P\{M_d\} = 1 - \frac{K_2 - N}{K_2}$$

where again K_2 is a random variable with specified mean and variance, $P\{M_c\}$ is the motility constant for the species, and $P\{M_r\}$ is a two-valued function which recognizes the lower motility of the "established residents." A Monte Carlo decision is made on the basis of this probability.

Given that an individual is going to migrate in a Δt, the population which receives this individual must be determined next. In general, this will be a function of, (1) the location of the recipient population with respect to the donor population and, (2) the relative population densities of the recipient populations. The function relationship due to location can be established mathematically. The population density effect is implemented simply by allowing the individual to enter the donor population, but not allowing it to "establish residency" unless it stays a specified length of time, normally one Δt.

The probability of reproducing is computed for females only, where

$$P\{R\} = P\{R_a\} \cdot P\{R_d\} \cdot P\{R_c\}.$$

$P\{R_d\}$ is the age specific reproduction function,

$$P\{R_d\} = \frac{K_3 - N}{K_3},$$

and $P\{R_c\}$ is the fertility constant for the species.

When it has been established that a given female will reproduce in a Δt, a mate is selected randomly from the adult males in the population. If no adult males exist in the population, the female is passed over until the next Δt. The population is not taken to extinction at this time because of the possibility of immigration of a male before the death of all females.

The litter size is determined by a random variable drawn from a normal distribution with a specified mean and variance. After litter size has been determined, the specified number of progeny is generated by Monte Carlo methods, taking into account the genotypes of the parents, the probability of mutation, and the segregation ratios (gametic selection forces) specified for each gene pair.

EXPERIMENTAL INVESTIGATIONS

Although the experimental phase of the program is by no means complete, a number of simulations have been conducted. Since space limitations preclude a full treatment of this work, only a few experiments will be described. Experiments selected for description are those which illustrate the more prominent features of the model.

1. Growth of Confined Populations

The increase in numbers of a confined population frequently follows a characteristic time function known as the logistic growth function. Figure 2 illustrates this growth as observed in an actual biological experi-

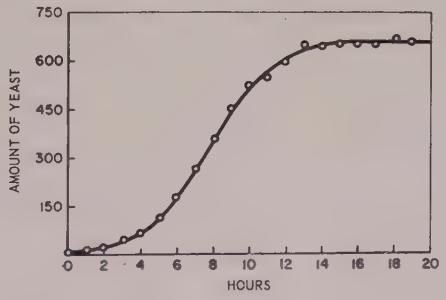


Fig. 2. The Logistic Growth of a Laboratory Population of Yeast Cells (From Ref. 7)

ment.[7] As can be seen, the initial rate of increase is initially low due to the small size of the parent population. As the size of the parent population increases the rate of growth increases accordingly. However, as the population size approaches the carrying capacity of the environment, the rate of increase falls off due to the density dependent reproduc-

tion and survival factors. (See pages 208 and 210 for the mathematical formulation of these factors.)

Figure 3 shows the results of an experiment in which the growth of the populations followed the typical logistic function with minor exceptions. Each subpopulation was initiated with two males and two females. Note that during the first seven time intervals the rate of increase

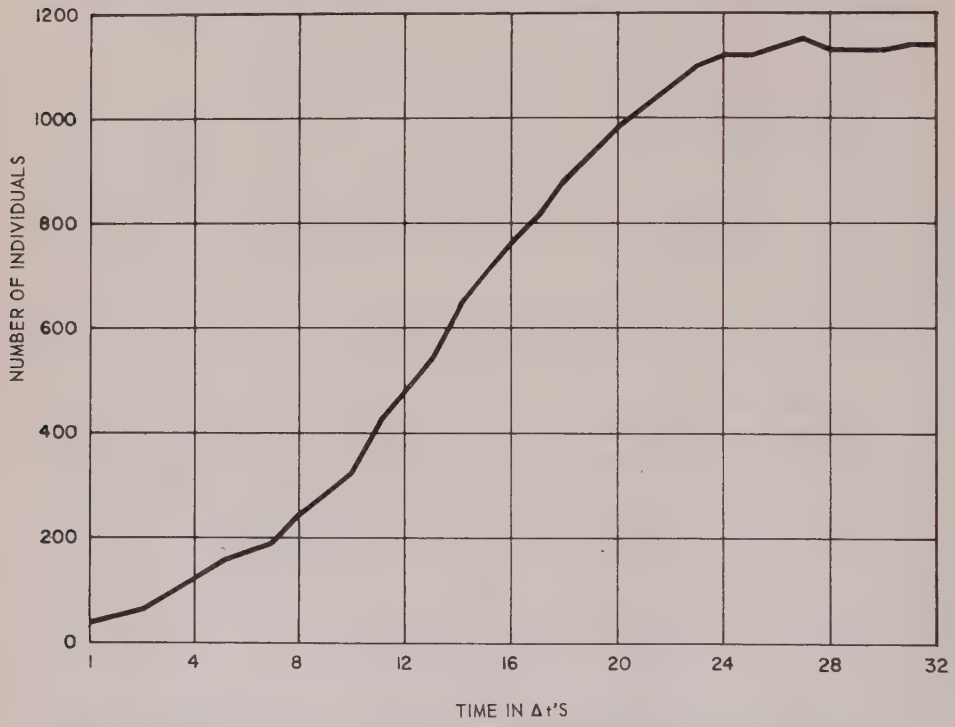


Fig. 3. Typical Logistic Population Growth in the Model with Normal Period of Reproductive Immaturity. Number of Individuals Summed over 8 Populations. One $\Delta t = 20$ days of Simulated Time.

of the population remains essentially the same in spite of the fact that the population size is increasing. This is because the newborn animals do not begin to reproduce at a significant rate until they are several time intervals old. This time lag introduces a slight discrepancy from the theoretically ideal logistic growth function. In a later experiment the time lag due to the period of reproductive immaturity was eliminated by changing the age specific reproduction function. Here the organisms

reached full reproductive maturity in the interval immediately following their birth. A more typical logistic function resulted since the rate of increase of the population responded immediately to the increase in the size of the parent population.

The exception from the theoretically ideal logistic growth function should not be considered as an artifact of the simulation, but rather it is precisely what could be expected in a living population under similar conditions of growth rate, time to reach reproductive maturity, and environmental carrying capacity.

2. Growth in an Unconfined Population and the Establishment of New Populations through Migration

This experiment is intended to show the effect of migration on the growth of a population and the manner in which new populations may become established through migration. Population No. 0 was initially

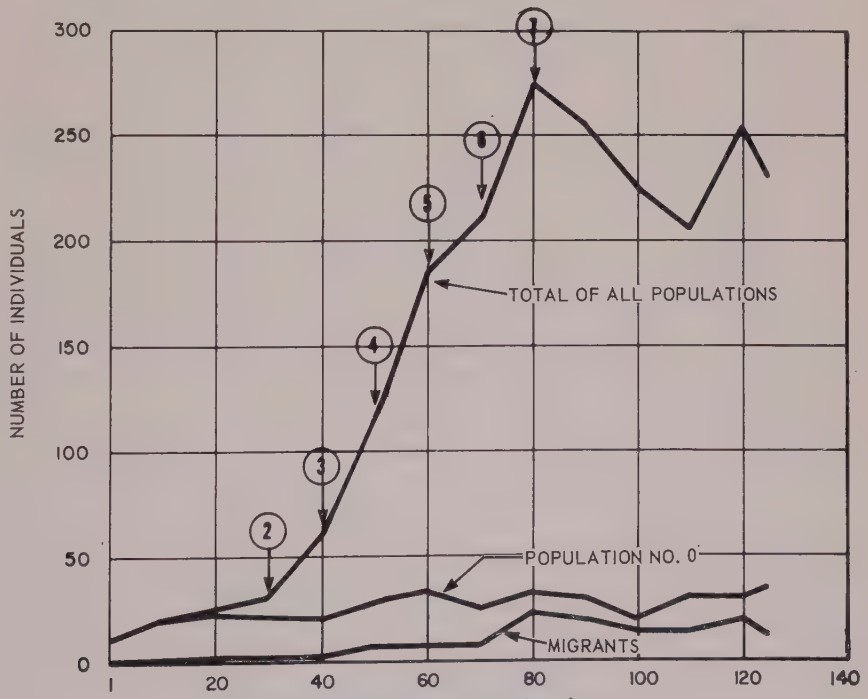


Fig. 4. Population Growth Migration.

established with 4 males and 4 females each, and the remaining seven populations were left vacant. Results of the experiment are shown in Figure 4. Numbers over the top curve indicate the number of populations actively reproducing at that time. Eighty time periods were required before all eight populations were actively reproducing. The experiment established that the model behaves as predicted with respect to population growth of unconfined populations.

3. Genetic Equilibrium in a Balanced Gene System

The experiments described above were conducted with populations consisting of all homozygous individuals. That is, the effects of genetic diversity were eliminated in those experiments. In the experiments described below genetic diversity was introduced, thus allowing natural selection to come into play.

As described on pages 207–210 above, the genetic system modeled is a balanced lethal system maintained by the interaction of positive gametic selection and negative zygotic selection. It can be shown[8] that this system will approach a stable gene frequency equilibrium, the value of which depends upon the degree of positive gametic selection. To demonstrate this phenomenon, an experiment was performed where the populations were initiated with 5 males and 5 females each and where the t^w gene frequencies of the populations varied from 0 to 50%. (The symbol + is used to signify the normal gene and t^w to signify the lethal gene.) Figure 5 shows the approach of the gene frequency to the predicted equilibrium of 40%. Other experiments using different values of gametic selection have demonstrated that the model behaves as predicted on the basis of the biological prototype.

FUNCTIONAL EQUIVALENCES

The current study and modeling of evolutionary dynamics from the bionics standpoint was prompted by certain similarities that were noted in the mathematical theories of evolving genetic systems and self-organizing binary logical networks.

To elucidate further the relationship between these two types of self-

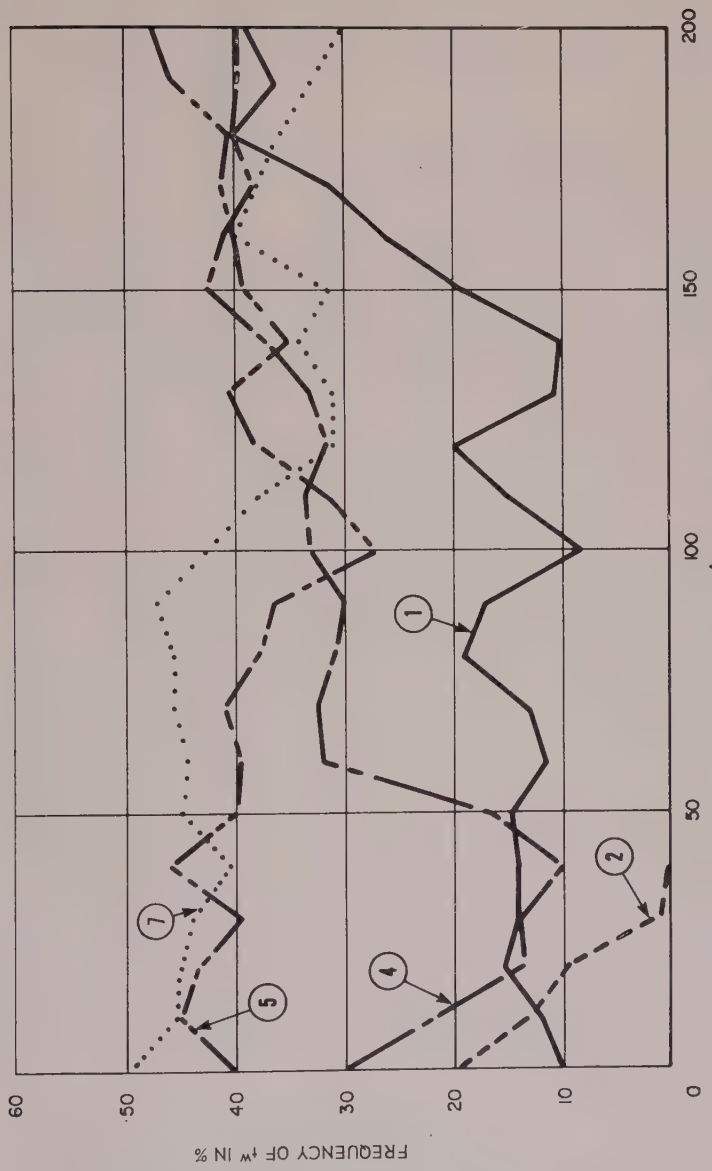


Fig. 5. Gene Frequency Trajectories Versus Time.

organizing systems, the functional equivalences between them must be identified. This section will attempt to identify elements in each system that have the same or equivalent functions, and will compare and contrast the operation of these functional equivalents.

1. Review of Self-Organizing Binary Logical Network (SOBLN)

The SOBLN concept is a generalization of the original ARTRON concept developed at Melpar, Inc. under Air Force support. A SOBLN is a probability state variable device that can be described at any time by its probability state and its logical state. The logical state assumed on any "trial" is the result of one or more random decisions with probabilities determined by the probability state of the network. The probability state is modified in turn, depending upon the interaction of the logical state with environmental variables and a "goal circuit." These relationships are shown schematically in Figure 6.

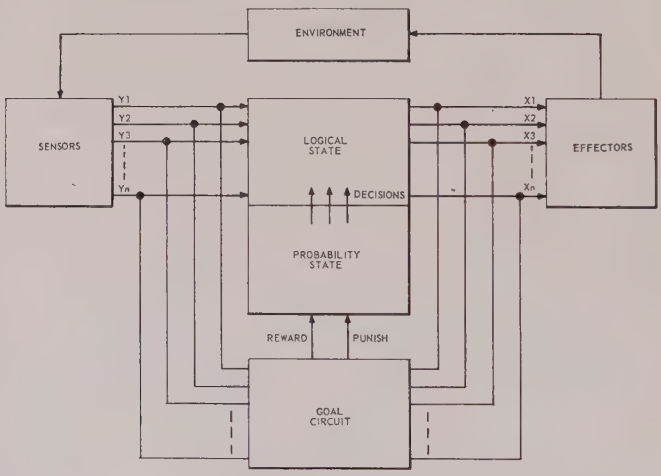


Fig. 6. Schematic Representation of Self-Organizing Binary Logical Network.

The logical state of the network serves to transform a particular binary input vector $Y_1 - Y_m$ derived from the sensors into a particular binary output vector $X_1 - X_m$. The output vector in turn acts on the environment through the effectors to produce a (hopefully) desirable change.

The "goodness" of the decision is determined by the goal circuit, either by looking at the binary input and the resultant output, or by looking at the change in the environment-machine relationship as a result of an output. If the decision (logical state) was "good," the probability state is biased in favor of that state (reward). If it was "bad," the probability state is biased away from that logical state.

The network is composed of elements called "statistical switches". Only a small portion of the total complement of statistical switches may contribute to a given reaction. The input vector can be thought of as an address that accesses a particular group of statistical switches and their probabilities. Generally, the only probability states modified by the reward/punish signals from the goal circuit are those controlling switches that actually participated in the prior decision.

2. Review of Evolutionary Concepts*

Under the Neo-Darwinian concept of organic evolution, the basic unit of evolution is the population of interbreeding organisms. Evolution consists of changes in gene frequencies within populations. These changes are brought about by so-called evolutionary forces, acting singly or in combination. These forces are:

1. Zygotic Selection
2. Gametic Selection
3. Mutation
4. Migration
5. Genetic Drift

Zygotic selection operates through genetically based differences in vitality and fertility of individuals in the populations. If individuals of one genotype contribute more offspring, on the average, to the next generation than individuals of another genotype, then the genes represented by the first genotype will tend to increase in frequency with time.

Gametic selection operates in a similar manner to zygotic selection, but it acts upon germ cells rather than individuals. If genetically based

* The discussion in this section assumes a basic knowledge of Mendelian genetics. For a concise tutorial treatment see chapter 2 of "The Process of Evolution," Paul R. Ehrlich and Richard W. Holm, (McGraw-Hill Book Co.), 1963.

differences in the virility of spermatozoa or ova exist, the more virile sperm or eggs will tend to propagate their own genotypes and this will cause a corresponding shift in gene frequencies. An example of this is the t^w allele, which has a high gametic selective advantage over the normal (+) allele in the sperm of house mice. The so-called "Meiotic Drive" phenomena are also classified under gametic selection. In mammals, where the gametes are normally continually in the protected environment of the male or female reproductive tract, there can be little impact of the "external" environment through gametic selection. However, in lower organisms where the haploid stage frequently has an independent existence for a period of time, gametic selection must play a prominent role in the continued adaptation of the species under changing environmental conditions.

Whereas selection may be considered a directing process that tends to drive gene frequencies towards a more adaptive condition, mutation may be thought of as a random or noise-like process that is nonadaptive in nature. However, mutation is necessary to the overall process of evolution, since it provides the variability upon which selection operates. Mutation as such is not qualitatively responsive to the external environment, but it may be quantitatively responsive insofar as increased exposure to ionizing radiation will increase mutation rates.

Migration affects gene frequencies under conditions where local differences in environmental conditions have established differences in gene frequencies of subpopulations. In such cases, migration between subpopulations tends to prevent the local gene frequencies from reaching optimal levels for the immediate environment; however, migration provides the variability that allows the local population to react quickly to changes in environmental conditions. Unlike the variability provided by mutation, which is unselective or completely random in nature, the variability provided by migration is more apt to be of the "right" kind, since it is of a type which has been useful in adjacent geographical areas.

Genetic drift is a force that is effective only in breeding populations of small size. It is caused by the random sampling effects in population replacement or turnover. Since it is random in nature, it tends to drive gene frequencies away from the equilibrium points determined by natural selection. It contributes, however, to the overall evolutionary process, since it causes "hunting" of gene frequencies. This may allow the evo-

lutionary system to escape from local maxima in the adaptive surface and thus move to other higher maxima.

As a result of mathematical analyses made by Sewell Wright, J.B.S. Haldane, and others (primarily during the 1930's and 1940's), it is generally agreed among biologists that evolution is most rapid in a species that exists in numerous small populations partially isolated from one another, thus allowing a rich interplay of the forces of selection, mutation, migration, and genetic drift. This condition prevents genetic stagnation and creates a dynamic evolutionary system. These facts should be considered carefully in the design of self-organizing machines based on evolutionary principles.

3. Basic Equivalences

Several approaches may be used in drawing an analogy between organic evolution and self-organizing machines. These differences arise primarily from what is considered the basic unit of organization in each system. For the analogy summarized in Table I, we have chosen the population as the basic unit for evolutionary systems and consider this as analogous to the self-organizing networks in self-organizing machines. Just as a self-organizing machine may be made up of more than one network, frequently termed subnetworks, so also may a species be made up of more than one population, usually termed subpopulations. (The terms "population," "subpopulation," "network," and "subnetwork" are used rather loosely here, but the context should make the relationships clear in each case.) The discussion which follows elaborates further on the analogies summarized in Table II.

The probability state of the network is analogous to the gene frequency array of a biological population. The probability that a given logical state will be assumed on a certain "trial" of the network is specified by the probability state of the network just prior to the trial. Similarly, the probability that a given genotype will occur in a certain individual will depend upon the gene frequency array in the population just prior to the mating that produced the individual. Thus, the individual is equivalent to a "trial" (i.e., decision) of the network. There is, however, an important difference: in trainable logical networks as they have been implemented in the past, only one logical state exists at any moment; this is tested against the environment, and reward/punish

Table I: Comparison of Self-organizing Network and Evolving Genetic System

Self-Organizing Network	*Evolving Genetic System*
Probability state of the network	Gene frequency array
Logical state of the network	Genotype of individual
Range of counter states	Population size
Goal circuit function	Natural selection
Reward/punish value	Selective value and population size
"Learning curve"	Dependence of selective value on gene frequency
Training rules	Mode of expression of traits (i.e., dominant, recessive, etc.)
Statistical dependency between switches	Genetic linkage
Avoiding absorbing states by leaving a residual probability even when the counter is full	Gene mutation
Increasing reliability with redundancy	Subdivision of a population into subpopulations
Transfer of learning between redundant networks	Migration between subpopulations
Search for new solutions by random trial and error	Genetic drift

signals are generated, etc. In biological populations, many "trials" (i.e., individuals) may exist simultaneously, and these frequently compete against one another for resources of the environment. The goal circuit, in the form of natural selection, can then work on a comparative basis to select the better logical states to be rewarded (propagated).

In addition to allowing this comparative operation of the goal circuit, multiple simultaneous trials also speed up the evolutionary process under certain conditions. Because of the "time constants" of life, it takes a certain amount of time to produce an individual and to test this individual against its environment. Since a species exists as a population of individuals rather than as one individual, many more trials of the myriad of possible genotypes may be made in a given length of time than if only one individual existed at a time. A similar strategy might be used in applications of trainable logical networks to systems with long time constants.

In a trainable logical network, the logical state is expressed as a logical connective between the input from the environment and the output

to the environment. In the biological system, the genotype of the individual is impinged upon by the environment and expresses itself in the individual phenotype. This phenotype interacts with the environment in a unique manner determined by its input (previous and immediate experience) and its logical state (genotype). The interaction of the phenotype with its environment is bidirectional: on the one hand, the individual may alter its environment in a way that substantially changes the relationship of other individuals of the same or other species; and on the other hand, the immediate environment to a great degree determines whether or not the individual will successfully reproduce and thereby influence the gene frequency array.

This brings us to what is both the most difficult and the most fundamental aspect of the analogy: the goal circuit. The goal circuit of the trainable logical network expresses a goal or requirement of the engineering system in which it is embedded. It is a generally accepted precept of machine intelligence that the requirements implemented by the goal circuit logic should be as general as possible to ensure a high degree of reliability and adaptability. On the other hand, a goal circuit which is more generalized than necessary to meet all possible contingencies is undesirable, since it is probably less "efficient" than a less generalized goal circuit. We will see below that the goal circuit of organic evolution has a very generalized logic.

The basic logic of the goal circuit requires the continuity of the germ plasm: animals and plants exist in the form that they do today because these are the forms that have given the best protection to the germ plasm; other forms that do not exist today have somewhere or somehow fallen short in their ability to protect and ensure continuity of the germ plasm. The fact that evolution seems to be in the direction of higher complexity and a local decrease in entropy in the biological organisms is a consequence of the physical and chemical laws of nature. In fact this tendency manifests itself first in chemical evolution.

In the earliest stages of the evolution of life, the larger molecules, once formed, are less prone to be broken up by degradative forces, especially when they combine in aggregates with other similar molecules. Thus, we see that the tendency of organic evolution to produce more and more complex organisms is a consequence of the logic of the goal circuit operating under certain basic physical and chemical laws. The logic of the goal circuit itself is simply that the continuity of the germ plasm must be

maintained; that is, any individual (i.e., trial) of such a genotype (i.e., logical state) that it can maintain the continuity of its germ plasm into the next generation is "rewarded," since the presence of its progeny among the next generation will tend to move the gene frequency array (i.e., probability state) of the system in the direction of the genotype of the parent individual.

This is indeed a very fundamental logic. In any bionic investigation of organic evolution, it is important to keep the nature of the goal circuit in mind so that differentiation may be made between those characteristics of the system that are universal to self-organizing systems and those that are peculiar to organic evolution because of the nature of the goal circuit logic.

The functional equivalences discussed above cover the basic aspects of trainable logical networks. That is, we have considered the probability state, the logical state, input from and output to the environment, and the goal circuit. The equivalences described are clear cut and serve to demonstrate the common foundations of the two types of self-organizing systems. Of equal or greater interest from the bionics standpoint, however, are certain features of organic evolution that have ill-defined or no equivalences in trainable logical networks. It is these features which may suggest technological improvements in the design of trainable logical networks.

For example, the equivalence of mutation is somewhat uncertain. On the one hand, it may be considered equivalent to not allowing adsorbing states in a trainable logical network; that is, even though selection may drive a given locus all the way to fixation, back mutation provides the possibility of the locus escaping from fixation and responding to a change in the environment through natural selection. This is equivalent to allowing a statistical switch to respond to reward/punish signals even after it has reached a 1.0 or 0.0 probability state. This is generally done in trainable logical networks unless it is desired to "fix" the network in a certain logical state after it is fully trained.

On the other hand, mutation can be considered as one of the forces that cause "hunting" in the biological system; that is, mutation acting at all of the loci provides the genetic variability which leads to occasional unusual genotypes being produced. This may occasionally produce the extraordinary individual that may be very important to the species in colonizing new territory, etc. This is equivalent to preventing the

probability states of switches from going all the way to 1.0 or 0.0 in a trainable logical network. This causes the network to react occasionally in a manner in opposition to its previous training. This exploratory operation, however, may lead to the discovery of a new reaction pathway, the possibility of which arose after the highly exploratory reactions of the initial training period.

The problem of achieving a balance between the exploratory and conservative functions is one which occurs repeatedly in organic evolution. Certain conservative mechanisms have evolved which ensure that a successful form will be likely to perpetuate itself. Some of these are high stability in molecular structure of DNA, linkage of genes on the same chromosome, and, where it occurs, parthenogenic (i.e., vegetative or asexual) reproduction. These conservative forces are balanced by the exploratory mechanisms of gene mutation, crossing over and recombination between chromosomes, and segregation and recombination of traits in sexual reproduction. The operation of mutation and its bionic significance has already been discussed above. The adaptive significance of sexual reproduction and linkage will be taken up below.

Consider first parthenogenic or vegetative reproduction. In this case, the progeny of an individual is genetically identical to the parent except at those loci where mutations occur during the reproductive process. Since mutation rates are typically on the order of 10^{-5} to 10^{-6}, this means that there is an extremely high parent/offspring correlation in the genetic constitution. Thus, the members of species which reproduce exclusively or primarily by this means show little variability among individuals. In the TLN, this is equivalent to a fully trained network in which each switch retains a residual probability of changing states only of about 10^{-6} to 10^{-5}. It should be pointed out, however, that the number of genes (switches) present in an organism is considerable, so that even with these low residual probabilities each new organism (trial) has a good probability of differing from its parent (previous logical state) in at least one locus. Nevertheless, this is a relatively conservative and inefficient evolutionary system.

Even though a species may consist of millions of living individuals (i.e., simultaneous trials), there is no genetic communication between them. Thus, if a beneficial mutation occurs at one locus in an individual, it may be passed on to the progeny of the individual; and since it is "beneficial", the number of individuals bearing that mutation which have

descended from the common parent will presumably increase relative to other strains within the species. (These groups of organisms with common ancestry are called "clones.") At any time, the species may consist of several successful clones, each propagating its own beneficial mutations. But without sexual reproduction there is no way of uniting the successful mutations of one clone with those of another, except through independent recurrence of the mutation—a very rare event.

Sexual reproduction, then, must have served initially as a means of transferring genetic information between successful clones. The success of this transfer of information as an adaptive mechanism is demonstrated by the almost universal occurrence of sexual reproduction in biological organisms.

The untrained or partially trained TLN of the SOBLN configuration behaves essentially as a sexually reproducing evolutionary system, since each decision is made independently and the full range variability of the network is available at each decision. However, it might be desirable to alternate "sexual" and "asexual" operation in the network as occurs in certain biological forms. For example, some aphids reproduce asexually during times of relatively low environmental stress, thus building up large numbers during spring and summer. In the fall, however, they revert to sexual reproduction. This allows an interchange of genetic information and serves to increase variability just prior to the winter period of heavy natural selection. This variability is essential to effective evolution. But the variability is effective only during the period of environmental stress. During the summer period of population increase, sexual reproduction might even be detrimental to the species, since with few or no selective forces in operation, the gene frequencies would tend to drift from their winter norms. Added to this is the physiological "cost" of providing sexual reproduction, which is to be avoided if possible.

In a trainable logical network something akin to this might be implemented by providing two modes of operation. In the conservative or holding mode, a high degree of correlation would exist between successive decisions (perhaps 1.0). In the exploratory mode, successive decisions would be made independently. It remains to be seen whether it would be practical for the network to "learn" when to use these different modes, or whether this would have to be predetermined by fixed logic.

ENGINEERING CONSIDERATIONS

The theoretical work to date has included a detailed analysis of the relationships between evolving biological systems and trainable logical networks as they have been conceived to date. These investigations have shown a certain similarity between the organization of an evolving population and self-organizing devices. However, more important from the viewpoint of designing improved self-organizing systems are the differences that have been discovered. Some examples are described below.

1. Multiple Existence of Information Content

Each individual (or male/female pair) in the population carries all of the genetic information necessary to produce a similar individual. This is a large proportion of the total essential informational content of the system. (There is however, other genetic information that may be said to be associated with the population as a whole rather than with an individual; for example, information on *gene frequencies* as opposed to individual genotypes.) This multiple existence of essential information increases the reliability of the system since the death of any one individual results in no information penalty, only an energy penalty associated with reproducing a substitute. In contrast to this, redundancy in trainable logical networks ordinarily involves only replication of basic elements, with informational content spread over these elements. Failure of one element results in informational loss which must be made up by further interaction with the environment. Thus even though the system is capable of reorganizing after the failure, there is an informational penalty for the failure.

2. Multiple Simultaneous Trials

Each individual in the population, in addition to serving as a carrier for the genetic information, can also be considered to be a "trial" of this organizational information against the environment. Thus, in contrast to presently conceived self-organizing devices which use sequential trials

only, in evolutionary systems there is a multiplicity of "trials" in existence simultaneously. This would appear to provide two advantages:

1. The goal circuit can work on a comparative basis to select the better logical states to be rewarded, and
2. The process of adaptation can be speeded up by allowing many more logical states to be tried in a given length of time, assuming some physical restriction on the time to cycle through one trial.

3. Variable Number of Steps in Probability Space

In previous implementations of trainable logical networks, the number of steps between probabilites of 0.0 and 1.0 is fixed. In biological systems the number of steps in the probability space is directly related to the population size. This means that when the population size is low and the system is in a poorly adapted exploratory state, the amount of "hunting" due to random changes in gene frequency (noisy goal circuit) is large. As the species becomes better adapted the population size increases, the number of steps in the probability space increases and the amount of "hunting" decreases, thus allowing the system to lock onto its adaptive optimum.

4. Information Structuring and Transfer

Multiple existence of information implies further structuring and development of means for transferring information between subunits. In fact, we find in biological systems a complex structuring of multiple information carrying and processing systems, with built-in means for causing and regulating the transfer of information between them in such a manner that highly adapted forms will tend to propagate their genetic information to other subunits of the system. No such mechanisms exist in present generation self-organizing devices, but it may be concluded that these capabilities will probably be required in the sophisticated high-capacity machines projected for the future.

5. Non-Specificity in Reward/Punish Signals

In present day trainable logical networks, reward/punish signals are applied only to those switches which actually contributed to an output. This restriction requires a considerable complexity of circuitry in order

to properly route the reward/punish signals. This complexity will increase many-fold as the size of the network becomes large. By analogy with living systems, it has been shown that this specificity in the routing of reward/punish signals is not a pre-requisite to the organization of the system. All that is required is that the states of the other switches in the network not be highly correlated with the pertinent switches, and that the states of the pertinent switches not change between the time the goal circuit makes its decision and the time the reward/punish signal reaches the network. This finding will have considerable impact on the design of future large self-organizing systems, as well as on the practicality of implementing such systems in the near future.

CONCLUSIONS

The investigation reported here has been an exploratory study seeking to determine which properties of evolving biological systems are responsible for their superior adaptive capabilities. The work has used a straightforward bionics approach of first, modeling the biological prototype as closely as possible, second, experimentation with and analysis of the model, and finally, of drawing conclusions concerning the characteristics of the system of fundamental importance to its capabilities. Because of this classical bionics approach, the work reported here complements other investigations[9, 10] into engineering applications of bionic principles which have bypassed the initial phase of high fidelity prototype modeling.

While these investigations can in no way be considered complete, initial results reported here as well as the results of the other investigations cited above all tend to reinforce our initial thesis, viz; that the presence and utilization of the random search approach (trial-and-error) organizational process in both individual learning and in organic evolution reflects the inherent adaptive power of the process, and that this adaptive power can also be utilized in man's cybernetic machines. However, it has become increasingly clear that the successful application of the random search method will require an astute differentiation between those characteristics of the biological system which are merely coincidental to the adaptive features and those which are fundamental and necessary.

It has also become apparent during the course of the investigation that the obvious and basic processes in the system (mutation, selection or

trial, reward/punish) are perhaps *not* sufficient to produce an efficient and effective adaptive machine. It may in fact be, that the less obvious characteristics, such as replication of information content, information structuring and transfer, variable subdivisions of the probability space, etc., are also essential to the system. The relative importance of these more subtle characteristics will have to be determined by further experimentation and analysis before the evolutionary principle can be applied to any large-scale engineering system.

REFERENCES

1. Guinn, David F. "Large Artificial Nerve Net (Lannet)," IEEE Transactions on Military Electronics, vol. MIL-7, 1963, pp. 234–243.
2. Carne, E. B., Connelly, E. M., Halpern, P. H., and Logan, B. A. "A Self-Organizing Binary Logical Network," included in *Biological Prototypes and Synthetic Systems*, edited by E. E. Bernard and M. R. Kare, Plenum Press, Inc., New York, N. Y., pp. 311–330 (1962).
3. Lewontin, R. C., and Dunn, L. C. "The Evolutionary Dynamics of a Polymorphism in the House Mouse," *Genetics*, **45**, No. 6, June 1960.
4. Justice, K. E. "A Method for Determining Home Range in Small Mammals," *Journal of Mammalogy*, **42**, 462–470 (1961).
5. Justice, K. E. "Ecological and Genetical Studies on Desert Populations of Musculus," Final Report for AT(11-1)-900 published by the Arizona-Sonora Desert Museum, Box 5602, Tucson, Arizona, 66pp. (1962).
6. Dunn, Beasley, and Tinker, "Polymorphism in Wild Populations of House Mice," *Journal of Mammalogy*, **41**, No. 2, 220–229 (1960)
7. Allee, W. C. *et al. Principles of Animal Ecology*, W. B. Saunders Company, 1949.
8. Bruck, David. "Male Segregation Ratio Advantage as a Factor in Maintaining Lethal Alleles in Wild Populations of House Mice," *Proceedings of the National Academy of Sciences*, **43**, No. 1, 152–158 (January 1957).
9. Fogel, Lawrence J., Owens, Alvin J., and Walsh, Michael J., "Artificial Intelligence through a Simulation of Evolution," *Proceedings Second Cybernetic Sciences Symposium: Biophysics and Cybernetic Systems*, Spartan Books, Washington D. C., 131–155 (1965).
10. Bremermann, H. J., Rogson, M. E., and Salaff, S., "Search by Evolution," *Proceedings Second Cybernetic Sciences Symposium: Biophysics and Cybernetic Systems, Spartan Books, Washington*, D. C., 157–167 (1965).

JEROME ROTHSTEIN†

Laboratory for Electronics, Inc.

Boston, Massachusetts

*Excluded Volume Effects as the Basis for a Molecular Cybernetics**

Abstract

It is shown by analysis of chain molecules on the quadratic lattice (beads and rigid links with links meeting at a bead at 90° or 180°) where motion is by "flips", that prohibition of multiple occupancy of a site by more than one bead splits up the set of chain configurations into mutually inaccessible families. Defining a catalyst as an agent permitting transitions between two such families leads to a kind of controlled chemical switching effect by varying entropy, and thus free energy, by controlling the volume of the accessible portion of phase space. Static information storage, as in DNA, is achieved by introducing several distinguishable species of beads, with specific catalysts associated with particular bead "messages." From analysis of how the mutually inaccessible families of configurations are generated it is concluded that the entropy associated with selection of a family is small compared to the entropy of the total set of configurations whenever the set of lattice sites is large compared to the number of beads and there is a large ratio of interior to boundary sites. Increased crowding increases the family selection entropy, which equals the configurational entropy when all available sites are occupied. This limit can be attained sooner if boundary effects are important. Some plausible implications for biological control mechanisms will be discussed, including a computer analogy in which DNA-type information corresponds to the program and catalytic selection of families and corresponding tapping of sources of free energy correspond to the rest of the computer.

† Present address: Ohio State University, Columbus, Ohio

* This research was supported under Air Force Contract No. AF 33(615)-1464, Wright-Patterson Air Force Base, Dayton, Ohio.

INTRODUCTION

At the 1963 Bionics Symposium we showed, on the basis of simple statistical thermodynamical considerations, that a macromolecular species capable of replicating itself by assembling smaller units in a simple environment was, with overwhelming probability, a chain molecule.[1] The argument led to a characteristic alternation between the roles of energy and entropy in the free energy "bookkeeping" determining the sequence of events. During the assembly process, when the model was behaving, in effect, like a template to guide the synthesis of a replica, the building up process was sufficiently exothermic to overcome the entropy decrease entailed by collecting parts from solution. After completion of the assembly process, model and replica had to separate. Were separation not to occur, replication would no longer be properly describable as such, but would be a growth process, like that of a crystal. To bring about the separation, the only available thermodynamic drive is the increase in entropy entailed by the separation process. It was shown that for chains the configurational entropy change is large enough to overcome reasonable binding forces between model and replica, while for other configurations, with high probability, this is not the case. Accordingly, a strong motivation was provided to explore the configurational entropies of chain molecules in relation to the problems of information storage and control of selective sequences of chemical events. It was hoped that this would give deeper insight into possible mechanisms of interest for molecular biology.

Progress, continuing and extending the above considerations, has borne out the early hopes and has uncovered many promising new leads for research. It has been found, for example, that the alternation between energy and entropy, discussed above for the overall process of self-replication, seems to be involved in the individual events out of which the overall process is constructed. An important consideration throughout is the peculiar nature of the configurational entropy of a chain molecule. As is well known, excluded volume effects make calculations of chain entropies notoriously difficult, resulting in Monte Carlo approaches to the problem in the physical, chemical and biochemical literature. We felt that the kind of averaging over all configurations characteristic of previous statistical mechanical approaches was inadequate here. The reason for this dissatisfaction can be made intuitive by noting that if one were

to restrict oneself only to the statistical properties of an ensemble of messages one would be entirely unable to grasp the meaning of any message in the ensemble. Because of the dominating role of specificity and selective control characteristic of biological phenomena we thought it likely that it was precisely in those features lost in the averaging process that the really interesting phenomena would be found. We therefore decided to embark on the obviously horrendous task of analyzing the configurations of chain molecules in detail, without averaging unless we were sure that the process would not wash away the features of greatest interest.

In their full generality the combinatorial problems of chain molecule configurations are almost stupefying. Cursory examination showed that geological time on a fast computer would be insufficient to analyze and classify the kind of detail contemplated for chain lengths characteristic of virus or any other form of life. It was therefore necessary to simplify both the chains and the "world" in which the chains would be put through their paces, in order to whittle down the possibilities to something more manageable.

The first restriction was to adopt a very simplified model of a chain molecule, and let it move about in a simple space. More specifically the chain was idealized to a set of "beads" joined by perfectly rigid "links". The beads are allowed only on lattice points, and for simplicity we considered only the simple cubic and plane quadratic lattices. We showed recently that one would not expect essentially different results on other lattices. As far as basic concepts are concerned the restriction to the lattices chosen does little, if any, harm. The choice of a lattice, rather than a continuum in space, is an idealization away from the real situation, but we do not believe it to be a serious one in the present state of development of the subject. The reason for this assertion is the high degree of local order in liquids or condensed matter generally; it seems to be a fairly good approximation to regard the condensed media as highly "damaged" solids.

Considering plane rather than cubic lattices is also a drastic oversimplification, for there is no doubt that real chain molecules move in a three-dimensional world. It was felt, however, that the inherent difficulties of the subject made it imperative to take advantage of every simplification one could make short of simplifying the problem away altogether. Work accordingly concentrated on chains flipping about on the quadratic lattice.

It turned out that this decision was fortunate for reasons in addition to those of mathematical simplicity. The excluded volume property i.e., the rule that only one bead could occupy a given lattice site at a time, led to a rich structure in the plane problem which tended to wash away in space. It was precisely this rich structure, based on exclusion, that provided an opportunity for bionic control processes. Because of its origin we coined the word stereocybernetics, the prefix alluding to the steric hindrance and stereoisomerism so long familiar to organic chemists, for control and communication processes at the macromolecular level based on excluded volume properties.

The structure in configuration space associated with chains in the plane permits selective control of chemical events by catalysts. Because the catalyst exerts its effects only where there is this kind of structure on which it can operate, a natural explanation arises for the importance of surfaces or membranes in biology and the fact that so many reactions can be controlled only when reactants are suitably bound to the surface of an appropriate substrate.

In the present paper we start with analyses of the configurations of short chains in the plane and obtain from them properties which can be rigorously shown to hold even for very long chains. Among them are precisely those needed for a molecular cybernetics. After one shows how the excluded volume property gives rise to these possibilities, one must look more deeply to find how stored information can utilize this machinery. Here "DNA-type" information enters the picture, namely that encoded in the order of distinguishable beads in the chain. It is in the bead "messages" that one must find the catalytic agents capable of utilizing the configurational control machinery. From examination of how the mutually inaccessible families of configurations are generated it is concluded that the entropy associated with selection of a family is small compared to the entropy of the total set of configurations whenever the set of lattice sites is large compared to the number of beads and there is a large ratio of interior to boundary sites. Increased crowding increases the family selection entropy, which equals the configurational entropy when all available sites are occupied. This limit can be attained sooner if boundary effects are important.

In the sequel the discussion is given in broad outline only, as the details fill hundreds of pages of progress reports. The material presented here, however, makes it clear that excluded volume effects can serve as the basis

for a molecular cybernetics, and thus presumably a biophysics, consistent with thermodynamics and statistical mechanics.

BASIC DEFINITIONS AND PROPERTIES OF SHORT CHAINS ON THE QUADRATIC LATTICE

One can describe a chain configuration by choosing one of its ends as a starting point and then giving the sequence of steps necessary to reach each of the other beads in turn. The links lie along the edges of unit cells. A base 4 arithmetical notation is very convenient: denote unit steps along the positive x, positive y, negative x, and negative y directions by the digits 0, 1, 2, 3 respectively. With this convention a chain consisting of $(n + 1)$ beads, and thus of n links, is represented by an n-digit base 4 integer. Integers in which 0 and 2 or 1 and 3 are nearest neighbors are inadmissible as they imply retracting a step thereby putting two beads on the same lattice point. Similarly closed "loops" like 0123 or 001223 are excluded for the same reason. One of the chief difficulties in chain mathematics is the fantastically varied set of possible loops occurring for the longer chains and the absence of general algorithms for eliminating chains containing loops without testing virtually all subchains in turn.

It is convenient to refer to the chain with n links as an n-chain, and to refer to "bad" chains, i.e. those containing loops, as inadmissible. An admissible base 4 integer is sometimes called an admissible name of the n-chain and a general representative of the set of names related by symmetry transformations of the lattice or by a change in the direction of reading off the steps is called a shape. All shapes can be obtained as "words" formed from the symbols a, b, $\bar{a}$, $\bar{b}$, where the barred letters represent steps in directions opposite to the unbarred letters, the letter a is a step in any direction and b is either of the two steps perpendicular to the first choice.

The chains diffuse over the plane by means of "flips". These are of two kinds, end flips and corner flips. A flip is the motion of a bead from the site it occupies to an unoccupied site consistent with bond rigidity and at a distance no greater than the diagonal of a unit cell. Corner flips, as their name suggets, can occur for a bead at the junction of two links meeting at right angles. If the remaining corner of the square (three corners being occupied by the beads bounding the two links) is empty,

then the bead can jump across the diagonal to the empty site. It is easy to see that the final configuration differs from the initial one in that the two digits representing the links meeting at the corner have exchanged places. An end flip differs from a corner flip in that it is an end bead that moves. If the rest of the chain is fixed the end bead can flip into, at most, three positions. For example, in the chain 00..., flipping the first bead can transform the initial configuration either into 10... or 30... The last two can flip back to 00..., but two flips are required to flip from one to the other. Corner flips have unique inverses in the sense that after a corner flip has been performed the only flip open to that bead is to flip back to the initial configuration. For the end flip the "straight" configuration can flip to either of two "bent" configurations; the bent-to-straight transformation thus has no unique inverse.

Figure 1 shows the complete name and shape transition diagrams for 2-chains while Figure 2 gives the same information for 3-chains. Figure 3 gives the 4-chain name diagram and, as can readily be seen, the increase in complexity with increase in n is very rapid indeed. We have carried through the construction of detailed transition diagrams up to the case $n = 11$, and understand all the essential features of the transition diagrams for chains perhaps five or ten links longer than this. There are 120,292 admissible configurations for the 11-chain, and 17,245,332 for the 16-chain. The number of admissible configurations increases approximately exponentially with n (it is possible to bound the asymptotic number between two exponential functions of n and to push the two bounds

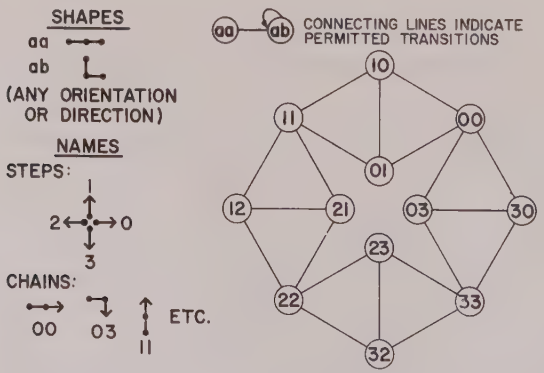


Fig. 1. Chain conventions and name and shape transition diagrams for 2-chains.

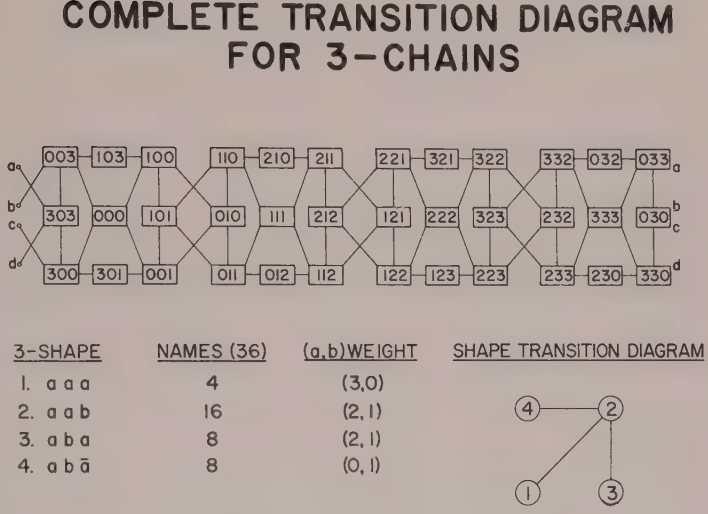


Fig. 2. Complete transition diagram for 3-chains.

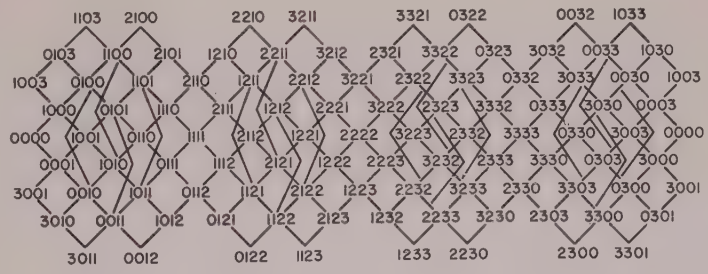
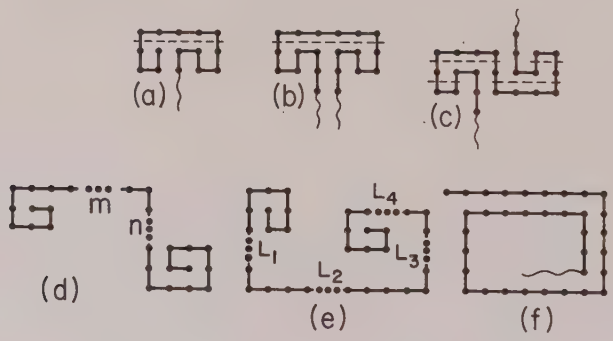


Fig. 3. Complete transition diagram for 4-chains.

fairly close to each other after a great deal of labor). Chains of biological interest, like DNA, contain many thousands of units.

It has been shown that for all chains on the quadratic lattice consisting of ten or fewer links the set of all admissible names is flip-connected. By this we mean that starting with any admissible name and using nothing but corner or end flips it is possible to reach the configuration described by any other admissible name. With the 11-chain a new phenomenon sets in, namely the existence of an admissible name (more generally a set of names) not accessible by flips from another set of admissible names. The shape in question can be written $ab\bar{a}a\bar{b}b\bar{b}aab\bar{a}$ and described as a chain consisting of the perimeter of a 2×3 rectangle with the middle of one long side deleted and replaced by the continuation of both ends of the remaining chain towards the interior by one unit. The chain is admissible but the ends are "tucked in" so that no flips are possible.

This phenomenon has been shown to generalize, and give rise to equivalence classes where two chains are equivalent if they can flip into each other. Figure 4 shows a half dozen of an unbounded set of families,



SESSILE CONFIGURATIONS

Fig. 4. Sessile configurations: wavy lines indicate continuing chains indefinitely prolonged. Dotted lines in (a), (b) and (c) indicate how sessile families can be generated simply by extending a pair of links so joined to subchains of the same length. An end-chain "anchor" is exemplified by (a), and two mid-chain anchors by (b) and (c). Arbitrarily large flip-connected sessile families are indicated in (d) and (e), where the dotted sections indicate runs of m, n, L_1, etc. links, and many configurations are connected by corner flips. A "hollow" end-chain anchor is shown in (f), where the indicated flips of the "trapped chain tail" are within the anchor configuration.

called sessile for reasons to be discussed later, each of which comprises a large set of flip-connected names, but none of which are flip-connected with each other. It has been shown that the number of such mutually exclusive families increases rapidly with increase in chain length, and that the "membership" of most families also increases rapidly with increase in chain length. Because the splitting of the set of names into mutually inaccessible families is entirely due to exclusion effects, one would expect that the presence of pre-empted sites, boundaries, or the like would also induce extensive splitting of the admissibles into smaller families. This is indeed the case and the complete splitting process has been calculated explicitly for chains on narrow channels (instead of the infinite plane). The results are precisely what one would expect from the discussion given above.

Many chains can migrate over the infinite plane by a succession of flips. One can arbitrarily select some bead, called the address bead, and follow the diffusion of the address bead all over the plane. Diffusion of the address bead is, of course, accompanied by name changes but as the number of possible names is finite, and the number of addresses potentially infinite, diffusion in the "name space" is quasi-periodic, while the lattice which is the "address space" is the arena of a genuine unbounded diffusion process. Those names with the ability to diffuse an unbounded distance, given sufficient flipping opportunities, are said to be mobile. Those doomed to remain within a finite area no matter how many flips they undergo are said to be sessile. The configurations of Figure 4 are sessile, generally because some part of the chain is so tightly packed that certain beads can never move at all. We suspect, but have not yet proved, that there are two and only two mobile families in the plane and that there are at most four mobile families on a channel. If this be the case, then virtually all of the phenomenon of splitting into families occurs in the formation of families of sessile admissible chains.

We know there are at least two families, when addresses are taken into account, for every family of names, which are said to differ in "parity". Their existence is a consequence of the flip rules on the quadratic lattice. The flip rules guarantee that any flipping bead hops across a cell diagonal. This means that the sum of the x- and y-coordinates of the bead remains odd if it was odd initially (odd parity), and remains even if it was even initially (even parity). A chain whose address bead is thus of given parity must therefore have its parity conserved no matter how it diffuses over

the plane. For every chain configuration, therefore, one has two families containing that configuration, each family containing the identical set of configurations found in the other, but with no possibility of flipping from one family to the other. The parity phenomenon does not exist on the triangular lattice. It is the reason we believe there are two mobile families in the plane. The four families on a strip stem from the additional doubling entailed by the fact that very long chains cannot turn around on a narrow channel. This result for the channel has been demonstrated; as short chains can turn around the four families reduce to two for these chains. By letting the width of the channel increase indefinitely one can argue that the four families become two in the infinite plane. The general proof has not yet been given because of the increasingly complex configurational situations arising with growth in channel width, though we believe it possible to supply the needed induction step.

Mention should also be made of "loops" and "tadpoles". While we have rejected chain configurations containing closed loops because they entail multiple occupancy of a lattice site, this does not prevent one from considering a case where the first and last beads of a chain are nearest neighbors. If one now joins these beads by a link one has a looped configuration without violating exclusion. Similarly one can join an end bead to an interior bead by adding an additional link, the resulting configuration being a loop with a tail, i.e., a tadpole. As mutual intertransformation of rings and chains, with or without appended side groups, are well known in organic chemistry, loops and tadpoles seem worthy of consideration in the present context. They have the interesting property of being sessile. This is easily seen for a tadpole, which has one bead where three links meet. It is obvious that this bead cannot flip without breaking a link. Loop immobility is more subtle because there is no bead, generally, which is immobilized. The general situation, however, can be grasped by considering a rectangular loop. It is "flipped out" as far as it can go; all flips place beads on the perimeter of the rectangle on interior sites. Occupancy of interior sites only reduces the freedom of beads on the perimeter to flip; it cannot make new sites available to them. Similar considerations apply to more complex loops.

To sum up, exclusion entails the splitting up of the family of configurations of an n-chain into a mutually exclusive and exhaustive set of families of configurations such that any two configurations within a family can flip into each other, and flips between families are forbidden.

The total number of configurations grows essentially exponentially with chain length. To see this, note first that if exclusion were neglected there would be 4^n possible configurations, and if only step retracing were prohibited there would be $4 \times 3^{n-1}$ configurations. Note next that there are 2^n ways of writing a "staircase" in the first quadrant, starting from the origin, with a total of n horizontal or vertical links, and all of these configurations are admissible. The number of admissible configurations of an n-chain is thus bounded between $p \times 2^n$ and $q \times 3^n$ where p and q are fixed integers. The upper and lower bounds can be squeezed together by further lowering and raising respectively, but computation of successive bounds becomes progressively more difficult.

CATALYSTS, STORED INFORMATION, AND CONTROLLED SEQUENCES OF CHEMICAL EVENTS

It is well known in chemistry that many reactions which are possible energetically occur at a negligible rate unless an agent called a catalyst is present. The role of the catalyst, like turning a valve or unlocking a door, permits the reaction to go much more rapidly. In many cases reactions go slowly because molecular shapes are such that the reacting subunits cannot get close enough to each other. The role of the catalyst is then either to induce a change of molecular shape, as when adsorption on a surface occurs, so that reactants become mutually accessible, or to mediate some other process like charge transfer. In either case, one can say that some form of hindrance, often steric hindrance, is overcome by the catalyst.

In the present context there is a natural area for catalytic activity in making mutually inaccessible families of configurations mutually accessible when the catalyst is present. In the formal mathematical model presently under discussion, a catalyst is then defined by giving the set of families it connects. A fundamental theory of catalysis requires a quantum mechanical investigation; here we are content with the above formal characterization, which is similar to using the bead-link picture instead of the quantum mechanics of molecular structure.

The exclusion-induced family splitting, together with the assumed existence of catalysts, provides a basis for selective chemical behavior. By this we mean that a sequence of chemical events can be induced to

occur which is specific in the same sense that selection of a particular message from an ensemble of possible messages is specific. We now consider how catalysts can enter into the chain picture and utilize the machinery of normally mutually inaccessible families to bring about such selective behavior.

Heretofore we have not endowed the beads of a chain with an individuality. Let us now explicitly admit some finite number of different kinds of bead so that one can distinguish different "messages" stored in the number and order of the various beads in the chain. To make this more intuitive, one can view the different kinds of beads as different letters of an alphabet, with the set of n-chains constituting the messages, using n letters, that one might select. This message is, of course, similar to a DNA code, being, in fact, motivated by the latter. We thus have two ensembles of possibilities associated with the n-chains, one the set of configurations earlier discussed, the other the set of DNA-like bead messages now being considered. There is no loss of generality, in a formal model like this one, if one regards bead messages as catalysts. The details of how a particular bead subsequence performs a particular kind of catalytic action is a deep problem of quantum mechanics, the solution of which is here assumed, in effect. Stored information affects what happens in our model "world" of flipping configurations by its catalytic effects only.

This is assuredly a drastic simplification of the real world, but it is one which appears to capture an important aspect of the real world. One can always make a simple situation more complicated by introducing other possibilities than those considered, but we use a strategy of maximum simplification for reasons earlier discussed. Even with these simplifications the problems are difficult enough; more complicated ones can wait until we have a better grasp of the simplest ones!

Reading stored information or writing information into a permanent record requires energy, whether we are using light to read a book, providing energy in a signal (with or without amplification) to drive a magnetic recording head, or whether the process occurs at a molecular level. Catalysis provides a conceptual control method, but there must be something to control which provides the necessary "drive". In thermodynamics this is supplied by some source of free energy (or some other thermodynamic potential appropriate to particular experimental conditions). For purposes of the present discussion it suffices to deal with

one of the simplest thermodynamic potentials, namely the Helmholtz free energy F, defined by $F = E - TS$, where E is the internal energy, T the absolute temperature and S the entropy. The direction of spontaneous chemical change is that of decreasing F, and this can occur by decreasing E, increasing S, or any combination giving the correct overall change in F (including those where either E or S alone goes the "wrong way" but is overpowered by the other). How does catalysis use free energy sources to accomplish "reading", "writing", and control in our model "world"? It is advisable to follow the process in a simple case to avoid making gross semantic errors. These arise from loose use of the terms information and entropy, without due regard for the differences between the worlds of thermodynamics and mathematics.

The mathematical information concept used in communication theory is perfectly straightforward. If one has a particular ensemble, generally called an ensemble of possible messages, then a mathematical function defined over the ensemble (namely the mean value of the logarithm of the probability of selecting any particular message from the ensemble), chosen to express how uncertain we are initially about what message will be selected, is defined as the amount of information conveyed by receipt of the message. This information function has the same mathematical form as the entropy of statistical mechanics and thus the name entropy was applied to it. Considerable confusion resulted from using "entropy as a measure of information", because the entropy of statistical mechanics corresponds to a measure of uncertainty in the theoretical specification of the state of the physical system, *without the uncertainty ever being resolved by receipt of a message.* For the physicist, receipt of a message is here the making of a measurement. Statistical mechanical entropy thus measures the missing information relative to a maximal specification of the state of the system.[2] This is a simple point (though it has caused much confusion), but the situation is made worse by the fact that in situations like the present one there are different entropy contributions which are certainly not equivalent informationally.

It is clear that in the ensemble of messages consisting of possible bead orders the informational concept and the physical concept are logically identical (remembering, of course, the difference in sign between information and entropy, as well as the existence of a constant, depending on choice of units, known as Boltzmann's constant in the physical case). But specifying the state of a chain involves specifying its configuration,

as well as bead order. This includes both the family of configurations in which it is flipping about and which configuration of the family it assumes at a particular time of interest. The latter information is typically not available in physical situations, and entropies calculated in statistical mechanics are simply measures of how uncertain we are, at equilibrium, about the microscopic configuration consistent with a given macroscopic specification of the system. The entropy of the family thus generally corresponds to "missing information". What about the third part of the physical entropy, corresponding to selection of the particular family in which the chain flips? Here we obviously have a vital part of the picture, intimately tied up with the whole process of controlling sequences of chemical events, and yet it is normally completely neglected in statistical mechanical discussions. More precisely, the selection of the family is included implicitly in the initial data, or in specifying the operations involved in preparing the system in a well defined macroscopic state, on the basis of which equilibrium entropies are then calculated.

The skeleton of the complex picture can now be seen. The information stored in bead order is equivalent to the catalysts that select the families. This means, in a very real sense, that the stored information controls the preparation of the system in the sense usually used when one talks of applying statistical mechanics to a system prepared in a particular macroscopic state.[3] But preparing a state means performing operations which require energy or, in the case of thermodynamic systems, free energy.

Let us now consider a specific example. Consider the synthesis of a very specific kind of molecular structure which does not form spontaneously in solution, but only in the presence of some complicated catalyst like one of our chains. The constituents thereby assembled are ordinarily sterically hindered from reacting, in addition to being dispersed all over the solution. Let the chain perform the traditional role of substrate or template in the following way. A particular constituent becomes bound to some site of the chain. This can only happen if the binding energy is sufficiently strong to overcome the entropy of solution of that constituent; the usual thermodynamic condition on F is satisfied thereby. But what has happened to the chain? A change has been made in its configuration space due to the binding of some of its units to this new constituent. Its whole structure of accessible and inaccessible configurations can undergo profound changes. In that new situation its beads can perform catalytic actions which they could not perform (or which were not even

defined) before the binding of the foreign constituent. Indeed, the foreign constituent can play a role like that of a new bead in helping to permit catalytic actions previously not available. In this new situation one can therefore see the possibility of exothermically condensing a second highly specific constituent of the molecule to be synthesized from the solution. A new configurational situation again results, permitting sequential condensation of a third constituent from solution, and so on, until the entire molecule is constructed. At this point one must separate the molecule so synthesized from the chain or template governing its construction. The thermodynamic drive is now to be sought in the entropy, rather than the energy part of the free energy. The bonds loosen because breaking them raises the free energy by a smaller amount than the free energy decreases because of the increase in entropy consequent to the separation of the controlling chain or template and the molecule which has been synthesized. The whole process is virtually a paraphrase of the earlier discussion on self-replication,[1] which led to the conclusion that the self-replicating species was, with overwhelming probability, a chain molecule. We can now see that a similar argument leads to the conclusion that when a specific compound is synthesized, depending on a long sequence of selective chemical events, and the compound produced is not itself a chain, then the template was a chain, with very high probability.

This is a very pleasing result, for just as the original argument indicated that Nature was not being arbitrary when she chose DNA to be a chain, so the present one indicates that she is not arbitrary in making messenger RNA a chain either. Also, because proteins surely play important catalytic roles in biosynthesis, it is no accident that proteins are also chains.

One should not conclude from this that of the pair, template plus product, the template is necessarily a chain. The same argument holds if the template is not a chain but the product is. This permits, for example, surfaces or membranes to perform similar catalytic roles. This possibility is almost surely a very significant one in many aspects of metabolism, particularly for those other than self-replication. The rich structure for chain configuration families induced in the plane suggests that actions sterically hindered in solution can occur in selectively catalyzed sequences on a surface, and may be involved in biogenesis.

It has been emphasized on many occasions how exclusion engenders family splitting. Perhaps the most drastic case occurs when one goes from a volume to a surface; this can be viewed as excluding all lattice

points in the volume except those on the surface chosen. The plane family structure completely washes away if one permits the lattice to fill a volume. There is, of course, a family structure in space also, but it is much more complex, involving things like knots in the simplest cases, and seems to be an undeveloped field even from the viewpoint of pure mathematics. The family structure in the plane, as mentioned previously, splits up even further if one introduces constraints like boundaries or pre-empted sites in the plane. The case of a narrow channel has been investigated in some detail leading to the general conclusion that the more restrictions introduced in this manner, the larger the splitting up into families, and the smaller the number of configurations, on the average, within a family. The "family selection entropy" thus increases as the situation becomes more crowded in the sense of having a larger fraction of the total number of sites occupied by beads, or more restricted by site pre-emption. In the limit when all available sites are occupied or made inaccessible, every family consists of a single configuration.

The complex interplay between family entropy and family selection entropy is a difficult, but obviously significant subject from the viewpoint of bionics and molecular cybernetics. Historical developments have heretofore concentrated most interest on the DNA type of information storage, but we make bold to prophesy that family selection entropies, family entropy, and their interaction with catalysts will occupy an increasingly important place in theoretical biological research. If the DNA code can be compared to a computer program, the present developments can, with considerable be justice, compared to the rest of the computer.

CONCLUSION

It is perhaps worthwhile to close with a few general observations. It is well known that the exclusion principle, in quantum mechanics, is what gives rise to the periodic table, valence and chemical structure and the stability of matter in bulk. Extending it to configurations too complex to calculate specifically according to quantum mechanical rules can be seen to lead to the excluded volume property of interest for chain molecules (and in many other contexts like crystallography or metallurgy). From this point of view it is not surprising that the excluded volume properties of matter give rise to a rich structure on which it is possible

to base a molecular cybernetics. We feel it is now not too much to expect that molecular biology can be made a branch of physics, in principle, in the same way that quantum mechanics makes all of chemistry or metallurgy part of physics in principle.

REFERENCES

1. Rothstein, J., "On Fundamental Limitations of Chemical and Bionic Information Storage Systems," IEEE Transactions on Military Electronics, Bionics Issue, (Eds. L. M. Butsch and H. L. Oestreicher) Vol. MIL-7, pp. 265–8 (April–July, 1963).
2. Rothstein, J., "Information, Measurement, and Quantum Mechanics," *Science* **114**, pp. 171–5 (1951).
3. This simple idea has deep ramifications. For an introduction to some of them see J. Rothstein, "Thermodynamics and Some Undecidable Physical Questions," *Philos. of Science* **31**, pp. 40–48 (1964)

Note added in proof: The existence of two and only two mobile families of opposite parities in the plane has been demonstrated, also the existence of at most four parity-polarity families on channels. Detailed treatment of chains on channels one link wide is given in J. Rothstein and P. James, "Families of Chain Configurations on the Quadratic Lattice and on Narrow Lattice Channels", Jour. Applied Physics, **38**, pp. 170–179 (1967). Further study of the family splitting phenomenon has shown it to be present for very general classes of flip rules, for drastic lattice modification, and even for weakenings of the excluded volume constraint short of permitting infinitely multiple occupation.

EDWIN R. LEWIS*

Librascope Group

General Precision, Inc.

Glendale, California

The Iron Wire Model of the Neuron: A Review†

INTRODUCTION

In conducting a literature search recently, a member of our laboratory discovered over 500 published articles devoted to neural modeling. The models discussed in these papers included large assortments of both passive and active electrical analogs, several mechanical analogs and hydraulic analogs, some pneumatic analogs, and a large number of mathematical models of various types. These models had been used to study many theoretical aspects of neural function and in many cases had led to concrete contributions to neurophysiology. Among the 500 and odd papers found during the literature search, well over 100, or 20% were devoted not simply to a single class of models but to a single model, a piece of iron immersed in nitric acid. Since the field of neural modeling was extensively reviewed recently by Harmon and Lewis[19], I will not attempt comprehensive coverage in this paper. Instead, I will devote most of my discussion to the 20% of the neural-modeling literature that deals with the iron-nitric acid model.

BACKGROUND

Speculation about the operation of the nervous system and construction of models to support that speculation is not exclusively a twentieth century game. This fact is demonstrated by Figure 1, which shows some of the mainstreams of neural modeling.

* Present address: University of California, Electronics Research Laboratory, Berkeley, California

† Research sponsored by Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force.

By the time of Galen, in the Second Century, the Aristotelian notions of the heart being the seat of sensation and will and the brain serving to cool the blood had been rejected. It was known that some nerves transmitted sensations to the brain and that some nerves transmitted motor commands from the brain to muscles. Galen synthesized the physiological and anatomical concepts of his day, and the resulting "Galenic Physiology" and "Galenic Anatomy" remained virtually unchallenged for nearly *fourteen hundred years*. Galen, often accused of having been dogmatic, but in many respects he was a scientific agnostic. He postulated that a force was transmitted through the nerves, but he was uncertain whether the force was "material" or "immaterial" and whether the nerves were hollow or solid.

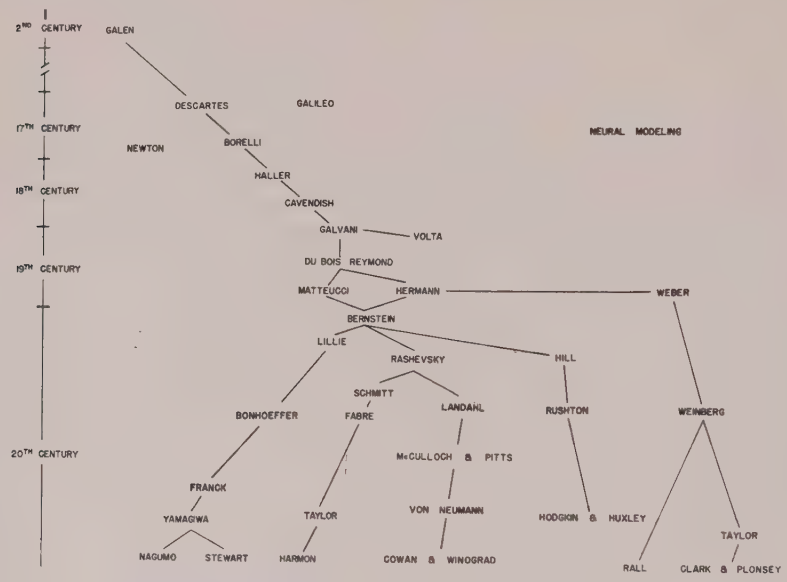


Fig. 1. Mainstreams of neural modeling.

Fifteen centuries later, Descartes was not so cautious. He committed himself to a concept which is often ascribed to Galen, but which probably originated with Erasistratus four hundred years before the time of Galen. Nerves were considered to be hollow tubes that conducted fluid or vapors from a central reservoir, the brain, to the muscles. These vapors filled the muscle, causing it to swell and at the same time con-

tract. Descartes viewed the nerves as a system of tubes, with valves to control the flow of fluid or vapor; and he considered the hydraulic automata that were in the gardens and fountains at that time to provide good models of nerve action. He used these machines to show that a fluid flowing through small tubes could elicit the rapid, powerful, coordinated actions typical of animals.

Influenced by the mechanistic views of Descartes as well as those of Galileo, Borelli set out to reduce all animal motion to purely mechanical principles. He also assumed that muscles contracted by swelling, and to demonstrate the plausibility of this notion, he proposed the rhombohedron as a model of the muscle fascicle. If the edges of a rhombohedron are fixed in length, the distance between opposite vertices will *decrease* over a considerable range of *increasing* volume. This is demonstrated in Figure 2, which shows Borelli's two-dimensional approximation to

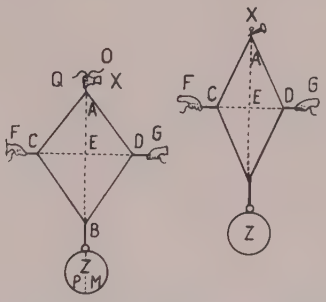


Fig. 2. Model muscle rhombs from Borelli.[8]

the rhombohedron. As indicated in the figure, he used ropes to simulate the individual fascicles and to calculate the expansive forces necessary for muscle contraction under load. On this basis he went on to calculate the forces involved in all known forms of animal motion, both visceral and skeletal, vertebrate and invertebrate.

By the last half of the Seventeenth Century, Glisson had demonstrated that muscle volume did not increase during contraction, and he had postulated that muscle contracts as a result of intrinsic irritability. This concept was made popular by Haller in the Eighteenth Century. Muscle was now thought to be an active device, needing only a stimulus or trigger to initiate contraction. Both Haller and Newton proposed mechanisms which could account for the propagation of the stimulus

through nerve. Haller proposed that nerve might be composed of a long row of spheres, each in contact with its neighbors. A sharp rap on the first sphere in the row would cause the last sphere to fly off immediately, accounting for the apparent rapidity of nervous conduction. Newton, on the other hand, supposed that nerves were translucent and that the stimulus was propagated as "optical vibrations".

Toward the end of the Eighteenth Century, the concept of "animal electricity" began to emerge. One of the first men to put forth a convincing argument for the existence of electricity in animals was Lord Cavendish. John Walsh had suggested that the shocks received from certain fish were the result of electric charge, but the results of his experiments to prove this were inclusive. For one thing, he could find no spark associated with the shock, and it was well known that sparks were associated with electricity. In addition, the shock produced no deflection in even the most sensitive electrometers. Finally, the shocks of the ray, *Torpedo*, were as strong in saltwater, a good conductor, as in air, an insulator.

In order to show that these observations did not exclude electricity, Cavendish[11] constructed a model ray. It was made of laminated wood, with pewter plates attached to the top and bottom. Wires led from each plate through glass tubes. The entire model was covered with sheepskin and soaked in saltwater for several days to increase the conductivity of the wood. Cavendish placed the model in a trough of saltwater and placed one hand over each pewter plate while an assistant touched the wires to a battery of charged capacitors. He found that he received the strongest shock from capacitors with low voltage and high charge. Since the voltage was low and the discharge was transient, he was unable to obtain either a spark or electrometer deflection. Also, since the body of the model was a very good conductor, the shock in saltwater was not much weaker than the shock out of it. Cavendish thus had used a model to answer all of the objections to Walsh's hypothesis, and animal electricity became an accepted phenomenon.

Eighteen years after Cavendish published his results, Galvani published the results of an experiment which demonstrated the presence of electricity in muscle. As a consequence of these results, he postulated that a muscle was similar to a charged capacitor and that the role of nerve was to discharge the capacitor, causing a contraction. Unfortunately, his interpretations of the experimental results were discredited by Volta,

who attributed them not to electricity intrinsic in muscle but to the generation of electricity by contact between dissimilar materials.

The controversy between Volta and Galvani was not settled until the middle of the Nineteenth Century, when du Bois-Reymond was able to measure electric currents associated both with nerve and with muscle activity. In order to do this, he had to construct the two most sensitive galvanometers in existence. He also constructed several electrochemical analogs in order to help explain some of his measurements; two of these are shown in Figure 3. Figure 3A shows an array of copper cylinders with zinc strips soldered to either side. Figure 3B shows 72 of these cylinders

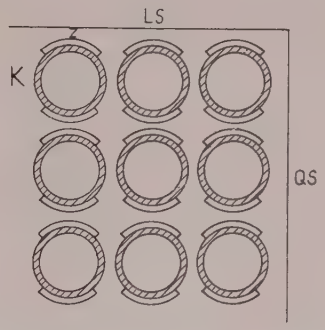


Fig. 3A. Model peripolar molecules.

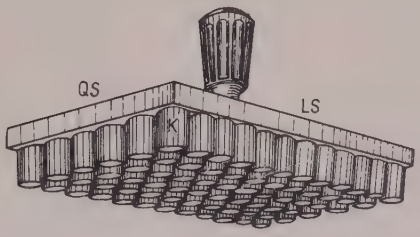


Fig. 3B. Model muscle rhomb from Du Bois-Reymond.[14]

mounted to form a model of a muscle fascicle. The cylinders were submerged in spring water so that steady electric currents flowed among them. The resulting voltage distribution replicated that found by du Bois-Reymond in muscle.

He used a whole series of models of this type to develop his "peripolar molecular" theory which was the forerunner of Bernstein's ionic hypo-

thesis. With the help of these models du Bois-Reymond was able to explain several observations including what he called "electrical antagonism" in muscle, "electrotonus" or electrotonic spread in nerve, and the propagated "negative variation" in both nerve and muscle. His analogs were the predecessors of a long series of electrochemical neural models, many of which are still in use today.

In order to examine the problem of electrotonus in more detail, Matteucci and Hermann both used electrochemical models. Mateucci's model consisted of a platinum wire wrapped with a cloth sheath and submerged in an electrolytic solution. Hermann's model was a bare wire immersed in an electrolytic solution. These two "core conductor" models were used to show that neural phenomena did not depend on the assumption ad hoc of peripolar molecules, and thus they led to the rejection of du Bois-Reymond's theory. Matteucci and Hermann were followed immediately by Bernstein, who along with Ostwald, Nernst and others developed the membrane theory and the ionic hypothesis. At this point the tree in Figure 1 branches. The remainder of this paper will be concerned primarily with the branch on the far left. Before discussing the iron wire models of Lillie et al, however, we will review briefly the other branches.

Expanding an earlier theory by Blair, Rashevsky[45] proposed a purely mathematical model of excitation in nerve. This model was described simply by two coupled, ordinary, first order, linear differential equations. The dependent variables were excitation and inhibition respectively, and the independent variable in each equation was time. The model accounted for essentially all of the excitation phenomena then known, including the triggering of a spike on either the leading edge of a negative pulse or the trailing edge of a positive pulse and the relationship between spike threshold and stimulus duration.

The equations describing Rashevsky's model were not always solved mathematically. Hill[24], who had independently proposed an equivalent model, offered an hydraulic analog for his system. This is shown in Figure 4. Here the height (V) of the water in tank *B* is equivalent to membrane potential and the height (*U*) of the hole (*D*) in the piston is equivalent to the threshold. Excitation is assumed to occur when *V* equals *U*. A pump (*P*) can produce a flow, or current, in either direction. If taps *K* and *L* are open and the pump is inactive, the system is at equilibrium with equal levels in the two tanks. An excitatory stimulus

is simulated by P pumping water from tank A to tank B. As the level in tank B increases, the piston (C) will rise with a rate determined by the viscous resistance of tap L. The system thus accommodates to slow increases in V, while rapid increases will serve to excite.

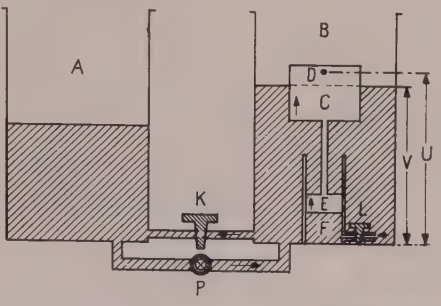


Fig. 4. Hydraulic nerve analog from Hill.[24]

Another method of solution for Rashevsky's equations was proposed by Rushton. He designed a mechanical device which could be used to plot a piecewise linear approximation to the solution. Monnier, on the other hand, used a passive electric circuit analog to solve the equations. Schmitt used an active electric circuit that included his well-known "Schmitt-Trigger". The branch in Figure 1 which begins with Schmitt is that of electronic neural models. It passes through Fabre in the early 1940's then through Taylor in the 1950's and on to many more, of whom Harmon is one current example.

Looking back to Rashevsky, we find that he and his colleagues, including Landahl, proceeded in the application of differential equations and continuous mathematics, attempting to extend their modeling efforts to include large systems of nerves and phenomena such as perception and discrimination. In 1943, McCulloch and Pitts[41] altered this course, applying discrete rather than continuous mathematics to the study of neural nets. Under the influence of Von Neumann[56] as well as that of McCulloch[40], this branch of the tree in Figure 1 has turned toward the problem of constructing reliable nets from unreliable elements. The work of Cowan and Winograd[13] provides a recent example of this trend.

The branch of the tree that began with Hill and Rushton extends to the work of Hodgkin and Huxley in the early 1950's. The four nonlinear differential equations of Hodgkin and Huxley[25] constitute probably the

most well known of all neural models. Three of the equations describe time and voltage dependencies of potassium and sodium ion currents across the squid axon membrane. The fourth equation relates the membrane potential to all of the other state variables of the system. The four equations, along with the data on which they are based, constitute the present successor to Bernstein's ionic hypothesis.

The branch of the right-hand side of the tree in Figure 1 begins with Weber[57] and his formal theory of the core-conductor. This work was extended by Weinberg[58] and has been reviewed recently by Taylor[54] and by Clark and Plonsey[12]. Interesting modifications of the core conductor work are provided by Rall[44], who modeled the spread of potentials in complex, passive dendritic trees.

The remaining branch, on the left-hand side of Figure 1, begins with Lillie and consists of individuals who have studied the passivated iron wire as a model of nerve.

NINETEENTH CENTURY STUDIES OF PASSIVE IRON

The phenomenon known as passivity probably was first described for iron by Keir[30] in 1790. He noted that after iron had been exposed to strong nitric acid or silver nitrate it behaved more like a noble metal than like iron. Normal iron, for example, would precipitate silver from a silver nitrate solution; the exposed ("passive") iron would not. Ordinary iron would precipitate copper from sulphate or nitrate solutions; passive iron would not. In other words, the solubility of iron was decreased by the action of the nitric acid. Keir observed that when passive iron was scratched or brought in contact with normal iron, it immediately became "active" (i.e., the solubility returned to that of normal iron).

For nearly forty years after Keir's publication, no mention was made of passivation of iron. Then, in the 1820's Wetzlar[59] rediscovered the phenomenon. In addition, he noted that when drops of copper acetate solution were placed on passive iron, copper deposition took place on the iron; but it began at the perimeter of the drop and spread toward the center. The change of state from passive iron to active (normal) iron had thus propagated inward from the edges of the drop. Observing that a change of state was accompanied by a temporary change of color on the iron surface, Herschel[22] soon discovered that the change to a passive

state was propagated as well. He also noted that propagation of the active state along an iron wire could be blocked by coating a small section of the wire with wax. Herschel concluded that the passive state was due to a "permanent electrical condition of the surface".

The work of Wetzlar and Herschel generated considerable interest, particularly because of its potential application to corrosion resistance; and between the time of Wetzlar's first paper on the subject (1827) until the end of the 19th Century, at least two hundred papers were published on the passivation of iron. Most of these were concerned with explanations of passivity. Faraday, for example, proposed that the surface of passive iron was oxidized. Helmholtz[21] thought passivation resulted from a thin coat of iron nitrate, while Andrews[3] believed it was due to an extremely clean surface. Other explanations for passivity included mechanical alteration of the iron and protective films of oxygen, nitric oxide, nitrogen, or cyanogen.

Whatever its nature, passivation could be induced by any one of several methods. As mentioned previously, Keir passivated iron with silver nitrate as well as with strong nitric acid. Other workers successfully used gases such as nitric oxide and nitrogen peroxide. In addition, it was found that passivation was enhanced by making the iron anodal in the appropriate solutions. The active condition was restored by making the iron cathodal. Other methods for restoring the active state included vigorous stirring of the liquid surrounding the iron, scratching the surface of the iron, and adding halide ions to the solution.

Herschel had noticed that under certain conditions oscillations of passivity and activity would occur on the surface of a piece of iron in nitric acid. Successive waves of passivity and activity would propagate back and forth over the surface. If two pieces of iron were connected to opposite terminals of a galvanometer and then immersed in nitric acid, oscillations of current were observed. In spite of these observations, passive iron generally was thought to be nonreactive and insoluble in nitric acid. Toward the end of the 19th Century, however, Nichols and Franklin[43] showed that a continuous reaction took place at the surface of the iron, and passive iron did indeed dissolve slowly but quite measurably in nitric acid.

By 1900 many of the characteristics of passive iron had been ascertained, but with respect to the origin of those characteristics no single theory prevailed. Two questions were prominent: what was the compo-

sition of the passive surface, and why were passivity and activity propagated? At the suggestion of Ostwald, Heathcote[20] began an intensive study of passivated iron.

Mousson[42] and others had suggested that propagation of passive and active states depended on local electric currents. Noting that local electric currents had been observed in nerves during propagation of the nervous impulse, Ostwald suggested that propagation in passivated iron might be analogous to that in nerve; and Heathcote went on to examine this analogy in considerable detail.

1. Heathcote – Nerve as an Analog of Passive Iron

Heathcote was interested primarily in determining whether or not passivation was an electrolytic process, and a study of transmission of active and passive states was essential to his objective. He was concerned with how the action of a small portion of iron in one state could effect a change of state in the large remainder. Examining the transmission of the active state over the surfaces of passivated iron rods and iron wires in nitric acid, Heathcote performed experiments which were to be repeated time after time by neural modelers in the succeeding years. He found, for example, that if he touched one end of a passive iron wire with a piece of zinc, a state of activity would propagate along the entire length of the wire. Following this wave of activity, the wire would become passive for a short time, then another wave of activity would begin; and the system would oscillate in this manner until the zinc was removed.

In another experiment, Heathcote observed that the velocity of transmission of the active state decreased as the volume of nitric acid surrounding the iron wire was decreased. He also noted that the velocity was greater in dilute acid than in concentrated acid.

Perhaps his most significant result with respect to transmission was obtained with a 75 cm iron rod in moderately concentrated nitric acid. Touching this rod with a piece of zinc, he observed that the active state did not spread over the entire rod at one time, but propagated, as zone of activity 10 cm long, with passive iron in front and behind. Prior to this observation, Heathcote had assumed propagation in iron wire was analogous to the transmission of chemical action in gunpowder, but if the

analogy had been valid, activity would have spread in both directions from the zone and engulfed the rod. Heathcote was able to account for this difference with two observations. First, currents flow from an active region to neighboring passive regions, tending to activate the latter and passivate the former. In addition, immediately following repassivation, the surface of iron becomes highly resistant to activation; and this state of refractoriness decays slowly. The iron rod is thus able to recover passivity even as a wave of activity is progressing along its surface.

It was at this point that Heathcote learned from Ostwald of the possible analogy between nerve transmission and propagation of the active state in iron. Examining the available facts on nervous transmission, Heathcote found nothing in that phenomenon that could not be explained by postulating a passive envelope around the nerve or around some structure inside the nerve.

For one thing the velocity of propagation in the iron wire was nearly equal to the velocity of nervous transmission measured by Helmholtz in frog motor nerves. In addition, both the iron wire and the nerve consumed very small amounts of energy during transmission, and both had electric currents associated with transmission. Both were able to recover and transmit again, yet both were subject to fatigue. Finally, even Keir in his early experiments had observed that activity could spread from one piece of iron to another without actual contact between the pieces; this could be analogous to synaptic transmission.

With respect to the iron wire itself, Heathcote concluded that passivation was indeed an electrolytic process and that electric currents were required for a change either from the passive state to the active state or the reverse. He also concluded that passivity is due to a layer of iron oxide on the surface of iron, but this point was not settled nor is it settled today.[67]

2. Lillie—The Iron Wire Revisited

Not long after Bernstein proposed the ionic hypothesis, Lillie became its chief proponent in the United States, and in 1909[31] he introduced into his arguments the first of a long series of discussions of electrochemical analogies. He pointed out that Bredig and his colleagues[9] had

found remarkable similarities between pulsatile catalysis in the case of mercury surfaces in contact with hydrogen peroxide and the rhythmical process in cardiac muscle.

In 1914 Lillie[32] postulated that the normal nerve-membrane polarization prevented some critical reaction and thus stabilized the membrane, much the same way that polarization at the surfaces of plates in a battery blocks the reaction between the plates and the liquid. In 1915[33] he carried the analogy further, proposing a model for the injury potential in nerve. This was a galvanized iron wire immersed in dilute sulphuric acid. When the outer layer of zinc was removed at some point, a continuous current would flow from the iron to the zinc, and one could observe an "injury potential" which diminished with distance from the point of damage. Lillie drew an analogy here between the zinc coat of the wire and the plasma membrane of the nerve fiber.

Lillies's first mention of the similarity between nervous transmission and the spread of activation over passive iron appeared in 1916.[34] By 1922[35] he had extended and refined Heathcote's analogy to include the following points:

1. Both systems (nerve and passivated iron wire) consist of a conductive core inside an insulating or semi-insulating sheath surrounded by an electrolytic solution.
2. Local chemical, mechanical, or electrical alterations may initiate a propagated disturbance in either system.
3. Once initiated, the propagated disturbance in either system has a form and velocity determined by the system itself and independent of the characteristics of the stimulus.
4. Local regions in both systems recover following the passage of disturbance.
5. In both systems propagation is associated with local variations of electrical potential and local current flow.
6. The propagation velocities in both systems are of the same order of magnitude and exhibit similar, strong temperature dependences.
7. In both systems propagation may be blocked by certain chemical or electrical alterations.
8. Both systems exhibit stimulus thresholds for initation of the propagated disturbance.
9. In both systems effects of subthreshold stimuli can accumulate and trigger a propagated disturbance.

10. Following the passage of a propagated disturbance, both systems exhibit temporary refractory periods during which their thresholds are higher than normal.
11. In either system an electric stimulus may excite or inhibit, depending on its polarity.
12. Inhibitory electric stimuli increase mechanical and chemical thresholds in both systems and may block passage of a disturbance initiated elsewhere.
13. In either system an electric current must be above a certain amplitude in order to trigger, no matter how long it is maintained; and a current sufficient for triggering must flow for more than a certain time in order to be effective.
14. A current rising too gradually from a subthreshold value to a normally sufficient intensity will not trigger either system; both systems thus exhibit accommodation.

From this long list of similarities between transmission in nerve and transmission in passive iron, Lillie inferred that the basic underlying physical processes were identical. It was generally accepted (and still is) that passive iron was surrounded by a thin film of oxide. If this film was broken by any means, the exposed active iron surface became anodal with respect to the passive surface. If the anodal area was sufficiently large, the resulting current would reduce the oxide film on neighboring regions; and the anodal area would grow. Propagation thus depended on electrochemical reduction of the resistive oxide film. Lillie proposed that nerve had a similar, chemically alterable film. In defense of this thesis, Lillie, pointed out that transmission accompanied by rapidly reversed breakdown or loss of continuity in the plasma membrane had recently been observed in echinoderm eggs during insemination. He also proposed that the local reactions responsible for nervous transmission might be oxidation-reduction reactions mediated by electric current, just as in passive iron. To support this he noted that nerve is not exhausted by repetitive stimulation except in the absence of oxygen and that lack of oxygen also tends to prolong the refractory period.

Perhaps his most significant modeling results were obtained by Lillie in 1925 with regard to myelinated fibers. To simulate the nerve in a myelin sheath, Lillie threaded a passive iron wire through segments of glass tubing, leaving between segments short lengths of exposed wire to simulate nodes of Ranvier. With this configuration, he observed two changes in the pro-

pagation of the active state.First, activation no longer proceeded steadily along the wire, but jumped from intersegment space to intersegment space. Second, the velocity of propagation was considerably increased in this mode. On the basis of these findings, Lillie proposed that the segmental structure of medullated (myelinated) nerve fibers might have a definite relation to the high velocity of transmission in those fibers.

In addition to examining conduction in a wire covered by segments of glass tubing, Lillie studied conduction in wires inserted in continuous glass tubes. He found that the velocity of transmission was nearly directly proportional to the diameter of the tube. This would correspond to observations in nerve, but only if the glass tube were taken to be the plasma membrane of the axon, the iron wire being some intracellular structure.

In the late 1920's Lillie[36] used the iron wire—nitric acid system to model biological rhythmic or oscillatory phenomena. By placing a very pure i on wire in nitric acid and touching one end with a piece of zinc, Lillie was able to induce the same type of oscillations between passivity and activity that Heathcote had observed, but with much more regular rhythms. Examining the effects of temperature on the oscillations, Lillie found that their frequency increased by a factor of $2\frac{1}{2}$ to 3 for every 10 °C increase in temperature. This factor corresponded remarkably to that found for biological oscillatory systems such as cardiac muscle and cilia. Lillie also observed that the frequency of oscillation in the model increased as the length of the iron wire was decreased. This was consistent with the general observation that the frequencies of ciliary beats and heart beats increase as the linear dimensions of the tissues decrease. The potentials associated with the active phase of the oscillations in iron wire showed a rapid, regenerative rise, a plateau which occasionally exhibited small secondary oscillations, and a rapid, regenerative fall. Regenerative rise and fall separated by a plateau was also the basic form of cardiac potentials in many animals, in fact *Maja* and *Limulus* both exhibited oscillations in the plateau.

In another study related to oscillatory phenomena, Lillie suspended a ring of iron into a circular trough of nitric acid. When he touched the iron ring at one point with zinc, waves of activation began to propagate in both directions, but they were both annihilated as they met on the opposite side of the ring. When one of the waves was blocked by touching a point on its path with platinum (which prevented activation), the other

wave continued to propagate indefinitely in a circular motion around the ring. Both continuous circular propagation and mutual annihilation of oppositely traveling waves had been observed in circular rings of muscle tissue and in the subumbrellar tissue of medusae.

Finally, Lillie examined the possibility of simulating irreciprocal synaptic transmission with his model. He found that activation could be transmitted between two pieces of iron wire each half submerged in the acid but not touching each other, as long as they were electrically connected outside the acid. The observed transmission was indeed irreciprocal, occurring quite readily from the wire with the larger submerged area to the wire with the smaller area, and being difficult to obtain in the reverse direction.

Lillie published his last article on the iron wire model in 1936,[37] but that article was not to contain the final word on the subject. Both Bonhoeffer in Germany and Yamagina in Japan continued the iron wire studies and started schools of modeling that persisted well into the late 1950's.

3. Bonhoeffer—Iron-Wire Kinetics

Finding it incredible that two such apparently different chemical systems as passive iron and nerve should exhibit so many similar characteristics, Bonhoeffer[6,7] decided to attempt abstracting a kinetic description which would account for the common characteristics without specifying the underlying physical or chemical processes. Being an electrochemist, he first based his kinetic description on the iron wire, then generalized it to include nerve. He decided to ignore the transmission properties of iron wire and consider activation by itself. He began by noting four important characteristics of activation: threshold, accommodation, refractoriness, and oscillatory behavior.

Bonhoeffer assumed that the kinetics of activation could be described by two state variables, x and y, the solutions to a pair of coupled, first order differential equations.

$$dx/dt = f(x, y)$$

$$dy/dt = g(x, y)$$

He equated x with the degree of activation as indicated by the electrochemical potential, and y with refractoriness as measured by the thre-

shold. To account for accommodation, dy/dt was taken to be positive for positive values of x; in other words, refractoriness was increasing any time the wire was in an active or partially active state. Also, since activation was an autocatalytic process—as witnessed by the threshold and the oscillatory behavior of the iron wire, dx/dt had to be positive for some range of positive values of x. Using these and other, similar arguments, Bonhoeffer estimated the qualitative natures of $f(x, y)$ and $g(x, y)$.

Using a state plane (i.e., a plane with coordinates x and y, as shown in Figure 5) he plotted two curves, one being the locus of $dx/dt = 0$, the other being the locus of $dy/dt = 0$. The intersection of these loci represented the point of equilibrium, P_0, for the system.

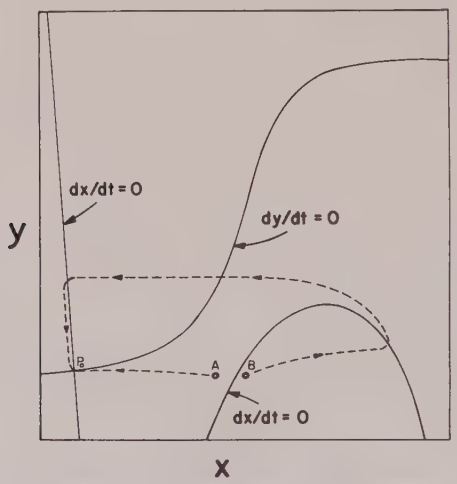


Fig. 5. State-plane diagram for the iron wire model, drawn from Bonhoeffer.[7]

Again using qualitative arguments, Bonhoeffer plotted the paths, or trajectories that the iron wire would follow through the state plane under various initial conditions. In the area under the bottom segment of the locus $dx/dt = 0$, the rate of change of x is positive, so the trajectories are directed toward the right, away from equilibrium. Everywhere else on the diagram, however, the trajectories are directed toward the equilibrium point, P_0. Bonhoeffer noted that changes in threshold were generally much slower than changes in the electrochemical potential. Except near the locus of $dx/dt = 0$, therefore, the trajectories are nearly

horizontal. A stimulus which leaves the iron wire in a state corresponding to point A is subthreshold; the trajectory proceeds directly back to P_0. A stimulus which leaves the wire in a state corresponding to B, on the other hand, is superthreshold; the trajectory proceeds through a phase of activation, then through a phase of repassivation (the top, horizontal portion of the trajectory), and finally through a refractory phase (the vertical portion on the left).

Using this graphical representation to describe the system kinetics, Bonhoeffer was able to explain all of the observations made by his predecessors on the iron wire model. More recently, Fitzhugh[17] has extended Bonhoeffer's model and has applied his graphical techniques, developing a very useful model of the nerve membrane.

Bonhoeffer, his student, Franck, and several of their colleagues, repeated many of the experiments performed by Lillie, using more sophisticated techniques and instruments, including the cathode ray oscilloscope. Franck, who published a review of neural modeling in 1956[18] was particularly interested in saltatory conduction. On the basis of his studies with the iron wire, he reasserted Lillies's hypothesis that conduction in myelinated fibers was saltatory and that this accounted for the high conduction rates in such fibers.

Franck also carried out careful studies of propagation velocity in his "saltatory nerve models," examining its depencencies on internodal distance and nodal area. He used his saltatory nerve models to build other, more complex models such as an interesting simulation of the electric organ and its innervation in electric fish. Passive iron electrodes were connected in series so their potentials would sum if all were fired synchronously. Branching saltatory nerve models were used to insure synchronization.

4. Yamagiwa—Propagation Studies

While Bonhoeffer in Germany in the early 1940's was studying the kinetics of activation in iron wire, Yamagiwa in Japan was studying its transmission properties. Yamagiwa was particularly interested in the length of the propagating active zone and the dependence of propagation velocity on that length.[62] Among other things, he found that the length of the propagated active zone was determined by the properties of the

wire and not by the properties of the individual stimulus. The principle determining factor in the wire was the time of passivation. If the wire was immersed in the passivating solution for a long time or if the intervals between the triggering of activation were long, the propagated zone was long and it propagated with a high velocity. If, on the other hand, the immersion time or the trigger intervals were short, the propagated zone was short and its velocity low. For very short immersion times or trigger intervals, conduction was decremental, the zone growing smaller and its velocity decreasing. Yamagiwa drew two interpretations from his results: first that continued propagation both in the iron wire and in nerve depends on an adequately long zone of activation or excitation, and second that the velocity of propagation is nearly directly proportional to the length of the active zone. He attempted to test his first conclusion both in nerve and in iron wire by artificially limiting the initially triggered area, but he was unable to control the initial area in either case. He has continued to hold the view, however, that triggering in nerve is determined by area rather than potential and that graded response is a small area activity.[66]

In other studies, Yamagiwa[60,61] examined interactions among parallel iron wire models in a common nitric acid bath. He found that in general the models exhibited completely independent conduction. When the wires were very close together, however, weak interactions did occur. First, if parallel wires of nearly equal intrinsic conduction velocities were stimulated simultaneously at one end, the active zones propagated together and at a velocity lower than that for conduction in the individual wires. Second, when two wires of markedly different velocities were stimulated simultaneously, the active zones propagated together at a velocity intermediate between the individual velocities. Finally, if one wire was stimulated immediately after the activation of a neighboring wire, its propagation velocity was increased. The first results were identical to those obtained by Katz and Schmitt[29], who observed that impulses in adjacent nerve fibers were occasionally locked in step and conducted at a reduced velocity.

Continuing his studies on interactions between adjacent iron wires, Yamagiwa[63] concentrated on the accelerating action of previous activity in one wire on conduction in its neighbor. He found that in an iron wire that conducted decrementally, the accelerating action increased the distance of conduction. In a wire that had recently been activated and was thus in a refractory state, the accelerating action reduced the re-

fractory time. The accelerating action itself declined exponentially with a time constant of several minutes. Finally Yamagiwa found that the interaction could be made irreciprocal. Several wires connected in parallel were much more effective in accelerating conduction in a single wire than the single wire was in accelerating them.

In another modeling study Yamagiwa[64] coated a length in the middle of the iron wire with paraffin. If the paraffin-coated section was sufficiently long, propagation of an active zone was blocked. Some effects of the blocked activation would be transmitted beyond the coated section, however, so that several successive waves of activation on one side would induce a single propagated zone of activity on the other side. Yamagiwa found that the number of successive waves required on one side of the paraffin to produce a single wave on the other side increased as the interval between waves was increased. These experiments and their results are reminiscent of the observations made forty years earlier by Lucas on nerve.

Finally, in studies essentially simultaneous with those of Franck in the early 1950's, Yamagiwa[65] examined saltatory conduction in iron wire. On the basis of his results, Yamagiwa also reasserted Lillie's hypothesis of conduction in myelinated fibers. Due primarily to the work by Tasaki[51] on myelinated nerve in toad motor fibers, that hypothesis is now generally considered valid; conduction in myelinated fibers is accepted as being saltatory and thus more rapid than conduction in unsheathed fibers. Tasaki himself later used the iron wire model, not in support of the theory of saltatory conduction, however, but in the support of this theory of the "mosaic membrane." This work is discussed in a later section.

5. Gunma Medical School – More Iron Wire Data

Inspired by Yamagiwa's work, Akiyama[1] and Matumoto[39], along with their students at Gunma University conducted a long series of experiments between 1954 and 1958 with the iron wire model. Most of these experiments simply confirmed previous results, but several interesting facts were uncovered for the first time.

One interesting result was obtained by Tajima,[50] who examined the dependence of conduction velocity on wire diameter. He found that velocity of propagation in the iron wire decreases with increasing wire

diameter. This is completely contrary to the relationship in nerve, where conduction velocity increases with increasing diameter. Tajima also noted that the propagation velocity for the iron wire increased as the volume of the surrounding nitric acid was increased. To replicate the dependence of conduction velocity on diameter in nerve, he suggested an iron tube with nitric acid on the inside. As mentioned in a previous section, Lillie had suggested an iron wire in a continuous glass tube filled with nitric acid.

Matumoto and, later, Matsuoka[38] observed that under certain conditions the propagating zone of activation on an iron wire was reflected on reaching a certain part of the wire and was propagated back in the reverse direction. They discovered that in cases where reflection occured, conditions existed whereby activation was prolonged at the site of reflection; and the reflected wave was the result of re-excitation of regions adjacent to that site. As a result of these observations, Kakinuma[26] attempted to find reflected spikes in the sciatic nerve of toad. Crushing one end of the nerve and stimulating the other, he did indeed observe reflections of the spike.

Several of the students at Gunma made detailed studies of stimulus strength required to trigger the iron wire as a function of stimulus duration. Strength-duration curves were obtained for iron wires of various diameters in nitric acid, for iron wires in silver nitrate, and for iron wires wrapped with silver wire and placed in nitric acid. In each case plots were made of strength vs duration, (strength) X (duration) vs duration and (strength)^2 X (duration) vs duration. In each case all three curves were essentially identical to the corresponding curves obtained in nerves and muscles. Zennyoji[68] put an end to these measurements when he measured the strength versus the duration of the current required to remove a coat of graphite from a piece of iron. Once again all the curves matched the data from nerve and muscle, so Zennyoji concluded that the strength-duration relation would not clarify the mechanisms of excitation in nerve or in iron wire.

The students at Gunma constructed several interesting modifications of the iron wire model. One of these was a block of iron on which small network patterns were painted with silver nitrate. Zones of activation could be observed traveling along the painted paths. In addition, some simple irreciprocally conducting junctions were developed between iron wires wrapped with silver wire.

Perhaps the most significant of all the results obtained by Matumoto

and his colleagues was in regard to local response in the iron wire model. Placing a small piece of passive iron in a solution of nitric acid and potassium ferricyanide, they observed the passive surface under a microscope as an activating current was applied. Activation at the surface was indicated by the appearance of a blue precipitate (Turnbull's blue). If the activating current was weak, small blue spots occurred randomly scattered over the surface. As the current was increased, the number of spots increased. When the density of the spots reached a critical value, activation spread over the entire surface. This result seemed to confirm Yamagiwa's notion of a critical area for triggering. It also provided support for a mosaic membrane theory proposed by Tasaki for nerve.

6. Tasaki – The Mosaic Membrane

Performing careful voltage-clamp experiments on the membrane of the squid giant axon, Tasaki and his colleagues[53] observed a repetition of discrete inward current pulses in response to a single step depolarization. They interpreted these results to mean that the excitable membrane of the squid axon is spatially non-uniform, each current pulse representing the full response of a limited area of the membrane. They went on to propose that each discrete area of the membrane has two possible states, a state of low permeability and a state of high permeability. Each discrete current pulse, therefore, represents a switching of states at a local area of membrane. They interpreted the apparently continuous increase in membrane current with increasing depolarization, as measured by Hodgkin and Huxley and others, as the result of a gradual increase in the area in the high permeability state. For any given depolarization the membrane is a mosaic of active and inactive patches, the ratio of active area to inactive area determining the gross membrane conductivity.

Hearing about the observations by Matumoto of discrete active patches in the iron wire model, Tasaki and Bak[52] set out to determine whether or not the iron wire system was analogous to the squid axon membrane. In performing their experiments, they inverted the usual iron wire analogy, taking the iron to represent the external fluid medium of the axon and the nitric acid to represent the intracellular fluid. With this convention, the potentials and currents in the iron wire model were of the same polarity as those in the axon. Plotting the peak "inward"

current against the clamping voltage for the model, Tasaki and Bak obtained a curve identical in form to the same function plotted for the axon. In addition, they observed the same sort of discrete current pulses that they had seen in the axon-records. They found, in fact, that all the characteristics of the iron wire-nitric acid system under voltage clamp closely resembled the corresponding phenomena in the squid axon. Coupling these observations with those of Matumoto, Tasaki and Bak submitted the iron wire model in support of their mosaic membrane hypothesis.

7. Recent Studies with the Passive Iron Model

Studies with the iron wire model and other, similar electrochemical models of nerve are apparently continuing. Carricarburu,[10] for example, has recently re-examined the oscillatory phenomena in the iron wire. Aldrich and others at MIT[2] considered the possibility of using passive iron models to simulate large neural networks; and Stewart has attempted to use passive iron spheres in nitric acid to simulate wave propagation in homogeneous aggregates of cells. Stewart has, in fact, proposed an adaptive system composed of mixtures of iron and glass spheres. The currents associated with the active waves would cause precipitation of gold or iron in the form of dendrites, and these would presumably strengthen the often used paths, providing a form of Hebbian learning. Stewart proposes eventually to reduce the size of the iron and glass particles in his system and thus increase its complexity. He envisions an electrochemical system that, after a prolonged period of training, will perform useful information processing functions. In attempting to develop his adaptive machine, Stewart has found it necessary to extend considerably the technologies associated with the iron wire – nitric acid system. Recently, he has begun to apply his electrochemical models to masses of nerve cells, examining such problems as the relationship between membrane conductance and whole-brain resistance.

Another interesting study is being carried out by Nagumo*, who is examining wave phenomena in two-dimensional iron wire meshes. He has observed both single passage and reverberation of waves of activity on iron screen in nitric acid. The wave activity in the wire screen is quite reminiscent

* Personal Communication; J Nagumo, University of Tokyo, Tokyo, Japan

of that observed by Farley and Clark[16] in computer simulated networks of discrete neural-like elements.

Finally reverting to the pre-iron wire model of Bredig *et al.*, Vis[55] has proposed to use the mercury-hydrogen peroxide system as a model of nerve.

DISCUSSION

Probably the most important single fact about the iron-nitric acid model of nerve is that almost everyone who has worked with it has placed great significance on its similarities to nerve. Lillie, for example, extrapolated from the iron wire model to propose a chemically reducible plasma membrane in nerve. Hill[23] pointed out, however, that complete reduction of the plasma membrane would set free far more energy than was observed in nerve on the passage of an impulse. With respect to the hypothesis of saltatory conduction, the iron wire results certainly suggested the rudiments of a valid hypothesis; but in the iron wire, conduction was from node to node and did not involve activation along nerve the intervening segment. That this probably could not be the case in was pointed out by Erlanger and Gasser[15] in the late 1930's; and it is generally accepted today that decremental conduction in the internodal segments plays an important role in transmission along myelinated fibers.

In spite of overenthusiasm on the part of Lillie and others, the iron wire model was well respected by many physiologists in the late 1920's and in the 1930's. Bishop[5] performed several experiments with it; Hill[23] said that physiologists could learn much from it if its limitations were kept in mind; and Katz[28] said that it helped illustrate the physical reality of Hermann's core-conductor theory.

This last point is well illustrated in Weinberg's paper[58], in which he expands Weber's mathematical statements of the core-conductor theory. Weinberg concludes, however, that the theory applies very well to the iron wire model but that its application to nerve is questionable. The problem is the core conductance of nerve, which is very much lower than that of the iron wire. The latter, in fact, can be assumed to be a perfect conductor in the core conductor formulation. Since the fluids surrounding an axon often exhibit a much greater conductance than the core (especially under laboratory conditions), the iron wire can almost be considered an inside-out model of nerve. As Tasaki and Bak[51] have pointed out, considering the model to be inside-out also corrects the

polarities of the voltages and currents. In addition, the relationship between conduction velocity and core diameter is inverted in the inside-out model and thus corresponds to the relationship found in nerve fibers.

Two strong critics of the iron wire analogy were Rosenblueth and Wiener.[46] They pointed out that experiments on the iron wire were at least as difficult to perform as they were with nerve and also that the phenomena of passive metals were not better understood than those of nerve and involved quite as much physical conjecture. They suggested that the useful model in the pair might be the nerve axon rather than the iron wire. The validity of these comments is at least partially confirmed by the fact that as many as fifty new papers appear in chemical journals each year proposing new theories of passivation; and the structure of the passive film has certainly not been determined.[67]

Regardless of whether or not the iron wire model is inside-out, and regardless of its present utility, it did serve one important role: it demonstrated that nervelike, decrementless conduction could occur in a purely electro-chemical system. In this respect it fulfilled the purpose for which it was proposed by Lillie; it supported the membrane theory and the ionic hypothesis—particularly against vitalistic arguments.

REFERENCES

1. Akiyama, I. The silver nitrate and iron system as an electrochemical model of nervous conduction. *Gunma. J. Med. Sci.* **4**, 41–46 (1955).
2. Aldrich, J. A., McCulloch, W. S., Lettvin, J. Y., and Pitts, W. H. The iron-wire neural nets. *Quart. Prog. Rep. Lab. Electronics M.I.T.* **64**, 303–305 (1962).
3. Andrews, T. On the action of nitric acid upon bismuth and other metals. *Phi. Mag.* **12**, 305–311 (1838).
4. Bennett, C. W., and Burnham, W. S. Passive state of metals. *J. Phys. Chem.* **21**, 107–149 (1916).
5. Bishop, G. H. The effects of polarization upon the steel wire-nitric acid model of nerve activity. *J. Gen. Physiol.* **11**, 159–174 (1927).
6. Bonhoeffer, K. F. Über die Aktivierung von passiven Eisen in Salpetersäure. *Z. Elektrochem.* **47**, 147–150 (1941).
7. Bonhoeffer, K. F. Activation of passive iron as a model for excitation of nerve. *J. Gen. Physiol.* **32**, 69–91 (1948).
8. Borelli, A., *De Motu Animalium* Rome: Bernado, 1680, Vols. I and II.
9. Bredig, G., and Weinmayr, J. Eine periodische Kontaktkatalyse. *Z. Physik. Chem.* **42**, 601–611 (1902).

10. Carricaburu, P., Oscillations de relaxation du nerf d'Akiyama. *Compt. Rend.* **251**, 906–907 (1960).
11. Cavendish, H., An account of some attempts to imitate the effects of the Torpedo by electricity. *Trans. Roy. Soc.* (*London*) **66**, 196–225 (1776).
12. Clark, J., and Plonsey, R. A mathematical evaluation of the core conductor model. *Biophys. J.* **6**, 95–112 (1966).
13. Cowan, J., and Winograd, S. *Reliable computation in the presence of Noise.* Cambridge: MIT Press, 1963.
14. Du Bois-Reymond, E. *Untersuchungen über tierische Electricität*, Vol. I. Berlin: G. Reimer, 1848.
15. Erlanger, J., and Gasser, H. S. *Electrical signs of nervous activity.* Philadelphia: University of Penn. Press, 1937.
16. Farley, B. G., and Clark, W. A. Activity in networks of neuron-like elements. In: *Information Theory* (*Fourth London Symposium*), edited by C. Cherry. London: Butterworths, 1961, p. 242.
17. Fitzhugh, R. Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1**, 445–466 (1961).
18. Franck, U. F. Models for biological excitation processes. *Prog. Biophys.* **6**, 171–206 (1956).
19. Harmon, L., and Lewis, E. R. Neural modeling. *Physiol. Rev.* **46**, 513–591 (1966).
20. Heathcote, H. L. The passivifying, passivity, and activifying of iron. *J. Soc. Chem. Ind.* **26**, 899–917 (1907).
21. Helmholtz, H. Messungen über den zeitlichen Verlauf der Zuckung animalischer Muskeln und die Fortpflanzungsgeschwindigkeit der Reizung in den Nerven. *Arch. Anat. Physiol. u. wissenschaft. Med.*, Berlin. pp. 276–304, 1850.
22. Herschel, Sir J. F. Lettre fur la maniere d'agir de l'acide nitrique fur le fer. *Ann. Chim. Phys.* **54**, 87–94 (1933).
23. Hill, A. V. *Chemical Wave Transmission in Nerve.* New York: Macmillan, 1932.
24. Hill, A. V. Excitation and accommodation in nerve. *Proc. Roy. Soc.* (London) B **119**, 305–355 (1936).
25. Hodgkin, A. L., and Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation of nerves. *J. Physiol.* **117**, 500–544 (1952).
26. Kakinuma, S. Studies on comeback of excitation by means of action potential: II, Experiments on the nerve of the toad. *Gunma J. Med. Sci.* **8**, 139–143 (1959).
27. Karrakida, Y. Studies on the transmission of excitation with the electrochemical model of excitation conduction. *Gunma J. Med. Sci.* **7**, Supplementum 6, 1958.
28. Katz, B. *Electrical Excitation of Nerve: A Review.* London: Oxford University Press, 1939.
29. Katz, B., and Schmitt, O. H. Electrical interaction between two adjacent nerve fibers. *J. Physiol.* **97**, 471–488 (1940).
30. Keir, J. Experiments and observations on the dissolution of metals in acids and their precipitation; with an account of a new compound acid Menstrum, useful and some technical operation of parting metals. *Phil. Trans.* **80**, 359–384 (1790).

31. Lillie, R. S. On the connection between changes of permeability and stimulation and to the significance of changes in permeability to carbon dioxide. *Am. J. Physiol.* **24**, 14–44 (1909).
32. Lillie, R. S. The conditions determining the rate of conduction in irritable tissues and especially in nerve. *Am. J. Physiol.* **34**, 414–445 (1914).
33. Lillie, R. S. The conditions of conduction of excitation in irritable cells and tissues and especially in nerve. *Am. J. Physiol.* **37**, 348–370 (1915).
34. Lillie, R. S. The conditions of physiological conduction in irritable tissues. *Am. J. Physiol.* **41**, 126–136 (1916).
35. Lillie, R. S. Transmission of physiological influence in protoplasmic systems, especially nerve. *Physiol. Rev.* **2**, 1–37 (1922).
36. Lillie, R. S. Analogies between physiological rhythms and the rhythmical reactions in inorganic systems. *Science* **15**, 593–598 (1928).
37. Lillie, R. S. The passive iron wire model of protoplasmic and nervous transmission and its physiological analogues. *Biol. Rev.* Cambridge *Phil. Soc.* **11**, 181–209 (1936).
38. Matsuoka, T. Studies on comeback of excitation conduction. 1. Studies with electrochemical model of excitation. 2. On the straited muscle. *Gunma J, Med. Sci.* **7**, Supplementum 3, 1958.
39. Matumoto, M., A new type of nerve conduction model. *Gunma J. Med. Sci.* **4**, 37–40 (1955).
40. McCulloch, W. S. Agathe Tyche: of nervous nets—the lucky reckoners. In: *Mechanization of Thought Processes*. London: Her Mayesty's Stationery Office, 1959, vol. II, p. 613.
41. McCulloch, W. S., and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
42. Mousson, A. Versuch einer Erklärung des Verhaltens der Salpetersäure zu den oxydierbaren Metallen. *Poggend. Annal.* **39**, 330–342 (1943).
43. Nichols, E. L., and Franklin, W. F. Destruction of passivity of iron in nitric acid by magnetization. *Kansas Acad. Sci. Trans.* **10**, 13–19 (1887).
44. Rall, W. Theoretical significance of dendritic trees for neuronal input-output relations. In: *Neural Theory and Modeling*, edited by R. F. Reiss. Stanford University Press, 1964, p. 73.
45. Rashevsky, N. Outline of a physico-mathematical theory of excitation and inhibition. *Protoplasma* **20**, 42–56 (1933).
46. Rosenblueth, A., and Wiener, N. The role of models in science. *Phil. Sci.* **12**, 312–321 (1945).
47. Stewart, R. M. Theory of structurally homogeneous logic nets. In: *Biological Prototypes and Synthetic Systems*, edited by H. F. Bernard & A. S. Kare, New York: Plenum Press, 1962, p. 370.
48. Stewart, R. M. Adaptable cellular nets. In: *Progress in Biocybernetics* vol. I. Amsterdam: Elsevier Press, 1963, p. 96.
49. Stewart, R. M. Electrochemical nerve models as neuristors. *Proc. Inst. Elect. Electronic Engrs.* **52**, 78 (1964).

50. Tajima, K. Studies on the relation between conduction velocity and the thickness of the nerve fiber with the electrochemical excitation model. *Gunma J. Med. Sci.* **6**, 171–191 (1957).
51. Tasaki, I. Conduction of the nerve impulse. In *Handbook of Physiology*. Sect. 1 *Neurophysiology*. **1**, 75–121 (1959).
52. Tasaki, I., and Bak, A. F. Voltage clamp behavior of iron-nitric acid system as compared with that of nerve membrane. *J. Gen. Physiol.* **42**, 899–915 (1959).
53. Tasaki, I., and Spyropoulos, C. S. Nonuniform response in the squid axon membrane under voltage clamp. *Am. J. Physiol.* **193**, 309–317 (1958).
54. Taylor, R. E. Cable theory. In: *Physical Techniques in Biological Research*, edited by W. L. Nastuk, New York: Academic Press, 1963.
55. Vis, V. A. The mercury-hydrogen peroxide system as an analogue of nervous transmission. *J. Gen. Physiol.* **38**, 17–29 (1954).
56. von Neumann, J. The general and logical theory of automata. In: *Cerebral Mechanisms in Behavior: The Hixon Symposium*, edited by L. A. Jeffress. New York: Wiley, 1951, p. 1.
57. Weber, H. Ueber die Stationären Strömungen der Elektricität in Cylindern. *Borchardt's J. Math.* **76**, 1–20 (1873).
58. Weinberg, A. M. Weber's theory of the kernleiter. *Bull. Math. Biophys.* **3**, 39–55 (1941).
59. Wetzlar, G. Ueber die Reduction der Metalle durcheinander auf nassem Wege. *Schweiggers J.* 50, 88–109 (1827).
60. Yamagiwa, K. The isolated and non-isolated conduction in Lillie's nerve model. *Jap. Med. J.* **1**, 452–461 (1948).
61. Yamagiwa, K Interactions between active elements (observations on Lillie's nerve model). *Jap. Med. J.* **1**, 557–567 (1948).
62. Yamagiwa, K. The conduction velocity in relation to the simulation intensity and to the size of the activated area (observations on Lillie's nerve model). *Jap. Med. J.* **2**, 217–229 (1949).
63. Yamagiwa, K. The interaction in various manifestations (observations on Lillie's nerve model) Part 1. The accelerating action. *Jap. J. Physiol* **1**, 40–47 (1950).
64. Yamagiwa, K. Iterative excitability in Lillie's nerve model. *Jap. J. Physiol.* **1**, 269–276 (1950).
65. Yamagiwa, K. The electro-saltatory transmission of nervous impulse (model experiments and theoretical considerations). *Jap. J. Physiol.* **2**. 79–92 (1951)
66. Yamagiwa, K. Again on the local response as a small area activity. *Jap. J. Physiol.* **10**, 456–470 (1960).
67. Young, L. *Anodic Oxide Films*, New York: Academic Press, 1961, p. 227.
68. Zennyoji, H. Studies on electric stimulation with various electrochemical models of excitation. *Gunma J. Med. Sci.* **6**, 279–294 (1957).

SECTION II

Biological Foundations

WOLF D. KEIDEL

Physiological Department of the University of Erlangen-Nürnberg
Erlangen, Germany

Information Processing by Sensory Modalities in Man

Johannes Müller discovered one of the most important principles involved in conveying information by the sensory modalities in human perception. Before his classical contributions to the physiology of senses the sense organs were considered mainly as some sort of doors or openings and holes through which the stimuli of the outside world could enter the body. The brain was thought of as some homunculus who "looks at" the images designed by the eyes, the ears, the tongue and the nose. No clear concept existed according to which the separation into the different and typical modalities (a word defined much later by v. Helmholtz) was performed by the "mind". There is a nice story according to which a nobleman in the 18th century, named "Münchhausen", when hunting ducks lost his flint, the striking of which produced a spark, which in turn ignited the powder of his gun. Now what he did in his misfortune was to strike his eye with his fist. Consequently, he—by inadequate stimulation—saw sparks and—according to his story—used them instead of the real ones to ignite the powder of his gun. This exactly was the status before Johannes Müller's conception.

Thereafter—when no homunculus could be found in the brain—it was possible to make considerable progress by simple and effective techniques. *Waller*, an anatomist, proved histologically that after the section of axons within the brain, all neurites beyond the cut degenerated in a few weeks. After having discovered this fact it was easy to obtain sufficient insight into the anatomical nerve tracts within the CNS to recognize that—besides some intermixing—there exists a clear organizational structure within the brain, and that nerve fiber tracts can be clearly separated anatomically for each sensory channel. These channels in turn belong functionally to given modalities.

Undoubtedly this concept led to quite a number of data and fitted fairly well the psychophysical results when dealing with the visual, the

somesthetic and even the auditory system in general. However, there were considerable difficulties in applying the "specificity"-concept to other sensory channels, as for instance taste or, even more, pain. *Weddell* and his group were among the first to discover the morphology of the *multiple innervation* of the skin, which then was found to be of a very general nature: The convergence–divergence principle of the connection of one neural layer to the next "higher" one (Mountcastle *et al.*) was proved to be valuable for explaining some peculiar neurophysiological problems such as that of "funneling" in audition, or as the contrast-phenomenon in vision which was later discovered to be mostly if not entirely based upon lateral inhibition (*Hartline*, *Ratliff*, *Reichardt*). Today we even possess mathematical equations to describe exactly the lateral inhibition in Limulus. In the meantime *v. Békésy* had performed some ingenious and manyfold experiments in nearly all types of sensory systems which together revealed the tremendous importance of inhibitory processes in the handling, transmitting and processing of information.

There is—among others—certainly a third aspect in addition to that of "*multiple innervation*", and "*the convergence—divergence*" principle

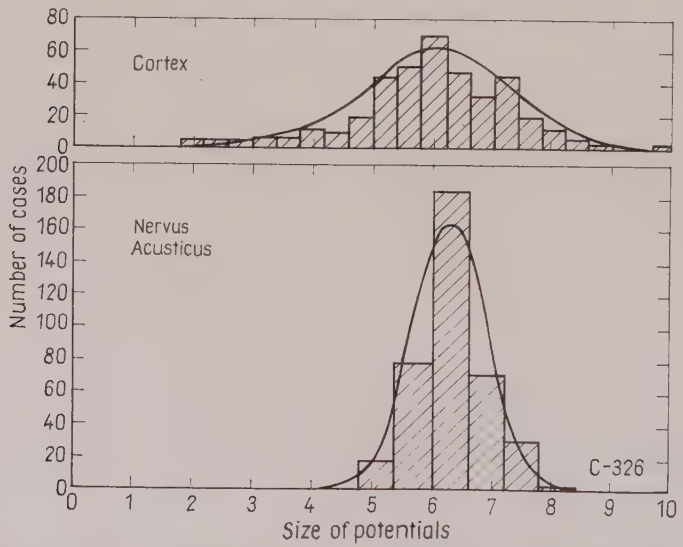


Fig. 1. A comparison of the distribution of the amplitudes of the compound action potentials of the auditory nerve and the auditory cortex in the animal. The statistical variability is considerably enlarged at cortical level. After *Macy*.

(including lateral inhibition), namely the *probabilistic principle*, which at least in part was promoted considerably at MIT. At first glance a very prominent feature of the CNS compared to the more peripheral parts of the entire neuronal system, is the fact that the precision of the information processing seems to decrease as the brain level under consideration increases. Figure 1 shows as an example the accuracy of intensity coding in the nerve supplying the sense organ compared with that at higher levels of the same sensory channel but within the CNS. It can be seen clearly that the overall performance of the coding process is much more precise in the auditory nerve than it is, for instance, at cortical level in audition. Although this neurophysiological behavior is more pronounced with respect to intensity coding processes and much less with respect to timing, this is the reason why modern computers are so useful in neuro- and sensory physiology. In older times it was thought that this behavior results from the lack of structural macromolecular features of the brain e.g., the existence of some residual, statistically variable permeability of the living membranes which—consisting of the chemical elements C,N, S, P, H, O—can not be built in the same precise way by nature as can be done by technological means, using completely impermeable insulators (*Ranke*). However today's more sophisticated insight provides a completely different concept.

As Figure 2 shows, the number of structural elements at the different levels of the CNS is clearly greatest at the cortical level. The bottleneck is in the primary neurons. As it is clear that no more information can be processed in the following, the "later" stages of the channels, the question arises why is there such a tremendous increase in the number of elements towards the cortex? This problem becomes even more pronounced when one considers the selection, the optimizing processes, taking place at the same time in the psychophysical processes corresponding to the same levels, as can also be seen in the figure. At present we do not know exactly whether this phenomenon is related to the fact that no computer has an even remotely comparable ability of compensating for failures of circuits as the human brain has. We do know e.g. that for the routine intellectual performance of aged people even 10 per cent of the organized neural structures in function suffice without a visible drop in performance, e.g. in sclerosis. We know on the other hand of the great efforts devoted to the technology of parallel multiple circuitry in satellite-electronics or in development of "self-healing" circuitry research which is performed

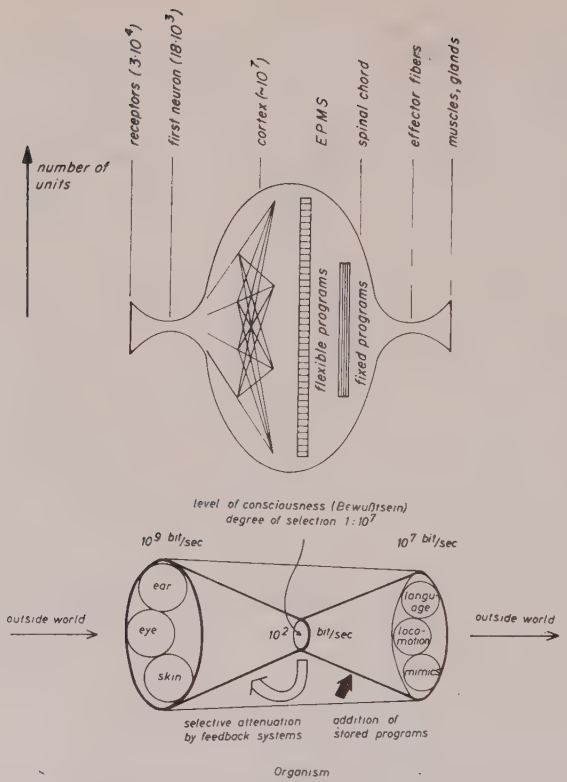


Fig. 2. The total amount of sensory information is reduced at a ratio of $1:10^7$ in man up to the conscious level (lower figure). The number of elements engaged in the processing of sensory information is much increased up to the cortex of the brain (upper figure).

in Europe e.g. by *Steinbuch* and his group. It is therefore tempting for biologists to look carefully to the morphological principles which give the brain such marvelous resistance and even regenerative ability against defective processes. Probably several hundreds of parallel circuits are the background of this phenomenal performance of the human brain in processing sensory information. A fourth point which I would like to mention briefly is the fact that the loss of sensory information does not impair, but sometimes even improves, the overall performance in sensation. Our group worked out a number of results regarding this problem, which is also related to the phenomena of adaptation and habituation or even to what we labeled "optimizing selection processes", (*Wolf D. Keidel*

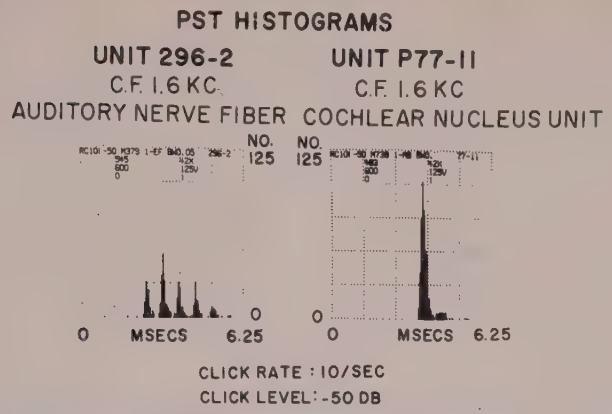


Fig. 3. A comparison of the post-stimulus-time histograms of the first neuron (left diagram) and the second neuron (right diagram) in audition (records from single units of the cat's auditory nerve and cochlear nucleus) show that the "unimportant" information about the single vibrations of the basilar membrane is completely lost at the second neuron's level (right diagram).

1963). I would like to show just one example, dealing with experiments performed by *Nelson Y.-S. Kiang*, which is one of the most brilliant experiments performed during the last years in physiological acoustics:

As Figure 3 shows, one can, using today's neurophysiological techniques, compare quantitatively post-stimulus-time-histograms of the first and of the second neuron in audition. The two histograms clearly show differences between the second and the first neurons. Between the cochlear nucleus and the auditory nerve all information related to the details of the individual vibrations of the basilar membrane are completely lost and the only remainder is time and stimulus intensity. These are certainly the only "important" parameters to be carried in the further information handling processes up to the cortical level. This might be considered as a very simple, but impressive, example of something the human brain is able to do best, namely to *select* information in a ratio of $1 : 10^7$ at conscious level and to generate such highly sophisticated concepts as "importance" in an individual, social and even evolutionary sense. This is different for different species of animals, and most developed in man, based upon his ability for "abstraction" and for "modelling" his surroundings by means of something as vague as what we call "consciousness", about which we technologically and biologically know nearly nothing.

Now let me come to my main point. As most of you know it is an old

author	differentiated	integrated
Merkel	$\Delta E = k \cdot \Delta R$	$E = k' \cdot (R - R_0)$
Weber	$\Delta E = k \cdot \frac{\Delta R}{R}$	
Fechner		$E = k' \cdot \ln \frac{R}{R_0}$
Plateau	$\frac{\Delta E}{E} = k \cdot \frac{\Delta R}{R}$	$\ln E = k' \cdot \ln \frac{R}{R_0}$
Stevens		$E = k' \cdot (\frac{R}{R_0})^n$

Fig. 4. Table of some types of possible functional relations between sensation magnitude and stimulus intensity. *Merkel's* law has proved to be wrong. The most important equation is that of *Stevens'* power function.

dream of all sensory physiologists to compare quantitatively both neurophysiological and psychological data. Figure 4 might remind you briefly of what types of magnitude scales can be obtained psychophysically by summing up a given number of measured just noticeable differences, a number which is related quantitatively to the stimulus intensity although not in a one-to-one, but rather in a multidimensional manner involving biological state, historical anamnesis, sensory interactions, vigilance processes and so forth. It was my friend and colleague *Reenpää* in Helsinki who first put emphasis upon this, in principle, incommensurability of psychophysical and stimulus parameters. There were many attempts to introduce some sort of approximation for magnitude scales, and the most important one certainly is that by *Stevens* and his group, based upon a very simple mathematical law, namely the power function

$$\Psi = k \cdot \Phi^n$$

When plotting stimulus intensity in a logarithmic scale versus sensation magnitude, also plotted logarithmically, with data obtained by a subjective doubling of stimulus intensity, the numerical value of the exponent n of that equation is clearly related to the modality of a given category of sensory stimuli. *Stevens* has published data, all collected by means of psychophysical experiments, which prove again and again the validity of this law and even the astonishingly small variability of these exponents. An example of this work is given in the next table and Figure 5. For

Table I: The Exponents (Slopes) of Equal-Sensation Functions, as Predicted from Ratio Scales of Subjective Magnitude, and as Obtained by Matching with Force of Handgrip

Ratio scale		Scaling by means of handgrip		
Continuum	Exponent of power function	Stimulus range	Predicted exponent	Obtained exponent
Electric shock (60-cycle current)	3.5	0.29–0.72 milliampere	2.06	2.13
Temperature (warm)	1.6	2.0–14.5 °C above neutral temperature	0.94	0.96
Heaviness of lifted weights	1.45	28–480 grams	0.85	0.79
Pressure on palm	1.1	0.5–5.0 pounds	0.65	0.67
Temperature (cold)	1.0	3.3–30.6 °C below neutral temperature	0.59	0.60
60-cycle vibration	0.95	17–47 db re approximate threshold	0.56	0.56
Loudness of white noise	0.6	55–95 db re 0.0002 dyne/cm^2	0.35	0.41
Loudness of 1000-cycle tone	0.6	47–87 db re 0.0002 dyne/cm^2	0.35	0.35
Brightness of white light	0.33	56–96 db re 10^{-10} lambert	0.20	0.21

practical use, the most important scale of all of those power functions undoubtedly is the *sone*-scale in audition.

Using the averaging technique, we were able to make electrophysiological recordings of cortical evoked responses in man which were clearly dependent upon intensity as well as other stimulus parameters. We started using auditory stimuli of tonal character and succeeded in detecting that only one of the several deflections within the very complex evoked potential seemed to contain the auditory information rather than signifying just the onset of stimulation. It was the so-called late response, somewhere around 90 to 150 msec after stimulus onset, which, when plotted also double logarithmically in terms of size of potential versus stimulus intensity, yielded a power function, with exponent, n, corresponding closely to that of the psychophysically obtained data. This holds even when different frequencies of the sinusoidal tones are used. A set of records of this type is shown in Figure 6a, and the total set of

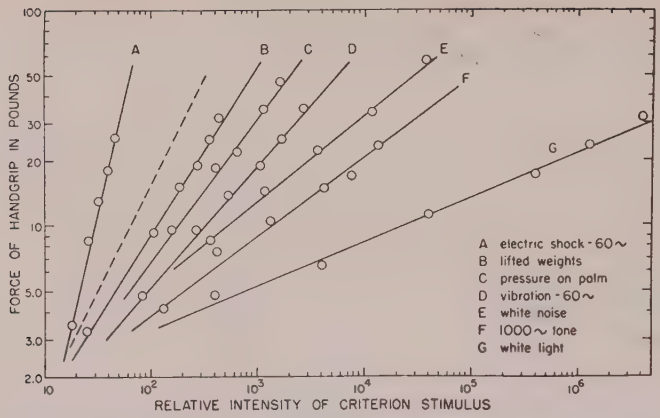


Fig. 5. *Stevens'* power-functions for different sensory modalities measured psychophysically in man. Their exponent n is greatest for pain and smallest for vision According to *S. S. Stevens.* Measured by comparison with force of handgrip. Conversion to magnitude estimation includes a factor of 1.7.

neurophysiologically obtained facts is outlined in the attached three-dimensional plot of Figure 6b. More details about these results which enable one to find objective threshold curves in deaf children and in mentally disturbed adults, are described elsewhere (see references). It was a great pleasure for us to learn at the Collegium ORLAS meeting of 1964 in Würzburg that *Hallowell Davis* and his group in St. Louis obtained very similar results using clicks.

In the meantime we examined other sensory modalities and obtained objective values of the exponent n for the neurophysiological power functions. We continued using *vibratory* stimuli of the type of Gaussian pulses of a duration of 75 msecs (1% amplitude), consisting of sinusoidal vibratory stimuli of frequencies between 50 and 400 cps. Here it could be found that the objective vibratory threshold seems to compare well with the subjective thresholds obtained by *Hugony* and *Setzepfand* years ago. The exponent n then is clearly greater than that in audition, being somewhere around 0.52 compared with 0.3 and less in audition. This is shown in Figure 7 for vibration. More details about that can be found in our paper in Pflügers Archiv (*Ehrenberger et al.*).

On the other hand *Spreng* and *Ichioka* (1964) succeeded in recording evoked potentials in man when delivering *painful electrical stimuli* to the teeth. For this modality the exponent n is of the order of 3.1. This

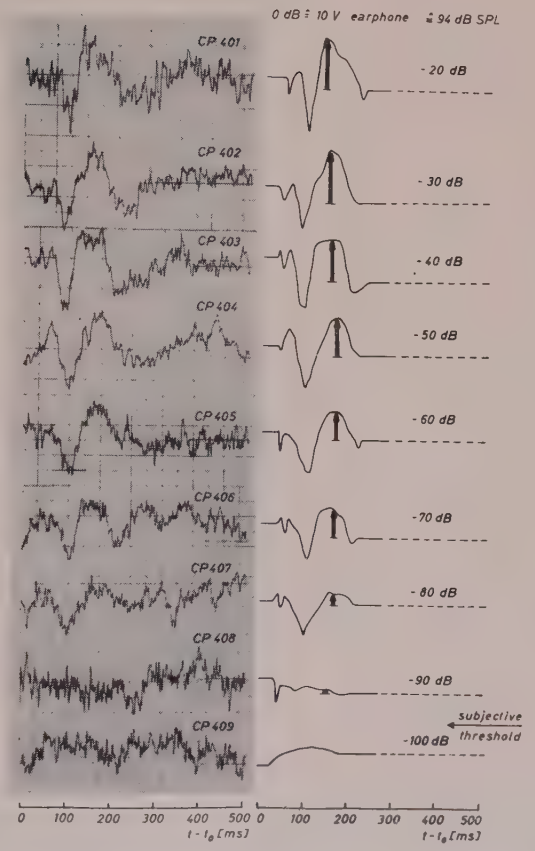


Fig. 6(a). A series of averaged evoked potentials in man in response to auditory stimuli of different intensities.

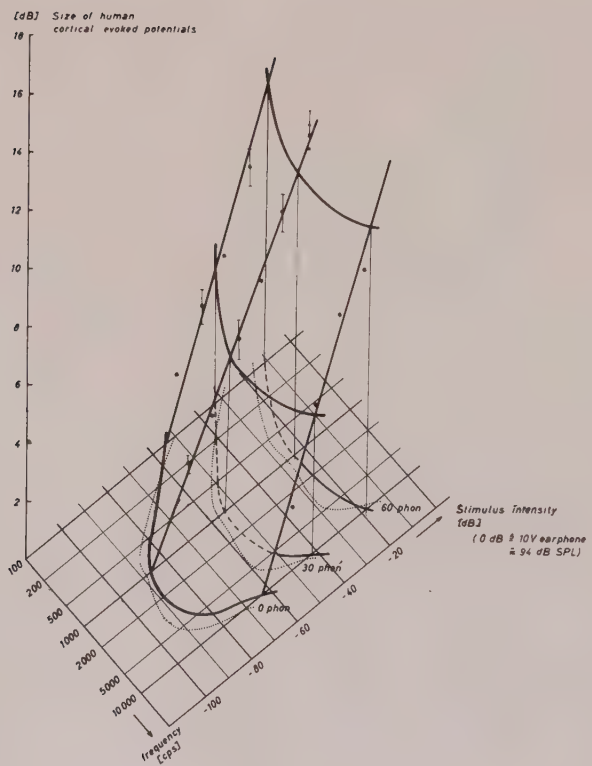


Fig. 6(b). Plots of the size of those potentials in a double-logarithmic scale result in power-functions which differ for different frequencies of the sinusoidal tones used as stimuli. The extrapolation of these intensity functions to zero yields the "objective threshold curves" in audition. According to *Keidel* and *Spreng* (1965).

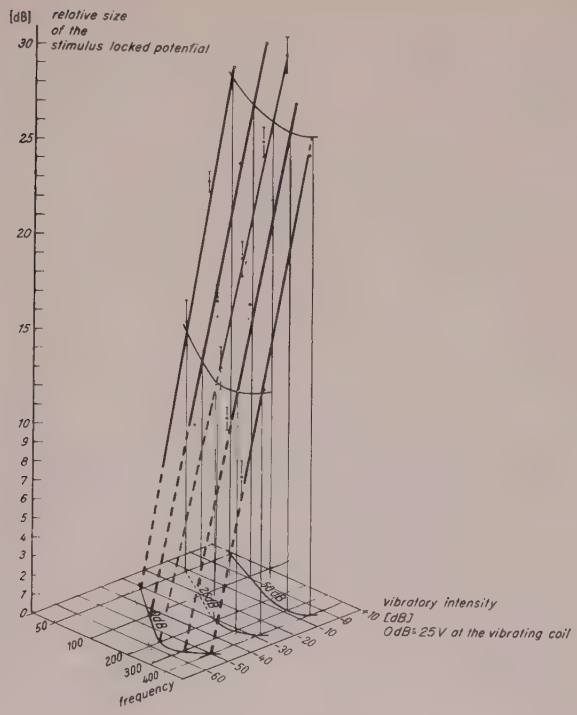


Fig. 7. Power functions neurophysiologically obtained in man by plotting the size of potentials versus stimulus intensities. The dotted lines are extrapolated. According to *Ehrenberger*, *Finkenzeller*, *Keidel* and *Plattig*.

means that the dynamic range of this modality is very small, the just suprathreshold stimulus intensity being only a few decibels lower than that for maximal amplitudes of the evoked deflections. In other words, in these cases an approximately all-or-none law is valid.

Just the opposite behavior, a very large dynamic range and therefore a great difference between maximal and minimal intensities which change the amplitude of evoked potentials, can be recorded in *vision* when using long duration monochromatic constant stimuli (not flashes) of different wavelengths. These experiments yielded, as shown in Figures 8a and 8b, an average exponent around 0.21 (standard deviation 0.04).

Plotting all data together for such different sensory modalities as pain, vibration, audition and vision makes a comparison of the power law

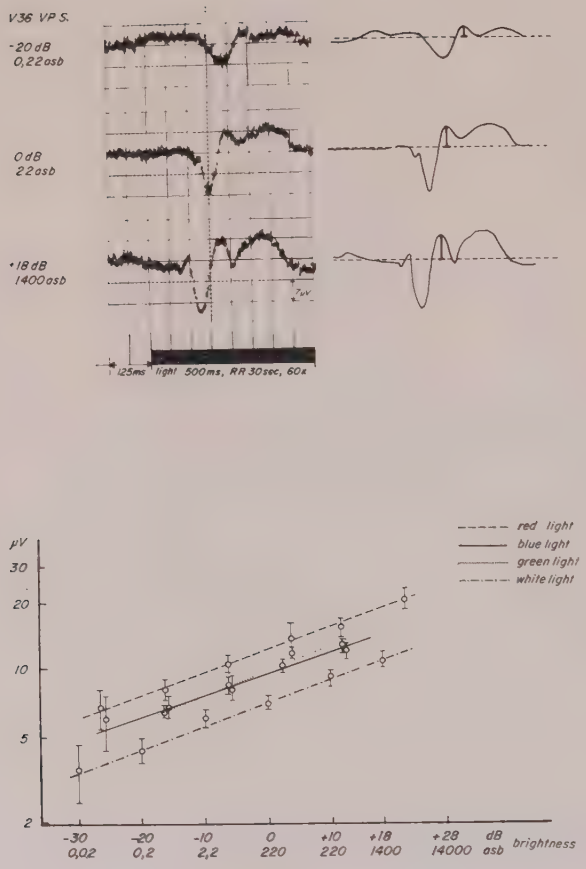


Fig. 8. *Figure a:* Averaged cortical evoked potentials to visual stimuli. *Figure b:* The resulting power functions for different wavelengths of visual stimuli compared with that obtained psychophysically by Stevens. According to *v. Loewenich.*

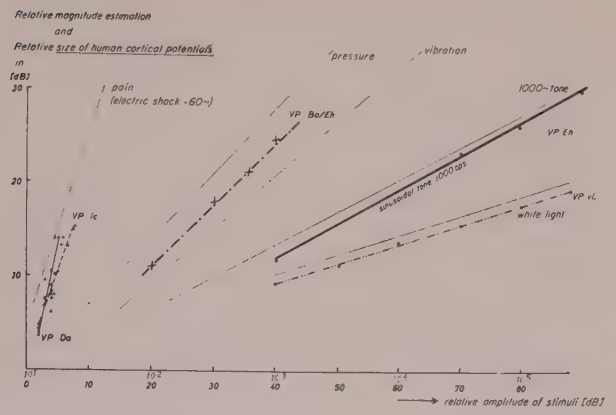


Fig. 9. All the neurophysiologically obtained power functions (solid lines) compared with those of psychophysical measurements of *Stevens* (dotted lines). The exponents in both cases are different and typical for each sensory modality. Psychophysical scaling by means of handgrip. See table I and Fig. 5.

exponents, obtained both neurophysiologically and psychophysiologically, possible, as can be seen in Figure 9.

It can be seen easily that the dynamic range for different sensory modalities, and therefore the reciprocal value of n of the power functions, differs considerably and systematically for the different sensory modalities the exponents being biggest for pain and smallest for vision. But in each case the correspondence between the "subjective" and "objective" data is striking.

However, this is true only when quite a number of experimental conditions are fulfilled such as type complexity duration and repetition rate of stimuli, for instance. I cannot go into too much detail on this point in this paper, but would rather say a few words about the problem of "specificity", "semispecificity" or "nonspecificity" of those neurophysiological data. Since the potentials can be recorded over nearly the entire skull at the same amplitude, at first it seemed they would not contain any "specific" information of a given sensory modality at all, at least with respect to their morphological distribution around the brain's cortex. But since the neurophysiological experiments are related in the same way as the psychophysical ones to the different modalities—at least for some typical deflections of the complex evoked potential in man—specific information must determine their dependence upon variations of inten-

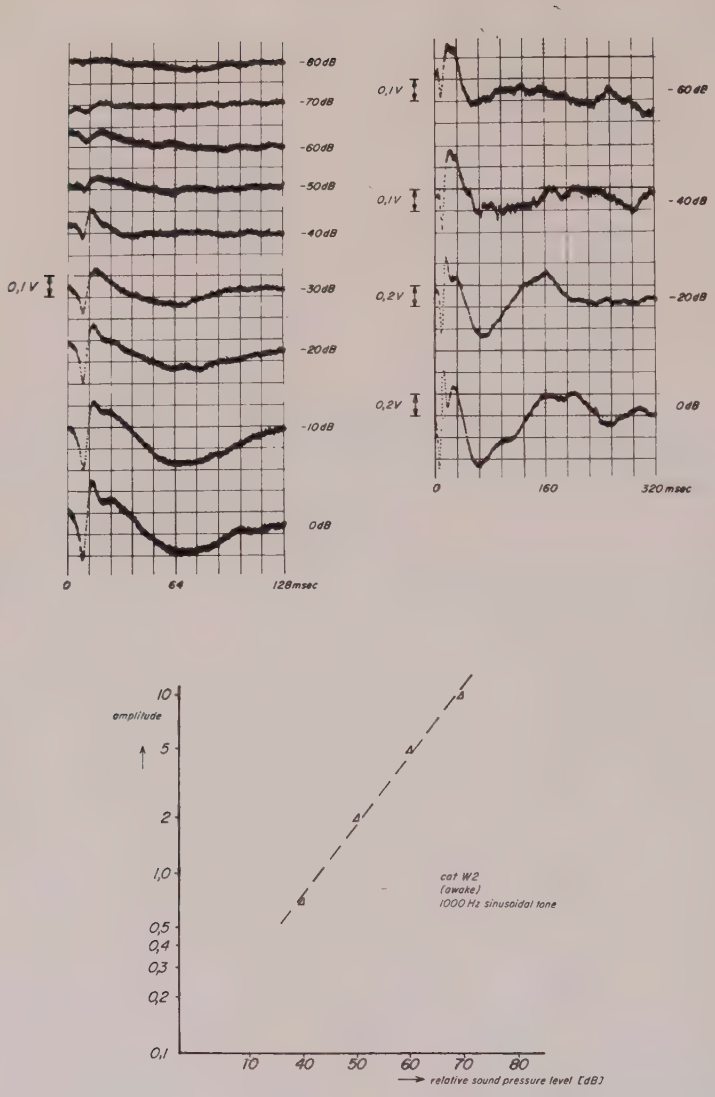


Fig. 10. *Figure a:* Cortical evoked potentials recorded by implanted electrodes in the unanaesthetized cat from cortical area A I. According to *E. David. Figure b:* Double-logarithmic plot of the late response of these potentials.

sities. Moreover, the complexity of the potentials as a whole was examined by *Bickford* who demonstrated that the early part of the response was of a muscular origin and could be recorded from nearly any part of the body above any muscle far away from the brain. The same – on the other hand—is *not* true for the late response which is restricted clearly to the skull. In this connection it might be of interest that with implanted electrodes it is possible to obtain comparable responses, with respect to their latencies and their dependence upon stimulus intensity only above sensory projection areas. Electrodes positioned above A I in the cat led to responses in the unanaesthetized animal, which are shown in Figure 10 as original curves (part a) and as a log-log plot (part b) with the exponent of its power function the same as was found in man (0.35). These experiments were performed at our department by *Dr. E. David.*

Another example of the specificity or at least the semispecificity (there might be some content of motor activity from the frontal and prefrontal cortex) can be obtained when *masking stimuli* are added to the normally unmasked stimulation. When delivering Gaussian pulses of 25 msec duration consisting of sinusoidal tones of 1 to 2 kc to one ear, variation of stimulus intensity led to the series of averaged responses shown in Figure 11 on the left.

By adding white noise it is possible to mask subjectively the tonal pulses which can be heard by the subject. At the same time, just one of the deflections of the complex auditory evoked potential is cancelled completely, which therefore must contain some representation of, or must be controlled by, the actual "specific" auditory information introduced by the pulse (right side of Figure 11).

Further, we were able to record averaged evoked potentials to olfactory stimuli in man by using a special technique of the stimulation and electrode positioning. These potentials, recorded by *Finkenzeller* reveal a different latency which seems to be related to the diffusion processes of the olfactory stimuli. Records which show a comparison with air blasts without any odorous gas blown to the olfactory epithelium, prove the reproducibility of the effect. Results of these experiments are shown in Figure 12.

Finally it was possible for *Dr. Plattig* and *Jauhiainen* (as guest from Helsinki) to obtain again different potentials to electrically induced taste stimuli of the tongue which reveal an intensity dependence with a steep-

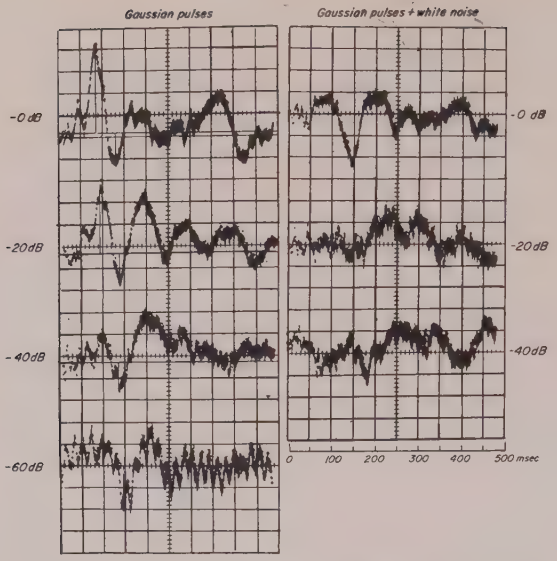


Fig. 11. Series of averaged cortical evoked potentials in man to Gaussian pulses of 25 msec duration delivered to the ear (left row). When white noise is added one typical deflection (late response) is cancelled when the pulse is masked by the noise psychophysically.

ness in a log-log-plot of somewhere around 1:0. This is shown in Figure 13.

The latter experiments are continuing.

After all, it seems to be clear that at cortical level the Stevens power functions indicate the way the sensory information is processed by the different modalities of the sensory channels in a rather "specific" manner, if certain experimental conditions are respected. Neurophysiology, on the other hand, has collected numerous facts and data about all sorts of coding processes within the sensory receptor cells of the sense organs themselves. All these experiments converge to the observation that at the peripheral end of each sensory channel, with one exception, (to be mentioned below) simple single-logarithmic relations—according to Weber's law—exist between each state of excitation (size of generator potential as well as number of spikes in a single unit) and each stimulus intensity. Figure 14a and b is just one example for visual receptors in the Limulus eye according to records of *McNichol* and his colleagues.

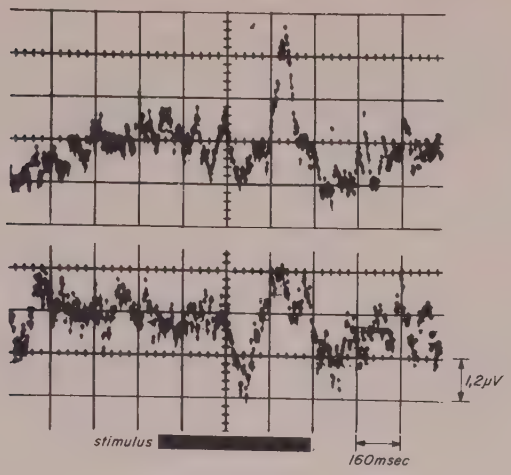


Fig. 12. Averaged evoked potentials in man to olfactory stimuli. According to *Finkenzeller*.

As many other examples as desired could be added. Such data prove that at the peripheral level, in general not the power function, (the double-logarithmic relation) but a simple single-logarithmic rule gives the best approximation of data.

However there is one exception which is worth reviewing briefly: for mechanoreceptors of the toad, sensitive to static and steplike vibratory stimulation, *Lindblom* could find a linear dependence of the impulse frequency of the nerve fiber supplying the receptor upon the displacement of the skin as shown in Figure 15a covering a displacement range of some 300 μ (*Lindblom*, 1963). On the other hand *Sato*, in 1961, was able to demonstrate that other mechanoreceptors in the cat, as the Pacinian corpuscles, behave exactly the way most other receptors do, namely in a single-logarithmic manner, according to Weber's law, as can be seen in Figure 15b.

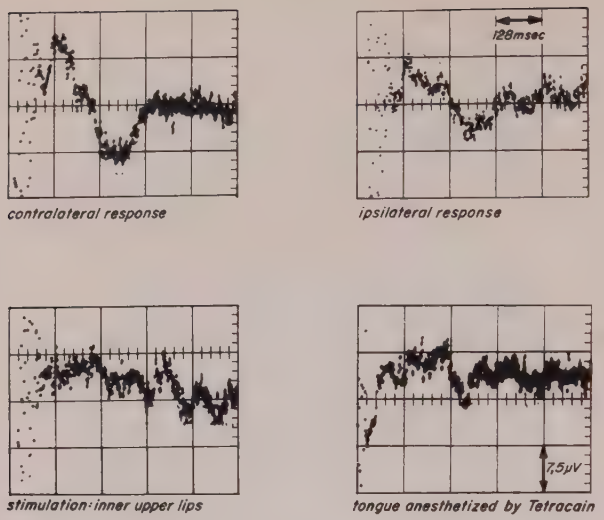


Fig. 13. Averaged evoked brain potentials in man to gustatory stimulation of the tongue. According to *Plattig* and *Jauhiainen*.

The difference is only that in the first case, the toad is stimulated mechanically, in the latter electrically. I do not think, however, that we need to drop the concept of the single-logarithmic relation of excitation versus stimulus intensity because of the only reported linear relation of toads' touch-receptor in skin. For if one considers the nonlinearity of the stiffness—indentation—function of the human skin, as it was measured by *Franke* in 1951, one should expect that a comparable nonlinearity of the toad's skin could change the linear function of Figure 15a to a normal single-logarithmic one, if this feature of the toad's skin is taken into account for the adequate stimulation of the touch receptor. So it seems fairly well established that at the receptor level a single-logarithmic law is usually the rule, while at cortical level the double-logarithmic power function is the normally found relation between state of excitation and stimulus intensity.

This raises however, questions concerning the level, within a given sensory channel, at which the logarithmic function changes to the power function and whether this change is gradual or abrupt, or whether it is only

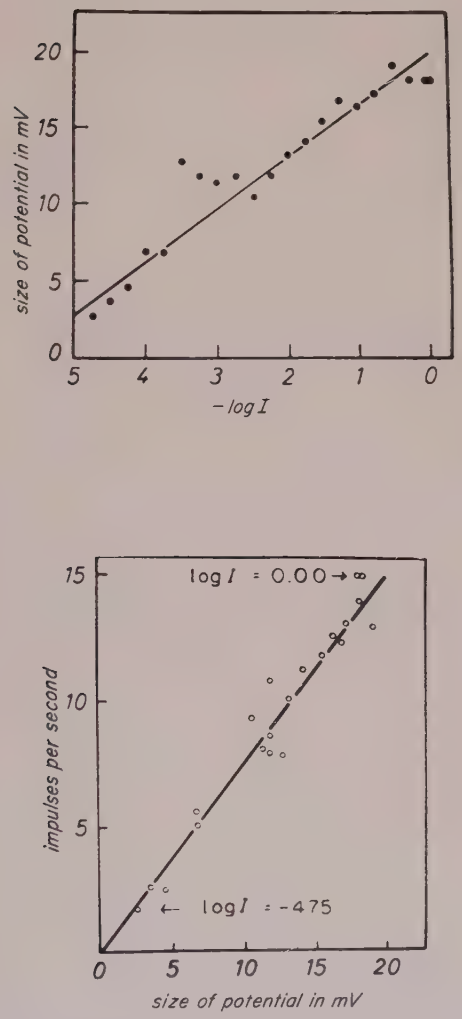


Fig. 14. *Figure a:* Size of generator potential versus stimulus intensity in Limulus reveals a single-logarithmic function similar to *Weber's* law. *Figure b:* Number of spikes per second in the nerve depends linearly upon the generator potential and single-logarithmically upon stimulus intensity. According to *McNichol.*

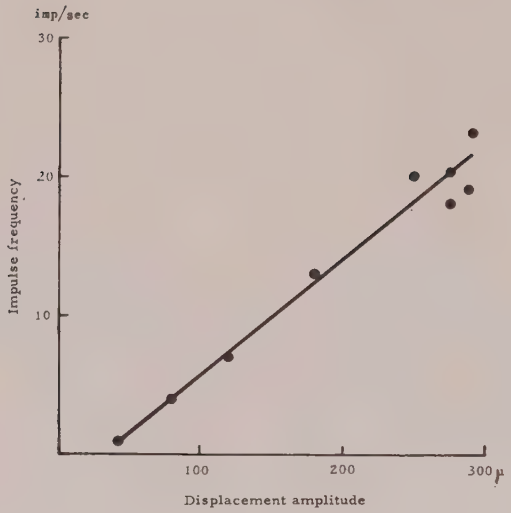


Fig. 15(a) Linear relation between the nerve impulse frequency and the skin displacement amplitude in the toad. According to *Lindblom.*

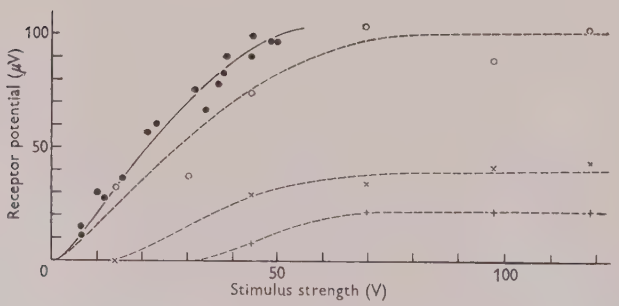


Fig. 15(b) Relation between receptor potential amplitude and intensity of vibration at various frequencies recorded on mesenteric Pacinian corpuscles of the cat. According to *Sato.*

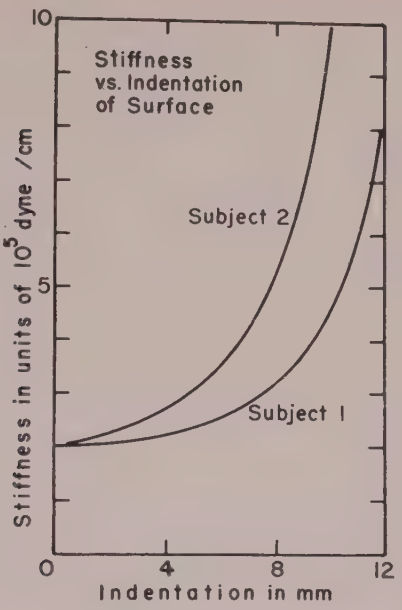


Fig. 15(c) Stiffness-indentation function of the human skin. According to *Franke* (1951) and von Gierke, Oestreicher, Franke, Parrack, von Wittern (1952).

a change associated with the decoding and information detection processes (see discussion of *Werner* and *Mountcastle*, 1965). This is exactly today's problem of research in information processing by sensory modalities. Let me conclude the discussion of this main point of my paper by presenting Figure 16.

This is obtained in the somesthetic system of macaque monkeys at the thalamic level by *Mountcastle, Poggio* and *Werner*. It shows clearly a power function with an exponent of 0.63 (stimulus: rotation of a joint) which is in fair agreement with the data obtained at cortical level neurophysiologically and in psychophysics. No clear measurements have been made up to now at the geniculate or the colliculus level of the CNS above which power functions can be observed and beneath which the Weber functions might prevail.

However, new measurements did prove that even at the receptor level for the somesthetic system (cats and monkeys; Iggo corpuscles) single cutaneous afferent can be found which obey clearly the Stevens law with exponents commonly less than 1 (Fig. 17).

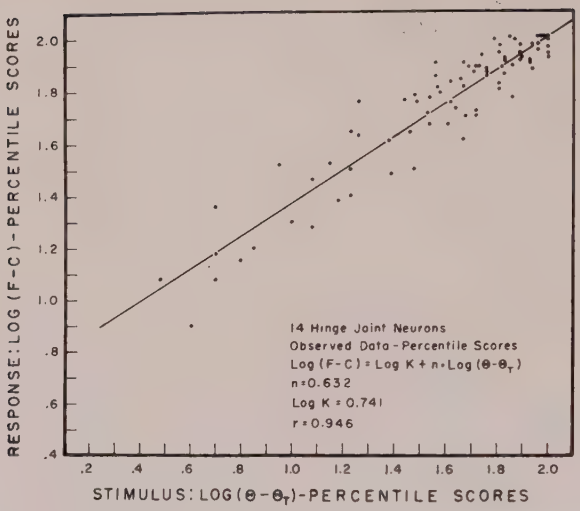


Fig. 16. Power functions of single units at the thalamic level recorded by *Mountcastle. Poggio* and *Werner.*

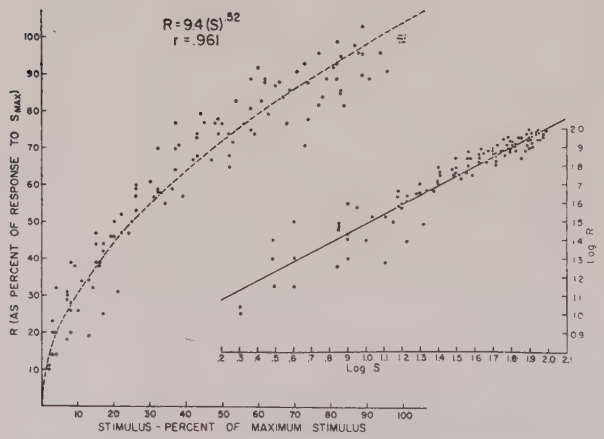


Fig. 17. Power functions of single afferent nerve fibers at the receptor level. According to *Werner* and *Mountcastle.*

At least for these receptors the conclusion can be drawn "that the serially superimposed neural transforms leading from first-order input to final output must, in sum, be linear for the intensity parameter" (*Werner* and *Mountcastle*, 1965). Since, however, some sensory channels of certain animals at other modalities do show Weber's law at the first order input, in good approximation, it is the problem of determining the site of the changeover from Weber to power law which must be solved in order to get a better insight into the information processing by sensory modalities.

REFERENCES

v. Békésy, G. Interaction of paired sensory stimuli and conduction in peripheral nerves. *J. Applied Physiology*, **18**, 1276–1284 (1963).

Davis, H. Slow Cortical Responses Evoked by Acoustic Stimuli (paper given at Collegium ORLAS 1964 Würzburg). *Acta Otolaryngol.* (Stockholm) **59**, 179–185 (1965).

Ehrenberger, K., Finkenzeller, P., Keidel, W. D., and Plattig, K. H. Elektrophysiologische Korrelation der Stevensschen Potenzfunktion und objektive Schwellenmessung am Vibrationssinn des Menschen. *Pflügers Archiv* **290**, 114–123 (1966).

Finkenzeller, P. Gemittelte corticale Potentiale bei olfactorischer Reizung, 31. *Tagung der Deutsch. Physiol. Ges. Pflügers Archiv* (1966), Referat 40 R. 27.

Jauhiainen, T., und Plattig, K. H. Reizsynchrone langsame Rindenpotentiale beim Menschen nach elektrischer Reizung der Zunge. 31. *Tagung der Deutsch. Physiol. Ges. Pflügers Archiv* (1966) Referat 41, R 27.

Keidel, W. D., and Spreng, M. Neurophysiological Evidence for the Stevens Power Function in Man. *J. Acoust. Soc. Amer.* **38**, 191–195 (1965).

Keidel, W. D., und Spreng, M. Elektronisch gemittelte langsame Rindenpotentiale des Menschen bei akustischer Reizung (paper given at the Collegium ORLAS, Athens, 1962) *Acta Otolaryngol.* (Stockholm) **56**, 318–328 (1963).

Keidel, W. D., and Spreng, M. Audiometric aspects and multi-sensory Powerfunctions of Electronically averaged slow evoked cortical responses in Man (paper given at the Collegium ORLAS, Würzburg 1964) *Acta Otolaryngol.* (Stockholm) **59**, 201–208 (1965).

Lindblom, U. Phasic and Static Excitability of Touch Receptors in Toad Skin. *Acta physiol. scand.* **59**, 410–423 (1963).

v. Loewenich, U. Leuchtdichtenabhängigkeit der menschlichen corticalen Reizantworten ohne und mit Einfluß der Adaptation. 31. *Tagung der Deutsch. Physiol. Ges. Pflügers Archiv* (1966) Referat 44, R 29 and Pflügers Archiv **293**, 256–271 (1967).

Mountcastle, V. B., Poggio, G. G., and Werner, G. The Relation of Thalamic Cell Response to Peripheral Stimuli Varied Over an Intensive Continuum. *J. Neurophysiology* **26**, 807–845 (1963).

Sato, M. Response of Pacinian corpuscles to Sinusoidal Vibration. *J. Physiol.* (London) **159**, 391–409 (1961).

Stevens, S. S. The Psychophysics of sensory function. In: Sensory Communication (W. D. Rosenblith) Wiley, New York, 1961.

Spreng, M., and Ichioka, M. Langsame Rindenpotentiale bei Schmerzreizung am Menschen. *Pflügers Archiv* **279**, 121–132 (1964).

Werner, G., and Mountcastle, V. B. Neural activity in Mechanoreceptive Cutaneous Afferents: Stimulus-Response Relations, Weber Functions, and Information Transmission. *J. Neurophysiology* **28**, 359–397 (1965).

Franke, E. K. Mechanical Impedance Measurements of the Human Body Surface. AF Technical Report No. 6469 (1951).

von Gierke, H. E., H. L. Oestreicher, E. K. Franke, H. O. Parrack, W. W. von Wittern. Physics of Vibrations in Living Tissue. *J. Applied Physiology* **4**, 886–900 (1952).

Macy, I., Jr. A Probability Model for Cortical Responses to Successive Auditory Clicks. Thesis, Massachusetts Inst. Technol., Cambridge, Massachusetts (1954).

McNichol, E. F. Visual Receptors as Biological Transducers. In *Molecular Structure and Functional Activity of Nerve Cells*. Edited by R. G. Grenell and L. I. Mullins, American Institute of Biological Sciences, Washington, D. C. (1956).

I. C. WHITFIELD

and S. D. COMIS

Neurocommunications Research Unit,

University of Birmingham,

Birmingham, England

A Reciprocal Gating Mechanism in the Auditory Pathway

It is a matter of common observation that sensory stimuli which are well above threshold may go unnoticed if they are very familiar, or if attention is concentrated on some other sensory input. Thus we are often unaware of the ticking of a clock in the room with us until it stops, or until our attention is drawn to it in some way. Although the first result might be explicable in terms of adaptation of the receptors, the second clearly is not; shifts of attention must, we suppose, involve the nervous system itself. Until the development of electrical recording, it was not possible to determine how far up the sensory pathway an "unheard" response travels, and even when such techniques were established, it was further necessary that the investigations should be carried out on intact animals, since shifts of attention are not meaningful for the anaesthetized preparation.

In 1955, Hernández-Péon and Scherrer reported that the gross electrical wave response to a click stimulus, recorded from the cochlear nucleus of a cat, was greatly diminished if the animal's attention was diverted by showing it a mouse (Hernández-Péon and Scherrer, 1955) and in the same year Galambos, Sheatz and Vernier (1955) showed that this click response gradually diminished if the clicks were presented repeatedly over a period of hours; they further showed that the response could be restored at will by conditioning the animal to attend to the click by means of a mild electric shock. Since conditioning phenomena are not known to occur at the medullary or spinal levels, these observations suggested the existence of centrifugal pathways from higher centres, reaching at least as far peripherally as the cochlear nucleus.

The existence of such pathways had, in fact, been known for a great many years. As long ago as 1893, Held observed "recurrent" neurones entering the cochlear nucleus, and Lorente de Nó (1933) described the termination of such a pathway in the dorsal cochlear nucleus. Neither of these workers, however, located the source of the fibres.

Most of our knowledge of the details of these pathways comes from the work of Rasmussen. In 1946 he drew attention to the existence of an efferent component in the cochlear nerve, and, due to its origin in or near the superior olive, named it the olivo-cochlear bundle. Since then, further work (Rasmussen, 1955, 1958; Desmedt and Mechelse, 1958, 1959) has established anatomically the existence of a centrifugal fibre system paralleling the afferent auditory system throughout its entire length (Fig. 1).

Galambos (1956) first related the anatomical and behavioural findings by showing that electrical stimulation of the olivo-cochlear bundle in the floor of the 4th ventricle resulted in the suppression of the gross auditory nerve response to clicks. Evidently, activation of this pathway could by itself account for the behaviourally-induced reduction in activity in the

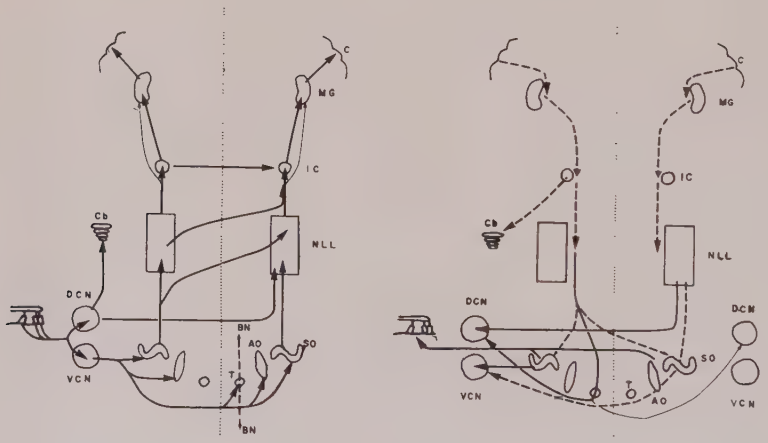


Fig. 1. The centripetal (left) and centrifugal (right) connections of the auditory system. Pathways not completely characterized anatomically are shown by dashed lines. DCN, dorsal cochlear nucleus; VCN, ventral cochlear nucleus; Cb, cerebellum; SO, lateral superior olivary nucleus; AO, accessory olive; T, nucleus of trapezoid body; NLL, nuclei of lateral lemniscus; IC, inferior colliculus; MG, medial geniculate body; C, cortex (from Whitfield, 1966b).

cochlear nucleus referred to above. However, Desmedt (1960) showed that electrical stimulation in the region of the lemniscal pathway could reduce the gross response in the contralateral cochlear nucleus while leaving the auditory nerve response unaffected, an observation which indicated a direct effect on the nucleus itself.

So far, the results of behavioural and physiological studies appeared to tie up rather well. However, Worden and Marsh (1965) were not able entirely to confirm the earlier behavioural results on the cochlear nucleus. They placed electrodes in several locations in both dorsal and ventral nuclei, but could find no correlation between placement and increase or decrease in the response; they were unable in some cases even to obtain consistent results on successive trials in the same location. Other groups of workers (Marsh, McCarthy, Sheatz and Galambos, 1961; Webster, 1962; Dunlop, Webster and Day, 1964) likewise obtained conflicting results in other nuclei along the auditory pathway.

It was becoming clear that the control system is not just a simple mechanism for shutting down the entire pathway in order to block unwanted inputs, but is a much more selective system, for which study of the summed response in the totality of input channels is far too coarse an index. The results just cited would, in fact, be more consistent with the occurrence of mixed inhibitory and excitatory fibres in the centrifugal system.

The first single unit studies were carried out by Fex (1962) on the olivo-cochlear bundle. His results confirmed that this bundle is uniformly inhibitory. However, Comis and Whitfield (1966), studying the effect of various drugs applied locally to single neurones of the cochlear nucleus, found that not only could a high proportion of neurones in the antero-ventral division of the nucleus be activated by acetylcholine, but that this substance markedly reduced the threshold of the neurone to peripheral sound stimuli, often by as much as 15 db. (Fig. 2). This nucleus, in the cat, gives a strong histochemical reaction for cholinesterase, and section of the auditory nerve does not result in the disappearance of the enzyme (Whitfield, 1967a). This is in accordance with the observation of Rossi (1961) that the afferent fibres of the auditory nerve are not cholinergic. Shute and Lewis (1965), on the other hand, have observed in the rat that lesions placed *central* to the dorsal cochlear nucleus result in disappearance of cholinesterase from that nucleus.

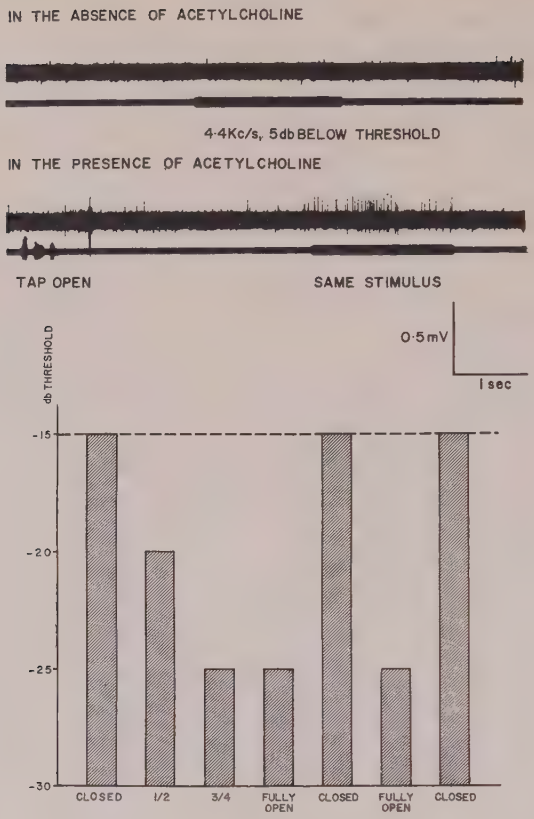


Fig. 2a. Upper trace: unit in antero-ventral cochlear nucleus responsive to a tone of 4.4 Kc/s, stimulated with this tone at an intensity 5 db below threshold, and therefore not responding; Lower trace: same unit in the presence of acetylcholine (tap open). The unit now responds to the previously sub-threshold stimulus.

Fig. 2b. The effect of acetylcholine on the threshold of response of a neurone in the antero-ventral cochlear nucleus to sound stimuli. When the tap is closed, the normal threshold is —15 db. Opening the tap and applying acetylcholine locally to the neurone lowers progressively the threshold to sound as the rate of application is increased. "Tap fully open" represents a rate of about 5×10^{-12} moles/sec. The vertical columns are for immediately successive records.

It seemed likely therefore that the application of acetylcholine was mimicking the action of a centrifugal pathway rather than that of the afferent, centripetal endings. Several such pathways terminating in the

cochlear nucleus have been described by Rasmussen (1960). Two of these pathways, one of which originates in the ipsilateral and the other in the contralateral lemniscal nuclei, terminate in the dorsal nucleus. A third originates in the lateral superior olivary nucleus (S-segment) and terminates in the ventral cochlear nucleus of the same side. Desmedt (1960) has described a pathway from the vicinity of the ventral nucleus of the lateral lemniscus to the contralateral ventral cochlear nucleus. Some of the loops formed by these pathways are summarized in Figure 3. It is the pathway from the lateral superior olive to the ventral cochlear nucleus in which we have been specially interested.

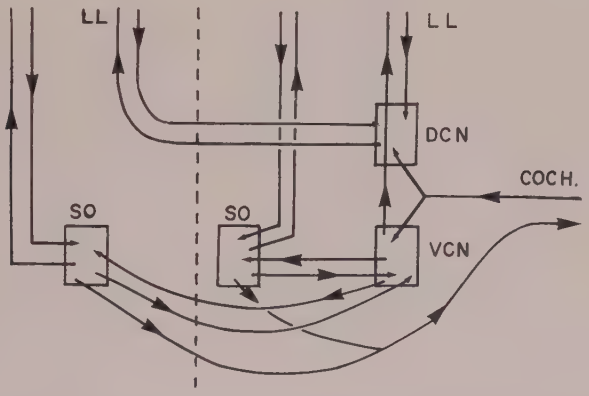


Fig. 3. Feedback connections of the cochlear nuclei COCH, cochlea; DCN, dorsal cochlear nucleus; VCN, ventral cochlear nucleus; SO, superior olivary nuclei; LL, lateral lemniscus. Midline indicated by dashes.

We have stimulated this pathway by means of an electrode inserted from the ventral aspect of the brainstem into the medial portion of the S-segment of the superior olive, while at the same time recording from single neurones in the ipsilateral antero-ventral cochlear nucleus. Direct current stimulation of cell bodies was used in order to minimise the possibility of antidromic stimulation of the centripetal fibres. The effect of such stimulation on the peripheral sound threshold is entirely comparable with the previously found effect of acetylcholine (Fig. 4). We have thus the first example of a centrifugal auditory pathway which enhances, rather than diminishes, the response to a peripheral stimulus.

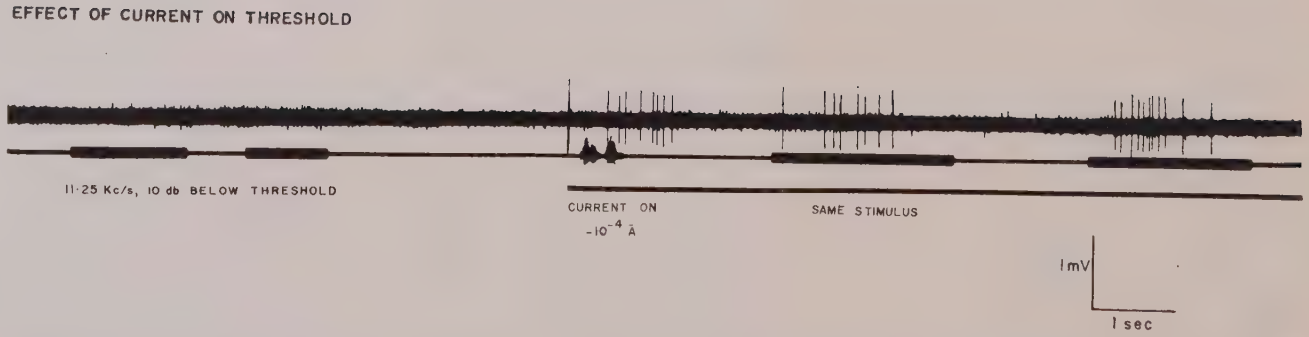


Fig. 4. Unit in the ventral cochlear nucleus which responds to a suitably intense sound stimulus of 11.5 kc/s. When this tone is presented at 10 db below threshold (thick bars of middle trace), there is, of course, no response. Stimulation of the ipsilateral superior olive with direct current (lowest trace) reduces the threshold of the neurone, which now responds to the previously sub-threshold sound stimulus.

If the effects of activity in the pathway are mediated by release of acetylcholine at its terminals, then it should be possible to block these effects with a cholinergic blocking agent such as atropine or dihydro-β-erythroidine. This can indeed be done, and Figure 5 shows the effect of atropine in blocking the response to stimulation of the superior olive. As might be expected, these drugs also block the response of the cochlear nucleus neurone to a previously effective sound stimulus (Fig. 6). It is postulated that this occurs because the gate, normally held open by activity in the centrifugal pathway, is thereby closed.

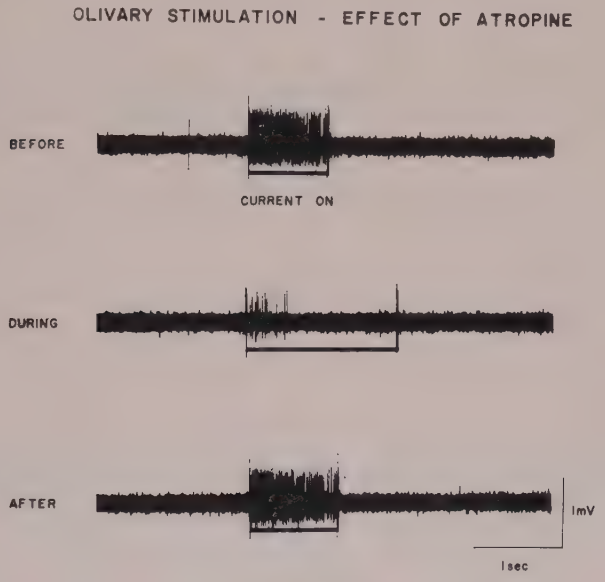


Fig. 5. Response of a neurone in the antero-ventral cochlear nucleus to direct current stimulation of the ipsilateral superior olive. When atropine is locally applied to the neurone its response is almost abolished, but recovers some minutes after the termination of drug application.

It would appear, therefore, that a double gating system is present in the cochlear nucleus. The pathway from the lateral superior olivary nucleus to the ventral cochlear nucleus is the "normally-closed" half of this gating system, and activity in its fibres is necessary to allow the neurones to respond. The pathway from the ventral lemniscal nuclei,

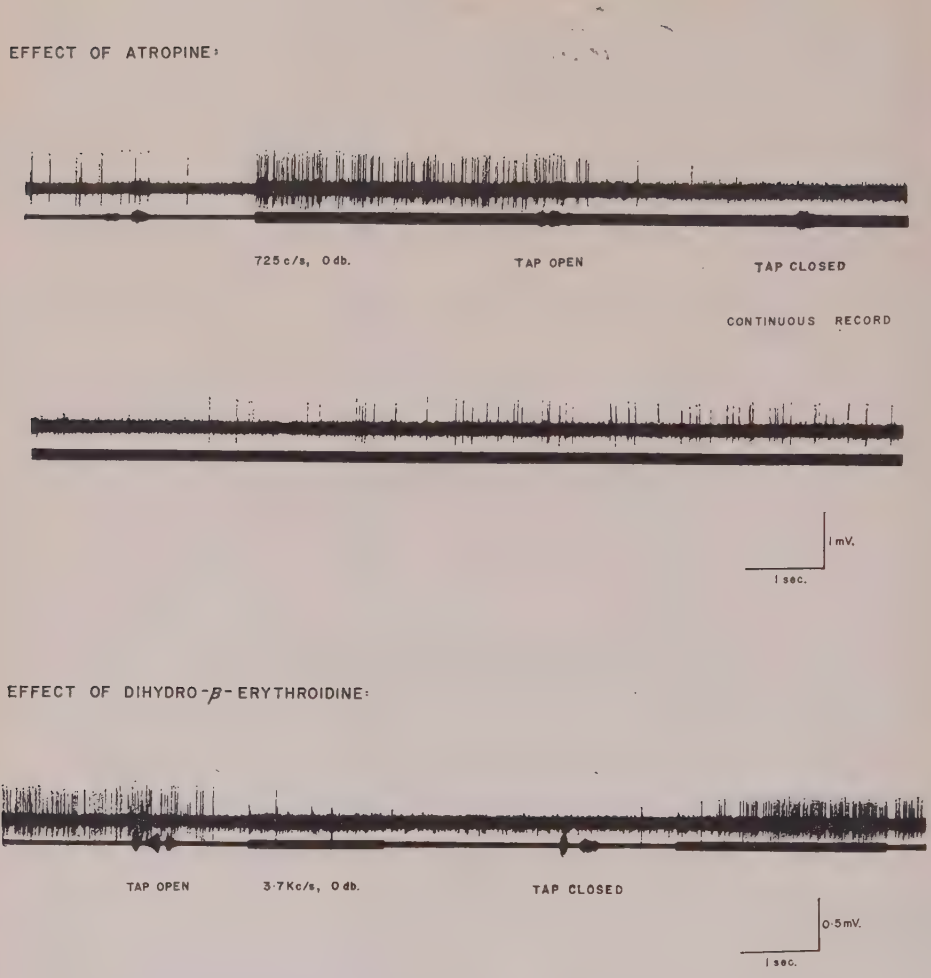


Fig. 6. The effect of blocking agents on the response of neurones in the antero-ventral cochlear nucleus to a sound stimulus. Above: response of a neurone to a continuous tone of 725 c/s. At "tap open" atropine was applied directly to the neurone, causing a failure of the response. After the tap is closed the response gradually recovers. The upper and middle traces form a continuous record. Below: the effect of di-hydro-β-erythroidine in blocking both "spontaneous" activity and activity in response to a tonal stimulus. Different unit from that illustrated above.

activation of which suppresses the centripetal response, is possibly the other half, although we have not so far been able to work out the details. Nevertheless we have foundthat whereas acetylcholine lowers the threshold to peripheral stimulation, *nor*-adrenaline has the reverse effect, and raises the threshold of about 80% of neurones in the ventral cochlear nucleus (Fig. 7). Although there are not yet sufficient lines of evidence to characterize the transmitter, the results are certainly suggestive.

We wish to postulate then, that there are at least two types of fibre which reach and control the neurones of the cochlear nucleus. One of these operates a "normally-open" gate, while the other operates a "normally-closed" gate. The balance between them determines the degree of peripheral activity which is allowed to proceed to higher levels. Thus the centrifugal pathways do not act simply to shut down the whole system, but control its activity in a precise and probably detailed way. The fineness of this detail remains, however, to be determined, and will require a point-to-point analysis of the distribution of the terminals.

It should be noted that the feed-back loop involved in the normally-closed gate system is potentially a very short one, since centripetal fibres run directly from the cochlear nucleus to the S-segment, and centrifugal fibres run directly back from the S-segment to the cochlear nucleus (Fig. 3). It is not known if the two connect directly within the S-segment itself. Even if there is such a direct connection, however, the loop delay would be quite long, owing to the latencies involved in the response of the centrifugal link of the loop. The response latency following stimulation of the olivary site is at least 30 msec even for strong stimuli; it is longer still for less strong stimuli. It seems unlikely, therefore, that the loop plays any part in the elaboration of the detailed pulse interval pattern, but rather that it tends to control the mean level of activity by integration over a short period. This would be in accordance with our view that the sequential distribution of pulse intervals is not relevant to information transmission, and supports the idea that the feedback paths serve to regulate the activity or inactivity of particular channels.

SUMMARY

1. It is possible to obtain either increase or decrease in the response of neurones in the cochlear nucleus to sound signals by stimulation of appropriate centrifugal pathways.

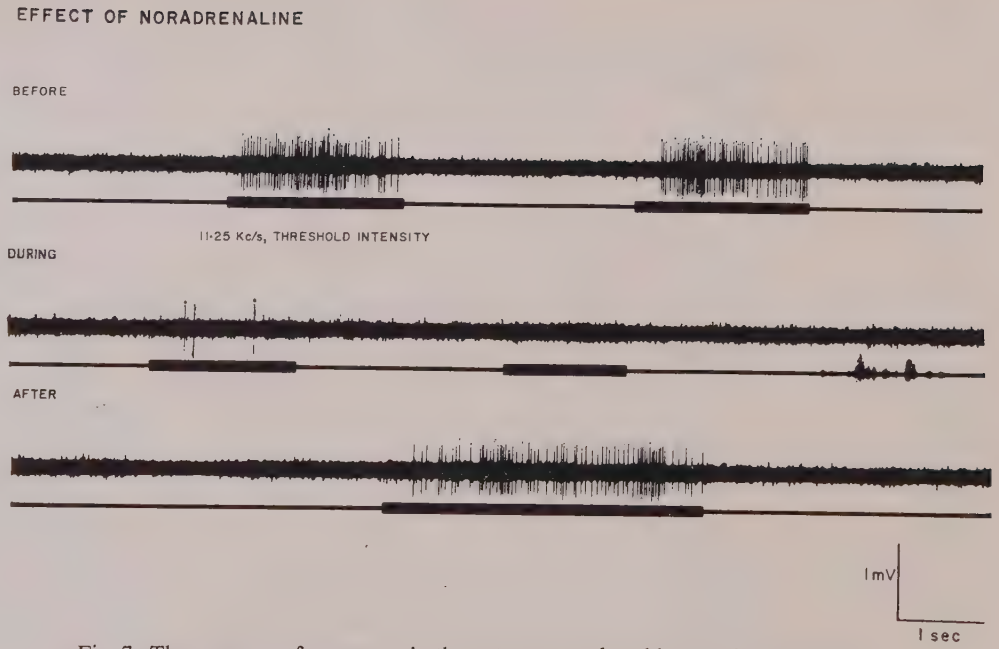


Fig. 7. The response of a neurone in the antero-ventral cochlear nucleus to a sound stimulus, before, during and some minutes after the local application of *nor*-adrenaline at the recording site. The intensity and frequency of the stimulus are the same throughout.

2. Lowering of the sound threshold can be brought about by stimulation of the lateral superior olivary nucleus, and this pathway appears to be cholinergic since the effects of stimulation can be mimicked by local application of acetylcholine and blocked by cholinergic blocking agents. The reverse effect, raising of the sound threshold, can be brought about by local application of *nor*-adrenaline.
3. The latency of the effect of centrifugal stimulation is quite long, which suggests that the pathways control the short-term level of activity rather than influence the details of the nerve impulse distribution.
4. It is proposed that a dual gating system operates on the cochlear nucleus to control the sensory input, one being a "normally-open", and the other a "normally-closed" gate.

ACKNOWLEDGEMENT

The research reported herein has been sponsored by the Air Force Office of Scientific Research under Grant AF EOAR 63-115 through the European Office of Aerospace Research (OAR) United States Air Force.

REFERENCES

1. Comis, S. D., and Whitfield, I. C. (1966). The effect of acetylcholine on neurones of the cochlear nucleus. *J. Physiol.* **183**, 22–23 P.
2. Desmedt, J. E. (1960). Neurophysiological mechanisms controlling acoustic input. Chap. 11 *Neural Mechanisms of the Auditory and Vestibular Systems.* Ed.G.L.Rasmussen and W. F. Windle. Springfield: Thomas.
3. Desmedt, J. E., and Mechelse, K. (1958). Suppression of acoustic input by thalamic stimulation. *Proc. Soc. exp. Biol. Med.* **99**, 772–775.
4. Desmedt, J. E., and Mechelse, K. (1959). Corticofugal projections from temporal lobe in cat and their possible role in acoustic discrimination. *J. Physiol.* **147**, 17–18 P.
5. Dunlop, C. W., Webster, W. R., and Day, R. H. (1964). Amplitude changes of evoked potentials at the inferior colliculus during acoustic habituation. *J. aud. Res.* **4**, 159–169.
6. Fex, J. (1962). Auditory activity in centrifugal and centripetal cochlear fibres in cat. *Acta physiol. scand.* **55**, Suppl. 189.
7. Galambos, R (1956). Suppression of auditory nerve activity by stimulation of efferent fibers to cochlea. *J. Neurophysiol.* **19**, 424–437.

8. Galambos, R., Sheatz, G., and Vernier, V. G. (1956). Electrophysiological correlates of a conditioned response in cats. *Science, N.Y.*, **123**, 376–377.
9. Held, H. (1893). Die centrale Gehörleitung. *Arch. Anat. Physiol.* Anat. Abt. 201–248.
10. Hernández-Péon, R., and Scherrer, H. (1955). Habituation to acoustic stimuli in cochlear nucleus. *Fedn. Proc.* **14**, 71.
11. Lorente de Nó, R. (1933). Anatomy of the eighth nerve. III. General plan of structure of the primary cochlear nuclei. *Laryngoscope, St. Louis*, **43**, 327–350.
12. Marsh, J. T., McCarthy, D. A., Sheatz, G. and Galambos, R. (1961). Amplitude changes in evoked auditory potentials during habituation and conditioning. *Electroen. Neurophysiol.* **13**, 224–234.
13. Rasmussen, G. L. (1946). The olivary peduncle and other fiber projections of the superior olivary complex. *J. comp. Neurol.* **84**, 141–220.
14. Rasmussen, G. L. (1955). Descending or "feedback" connections of auditory system of the cat. *Am. J. Physiol.* **183**, 653.
15. Rasmussen, G. L. (1958). Anatomical discussion of "Neural mechanisms in audition" by R. Galambos. *Laryngoscope, St. Louis*, **68**, 404–406.
16. Rasmussen, G. L. (1960). Efferent fibers of the cochlear nerve and cochlear nucleus. Chap. 8 *Neural Mechanisms of the Auditory and Vestibular Systems*. Ed. G. L. Rasmussen and W. F. Windle. Springfield: Thomas.
17. Rossi, G. (1961). L'acétylcholinesterase au cours du developpement de l'oreille interne du cobaye. *Acta oto-lar. Suppl.*. 170, 1–91.
18. Shute, C. D. D., and Lewis, P. R. (1965). Cholinesterase-containing pathways of the hindbrain: afferent cerebellar and centrifugal cochlear fibres. *Nature*, **205**, 242–246.
19. Webster, W. R. (1962). An empirical study of the electrophysiological correlates of the concepts of attention and habituation. Unpublished B.A. (Hons.) Thesis, Univ. Sydney, Australia. [cited by Dunlop, Webster and Day, 1964].
20. Whitfield, I. C. (1967a). The pharmacological behaviour of the cochlear nucleus. *Drugs and Sensory Functions*. Ed. A. Herxheimer. London: Churchill.
21. Whitfield, I. C. (1967b). *The auditory pathway*. Monographs of the Physiological Society. London: Arnold.
22. Worden, F. G., and Marsh, J. T. (1963). Amplitude changes of auditory potentials evoked at cochlear nucleus during acoustic habituation. *Electroen. Neurophysiol.* **15**, 866–881.

R. C. GESTELAND

Technological Institute

Northwestern University

Evanston, Illinois

J. Y. LETTVIN

W. H. PITTS

and S-H. CHUNG

Research Laboratory of Electronics

Massachusetts Institute of Technology

Cambridge, Massachusetts.

A Code in the Nose

Our central notion about receptor cells in the nose of vertebrates and chemosensory structures of invertebrates is that every cell has a different ordering principle or point of view with respect to the space of odors. The olfactory system constructs a space of many dimensions (no more than the number of different cells) but with low resolution in any dimension. This is a somewhat novel way of looking at sensory receptor codes but hardly a new notion. Leibniz in his *Monadology* describes the universe with this sort of a construction. His monads are a particularly apt description of olfactory receptors and perhaps the entire nervous system. Only the word, "cell", need be substituted for "monad". (The optical hologram is a simple example of this sort of an integral transform also. Every point on the photographic record of the diffraction pattern contains information about the entire visual field. Each point looks at the visual field from a different viewpoint and with low resolution.)

The olfactory code is more complicated. It is non-linear. Responses to mixtures cannot be predicted from the separate responses to the elements of the mixture. Further, the response is strongly contingent upon the recent history of the activity of the cell. An example of the kinds of responses we obtain from a few cells follows. We feel that there is a particular relevance of this work to the many current efforts to build electronic models of sensory systems. Most of these efforts pointedly ignore

what are, in our view, the crucial principles of organization of biological sensory systems.

Direct evidence for the complexity of representation of sensory phenomena comes from the olfactory system. This is because we can record the signals from the axons of the receptor cells in response to physiological stimuli, and the cells have no interconnections, hence the signal in each cell is independent of the signals in others. We use a fine-tipped metal microelectrode to record activity from the axons which comprise the first nerve of the frog. The frog is pithed in such a way as to insure vigorous circulation and the only other surgical intervention is to remove part of the dorsal surface of the nasal cavity to expose the receptor surface.

The recording electrode external to the axons will often pick up signals from several cells which are sufficiently close to the tip. The activity of different cells can be separated (as long as only a few are near the tip) by observing the amplitudes of the action potentials. The cell most closely coupled to the electrode signals with the largest spike. In the case of the records shown in Figure 1, three spikes of clearly different amplitudes are distinguishable. Simultaneously we pick up the slow potential from the surface of the mucosa which represents activity of a large number of receptors and add this to the signal from the microelectrode. This produces the base line deflections which are indication of the odor stimulus. The spikes from the single cells are passed through an amplitude selector and brightening circuit in order to produce a reasonably clean display. Variations in the amplitudes of each of the three spike groups are due to the ever present noise of the electrodes. The procedure and instrumentation are described in detail elsewhere (Gesteland, *et al.*, 1965).

Traces of an experimental sequence are displayed in Figure 1. Odors were presented in the order shown and each odor puff lasted for about one second except as noted in the caption. Stimuli were never given more frequently than one a minute. The odor intensity was selected to produce a noticeable slow potential, i.e., about one millivolt. This would be called a weak odor by a human but strong enough to allow identification of the substance. Since the slow potential measures the activity of a large number of cells, appearance of a slow potential means that a significant fraction of the receptor cells are affected by the stimulus. Traces of cell activity in between odor puffs also appear in Figure 1 in order to indicate

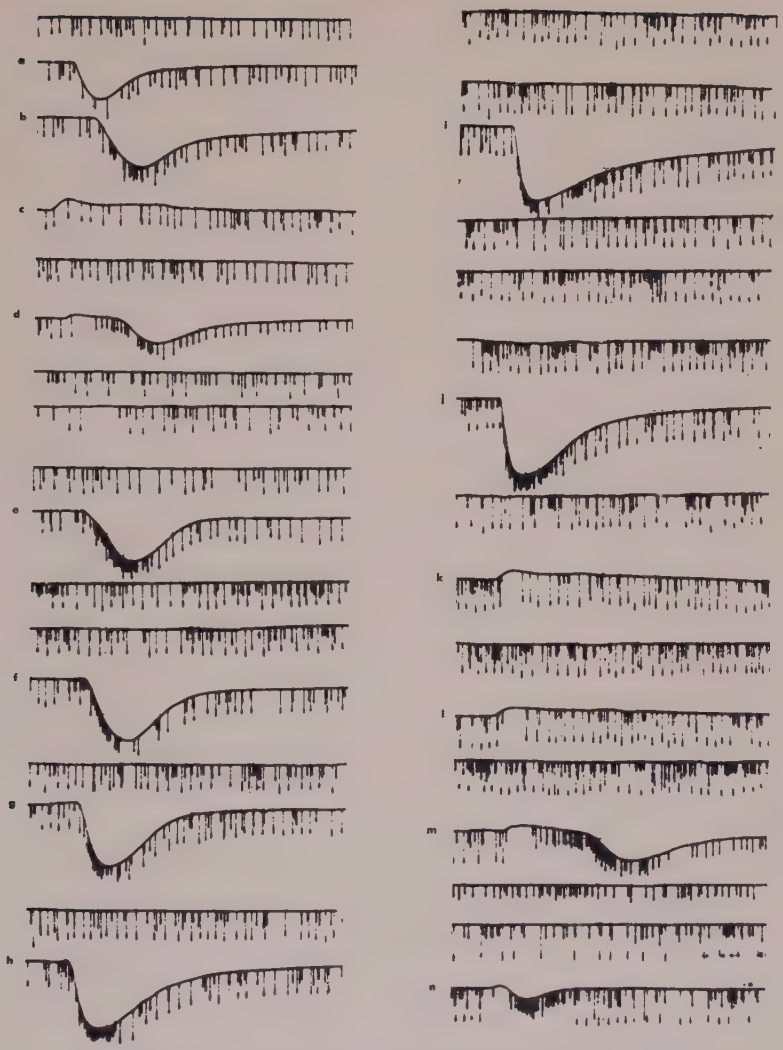


Fig. 1. Responses of single cells in an experimental sequence. Three cells can be distinguished by three different amplitude spikes. Each sweep is 10 seconds long and the base line is deflected to indicate the slow potential recorded with a gross electrode touching the mucosa surface. Lettered traces indicate stimulation with an odorous substance. Others show activity between stimuli. The odors were delivered in the sequence shown with at least one minute between stimuli. Puffs were of less than one second duration except as indicated. a. anisole; b. tetraethyltin; c. methanol; d. pyrrole (2 sec.); e. diethylaminoethanol; f. geraniol; g. limonene; h. menthol; i. camphor; j. menthone; k. methanol; l. methanol; m. pyrrole (4 sec.); and n. pyrrole ($\frac{1}{2}$ sec.).

aftereffects and background activity of the cells. Traces are ten seconds long.

If we look first at the cell whose responses are signalled by the spike with the largest amplitude we find the following behavior. This cell is almost inactive in the absence of a recently delivered odor. It fires once in the first trace and not at all in response to anisole, tetraethyltin, or methanol. It also does not fire at all during and about 20 seconds following a puff of pyrrole. Then begins a prolonged period of irregular activity with an average rate between one and two spikes per second. This now becomes the dominant mode of the cell and remains so for more than a half hour. It may be argued that the cell has been injured by the pyrrole but this seems unlikely since the activity of the other cells underwent no noticeable transition. The firing of this cell is then inhibited by diethylaminoethanol, geraniol, limonene, menthol, camphor, and menthone. There is only a short inhibition caused by methanol, a substance which inhibits firing in a majority of cells. The stimulus is repeated to demonstrate the identity of the response pattern for puffs of the same substance when these puffs are delivered in the same "odor context". There is a great difference between the responses to these two methanol puffs and the response to the puff delivered early in the sequence, before the cell was "turned-on" by pyrrole. The last two stimuli are both pyrrole, the second shorter and weaker than the first. Both are inhibiting, the first with a long aftereffect, the second lasting not much longer than the odor puff.

If we turn our attention to the cell whose response is signalled by the next-to-largest spike amplitude, we find a different ranking of stimuli and different interval patterns within the response. The cell is slightly inhibited by the first odor puff, anisole. It is not much affected by tetraethyltin, slightly inhibited during a puff of methanol and then excited following termination of the puff. It is strongly inhibited by pyrrole, slightly excited by diethylaminoethanol, geraniol, and limonene. Menthol has no effect. Camphor excites it and there are bursts of excitatory response following the puff. Menthone is also excitatory. Methanol is strongly exciting, starting late and lasting long after termination of the stimulus. Again this response repeats accurately in detail when the stimulus puff is repeated. These responses are much like the methanol response early in the series, in contradistinction to the first cell. Responses to pyrrole again demonstrate the importance of the odor context. The early

puff of pyrrole completely inhibited the medium-amplitude spike. The two puffs (long and short) delivered at the end of the series are transiently inhibitory, then strongly excitatory for the remainder of the puff period and for a few seconds following termination of the puff. There appears to be another period of excitation some 20 seconds later but we do not have enough record data to be very sure about the aftereffects in this case.

The cell signalled by the smallest spike also has its own private view of the odor world. Anisole does not affect it at the intensity used. Tetraethyltin excites the cell, methanol inhibits it. Pyrrole causes first, inhibition, then excitation, with the highest spike rate occurring near the peak of the slow potential following termination of the stimulus. Thence follow bursts of activity separated by quiet periods. The response to the two-second long odor puff lasts for minutes. Diethylaminoethanol, geraniol, limonene, menthol, camphor, and menthone all excite the cell. Methanol inhibits the cell during the puff. Turn-off of the odor causes the cell to fire at a rate considerably higher than its resting rate for at least 20 seconds. A second puff of methanol produces the same effect and this is not so different from the methanol response earlier in the series. Pyrrole also produces about the same response late in the series as it did earlier. The pattern preserved independent of odor duration.

We can summarize the results of many such experiments with the following statements (Lettvin and Gesteland, 1965):

1. Every fiber has an irregular base rate at which it fires in the absence of any introduced odor. The instantaneous rate, while varying over a fair range, tends to cluster around a rather low average of at best 1/sec., but usually much less. The fact that we have introduced no odor does not mean that the rate may not be governed by compounds emitted by the animal itself.
2. This noisy instantaneous base rate of firing can be increased (the fiber is exalted) by many compounds, decreased (the fiber is depressed) by many others, and slightly, if at all, affected by still many others. Exaltations and depressions may form a definite sequence when the receptor is exposed to a single odor, so that to characterize properly the response of a fiber we must talk of the sequence.
3. If, to simplify matters, we treat only the initial response as our significant measure, and then we can arrange all odors along a single

axis with respect to any single fiber, from those that exalt it most to those that depress it most.

4. If we have used, say, ten odors in studying a group of fibers seriatim, and for any nine such odors two fibers show the same ordering, it is unlikely that the tenth odor will have the same order position between the two fibers. Put in another way, we are saying that, given any set of stimuli that are ordered the same way by two fibers, it is easy to find an additional stimulus that discriminates the ordering done by the two fibers.
5. These ordering principles for any cell will only apply if the separate stimuli are delivered very far apart in time, since some cells will be affected for a long period following even weak stimulation with certain substances.
6. If we have the response of a fiber to one odor, and the response to another odor, then, whether the responses are different or the same, we cannot predict the response to a mixture of these odors, neither in

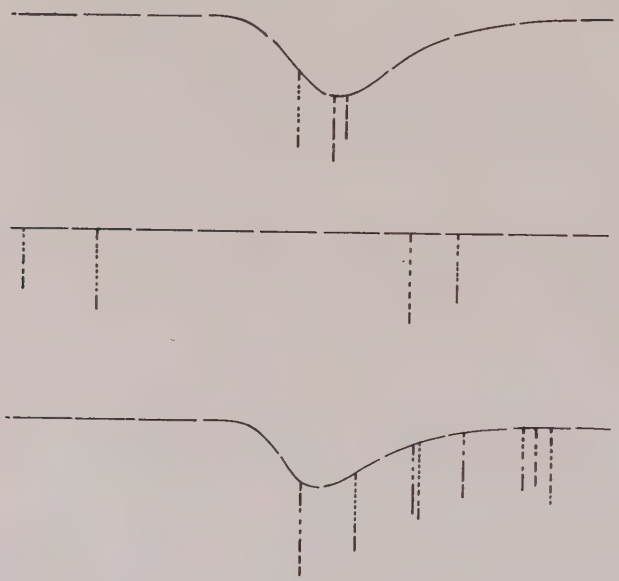


Fig. 2. Response of a single cell to *n*-butanol (top trace), musk xylene (center trace), and to both delivered simultaneously (lower trace).

magnitude nor direction. That is to say, we cannot tell that a fiber will be exalted by a combination of two odors that separately exalt it, or indifferent to a mixture of two odors that separately affect it little if at all. This is illustrated in Figure 2. Here the top trace shows a cell responding weakly to a puff of n-butanol. During the course of the second trace, a strong puff of musk xylene was delivered, producing no response. When the two stimuli are delivered simultaneously, several spikes in rapid succession are evoked.

These qualities that we have found in the olfactory fibers of the frog mirror those found in the "generalist" receptors of the bee by Schneider's group, and those in caterpillars by Dethier's group. When it is possible to say of the same sensory system that the same principles of encoding hold between so widely disparate animal types as bees, moths and frogs, we think it likely that these principles are quite general, and that the results reflect not so much a common incompetence of experimenters who fail to find specific receptors, but rather an unexpected yet legitimate mode of odor representation. The world of odor appears to the brain behind the nose as it is given in the fibers that communicate from nose to brain. There is no other pathway. If the information is given in terms as we just described, that is the informational system we must handle, and not some other that is confected out of odor theories and chosen because the structure is easily represented analytically.

What characterizes the response of an olfactory fiber is that it has a conspectus over the set of all possible odors with respect to which it utters a point of view that is not a measure in any ordinary sense. The most poignant case is when we see that the response to a mixture cannot be predicted well from the response to the components of the mixture given separately. Information so presented by a kind of integral transform is, in turn, unlikely to be handled by simple correlational methods that would only tend to smear, in this case, the resolution between odors. Unless one means by correlation what a judge does on weighing the testimonies of several witnesses, as opposed to what an instrument, marked "CORRELATOR", does on receiving sequences of numbers, we feel that correlation, as well as averaging, can be ruled out except in the first and most general sense. We have no idea how subsequent neurons higher in the nervous system are connected to accomplish this sort of correlation. Psychophysically we know that notions of groups occur and that a rose smells like a rose regardless of context. We should like

to suggest that congenital partial anosmias reflect defects in the second order fibers rather than in the primary receptors, purely by virtue of the sort of coding involved. Some people cannot smell cyanides, others cannot smell butyric acid, etc.; the anosmias being more "sharply tuned" than are receptor cells.

What we have found in the fibers of the olfactory nerve is, in a way, paradigmatic of what we find in most nervous elements. The form stands out most clearly when we handle a system like smell for which we do not have a preconceived mechanism in mind. If we address ourselves to a well-known system we find similarly complex codes in spite of our prejudice against such findings. We take the liberty here of pre-announcing without any supporting data some of the recent findings made by Humberto Maturana, Samy Frenk and two of us (S-H.C. and J.Y.L.) on the optic nerve of frog. From the contingencies crudely described in the original papers on the frog's eye, we were apparently unable to convince most readers of the difficulties in describing the action of even the simplest element in the optic nerve. (Lettvin, *et al.*, 1959.) It is clear that any second- or third-order neuron receives from many other neurons some inputs that inhibit and some that excite. Because of the asymmetrical nature of inhibition and excitation, one cannot sum inhibitors and excitors to explain the firing of the neuron. We may sketch the argument briefly thus: A subsynaptic excitatory patch acts, when it is activated, as a current activator locally introduced (i.e., Na^+-activation). It has an effect on the axon hillock (where firing originates) that depends upon its electrotonic distance from that hillock, for the nerve membrane over the cell and dendrites must be conceived as a nonlinear transmission line that, for small signals such as these, acts almost linearly. But, as Kuffler and Eyzaquirre showed (1955), an inhibitory subsynaptic patch when triggered does not act as a counter-current generator but rather as a shunt (K^+- or Cl^--activation) locally produced across the membrane so as to change the electronic characteristics of the membrane. Thus, the effects of combined excitation and inhibition in the dendrites as exerted upon the axon hillock depend as much upon the placement of excitors and inhibitors with respect to each other as on how many of them are active. One event, the excitatory one, where it occurs, is current generative into a nonlinear transmission line; the other, the inhibitory event, where it occurs, changes the distributed resistance in the line. From these facts alone and from knowledge of the complex anatomy of dendritic trees,

it ought to be possible to see, on a-prioristic grounds, that the transformation from input to output of a neuron is a difficult thing to conceive. We wanted to see what could be said at a minimum.

What we have found is this: any "dimming detector" axon fires at rates such that the intervals between pulses vary from about two milliseconds up to 2000 seconds. (This is significantly different from the rather narrow bandwidth of primary olfactory fibers in which the minimum pulse interval is approximately 100 milliseconds.) Within this dynamic range the firing fluctuates not only as a function of changes of light intensity and absolute light intensity upon the retina, but also according to the previous history of light. We presently have tracked reliably the effect of a bright flash for about two hours. A moving average of the pulse interval displays the measure of some of these parameters. But other functions, notably what we call "envelopes" on a continuous plot of pulse-interval against time, seem to measure not only different combinations of the same parameters but other variables not seen in the averaging. The facts that two or more different operations on the same time series of pulses yield different combinations of information and that some of the operations exclude some of the information while retaining the rest, suggested to us that the fiber was doing a kind of time-sharing multiplexing of the various kinds of information coming to it. We think that the meaning of a pulse interval is not negligible and that its signifficance with respect to the variables being transformed depends on context. Possibly the only analogy we can give here is what would happen if we were to take a similar display of intervals between baseline crossings in uttered speech plotted against time. Here the resulting dot figure would show different preferred interval regions at different times, and these would represent sum- and difference-frequencies, or formant modulations by pitch, as kinds of dotted lines. Such a line we call an envelope. Similar lines, clusterings, tendencies, occur in the firing pattern of the dimming detector, and they change in different ways between themselves according to the different kinds and different sequences of lighting. We are now able to track some of the parameters. It is as if one had a nonlinear oscillator with a kind of distributed control over all the coefficients of the higher order terms that are involved in describing the action of the element. It may be argued that our ability to recover information from a nerve fiber is no guarantee that the information is used by subsequent neurons. Our only reply is that, given the nature of

nervous connectivity, how does one decide what of the information to exclude? Indeed, there is a kind of impiety in holding that information is preserved up to a certain point and then ignored. For to what end would the element be so devised as to waste itself on resolvable ambiguity, frittering away its time with meaningless utterances?

Thus the neuron has a complex point of view and the categories in which it deals are not those that obviously fit the usual simple, time-invariant perceptual models based upon sensory experiences and mathematical conveniences.

ACKNOWLEDGMENT

This work was supported in part by the Bell Telephone Laboratories, Inc. and by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Contract AF 33(615)-3885. Dr. Gesteland is supported in part through the National Institutes of Health, Contract Fr-00018-03. Further reproduction is authorized to satisfy needs of the U.S. Government.

REFERENCES

Gesteland, R. C., Lettvin, J. Y., and Pitts, W. H. (1965). Chemical transmission in the nose of the frog. *J. Physiol.*, **181**, 525.

Kuffler, S. W., and Eyzaguirre, C. (1955). Synaptic inhibition in an isolated nerve cell. *J. Gen. Physiol.* **39**, 155.

Lettvin, J. Y., and Gesteland, R. C. (1965). Speculations on smell. *Cold Spring Harbor Symposia on Quantitative Biology*, **30**, 217.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proc. Inst. Radio Engineers* **47**, 1940.

V. BRAITENBERG

University of Naples,

Naples, Italy

On the Neural Optics Behind the Eye of the Fly

The symmetry and regularity of the arrangement of ommatidia in the insect compound eye (Fig. 1) leads us to expect an equally regular structure of the nervous apparatus in which the sensory inflow from the eye is further elaborated. The analysis of the texture of these organs and its interpretation in terms of functional schemes should therefore be facilitated, but it is not only the convenience of a clear periodicity which makes the visual ganglia of insects appealing to the histologist. In the well-

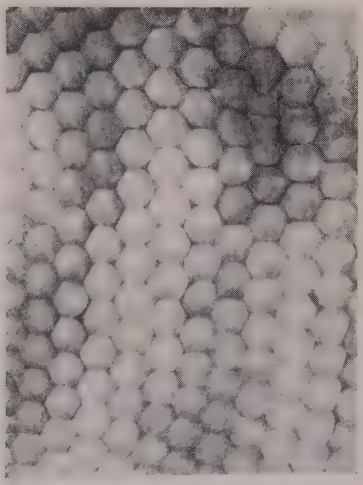


Fig. 1. Microphotograph of a piece of detached cornea of the fly to show the hexagonal arrangement of the lenses.

known experiments of Hassenstein, Reichardt and others some precise quantifiable motor reactions of insects have been seen to correspond to some equally well defined parameters of the visual input in a manner which was set down as a set of rules by Hassenstein and transformed into a mathematical model by Hassenstein and Reichardt in 1956, Reichardt 1957, Reichardt and Varju 1959. The resulting scheme is one of the very

rare instances of a model which is stated in terms of elements that should be readily identifiable with the elements (fibres, synapses) of the histology. Instead of relying on the arduous procedure of translating the histology into a functional "wiring scheme", we are here in the lucky situation in which me may look for a particular fibre pattern already predicted on the basis of the behavioural experiment. I shall not attempt to summarize the results of Reichardt and others but will only illustrate (Fig. 2) the pattern of cross connections between ommatidia which is to be ex-

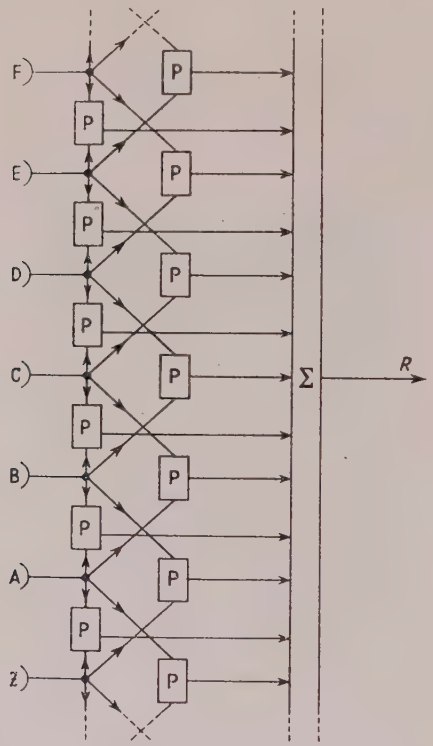


Fig. 2. (From Reichardt, 1962) Diagram of the functional relations between neighbouring ommatidia as required by Reichardt's model of the optomotor reaction in the beetle Chlorophanus. A, B, C... indicate different ommatidia in a row across the insect's eye. The boxes labeled *P* contain the postulated mechanism for the computation of velocity of movement according to the principle of cross-correlation of the inflow from the two ommatidia. Such boxes are postulated between any two neighbouring ommatidia and between any two neighbours but one. Σ represents summation of the outputs of all these boxes. *R* is the response.

pected on the basis of their experiments at some level in the ganglionic chain.

A few words about the gross lay-out of the visual apparatus of the fly are appropriate (Fig. 3). At the surface a layer of rather neat plane-convex

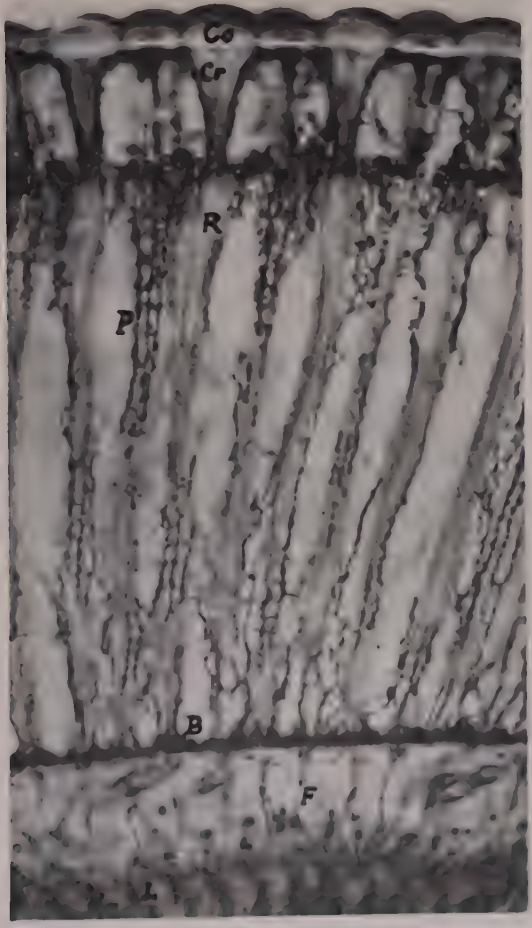


Fig. 3. Cross-section through the retina of the fly. *Co*: cornea with lense-shaped thickenings, *Cr*: crystalline cone. *R*: retinular cells. *P*: pigment cells. *B*: basilar membrane. *F*: fibres running from the base of the retinular cells to the first ganglion. *L*: lamina ganglionaris.

lenses followed by a layer of transparent cones, each with its base attached to the plane inner side of the corresponding lens, are without difficulty interpretable as the dioptric apparatus. Each cone goes over into an elongated structure which reaches down to the so-called basilar membrane, and which in cross-sections appears clearly to be composed of seven parallel columns. These are the seven retinular cells (an eighth retinular cell is added at the base of each ommatidium) for whose fine structure we refer to the electron-micrographs of Fernandez-Moran (1958) and which are beyond any doubt to be considered as the site of the photochemical processes leading eventually to nervous excitation. Thus through the layer in which the laws of geometrical optics are ruling and the layer governed by photochemistry, we can follow the transmission of optical messages into a layer of fibres arising at the base of the ommatidia, in which we may confidently expect the transmission to take place in the form of events (depolarizations, spikes(?)) familiar from neuro-physiology.

These fibres eventually lead into the first ganglion (lamina ganglionaris) where they will make intimate synaptic contacts, very impressive on electron-micrographs (Trujillo-Cenoz, 1965) with a set of other fibres leading to the next ganglion.

It is the first layer of fibres, between the ommatidia and the lamina (Fig. 4), which one would be tempted to identify with the set of functional relations between ommatidia postulated in Reichardt's model. In fact Figure 4a is well compatible with the number seven of fibres (the eight fibre is much thinner and may be disregarded here) arising from each ommatidium, since it may be interpreted as an indication of one fibre running straight down to the corresponding "neuroommatidium" of the lamina and six reaching the six neighbouring neuroommatidia. As can well be seen, the lamina preserves the periodicity of the retina and the "neuroommatidia" or "cartridges" of the lamina at least in first approximation are arranged in the same hexagonal packing order which characterizes the layer of ommatidia. Such a scheme would satisfy the requirements of Reichardt's model, based on Hassenstein's experiments which revealed interactions between neighbouring ommatidia and between neighbours but one, but no interactions between ommatidia farther removed. However, Figure 4b, obtained from a section cut obliquely through the eye, illustrates the disturbing observation that each ommatidium produces a fiber which runs downward-slightly backward

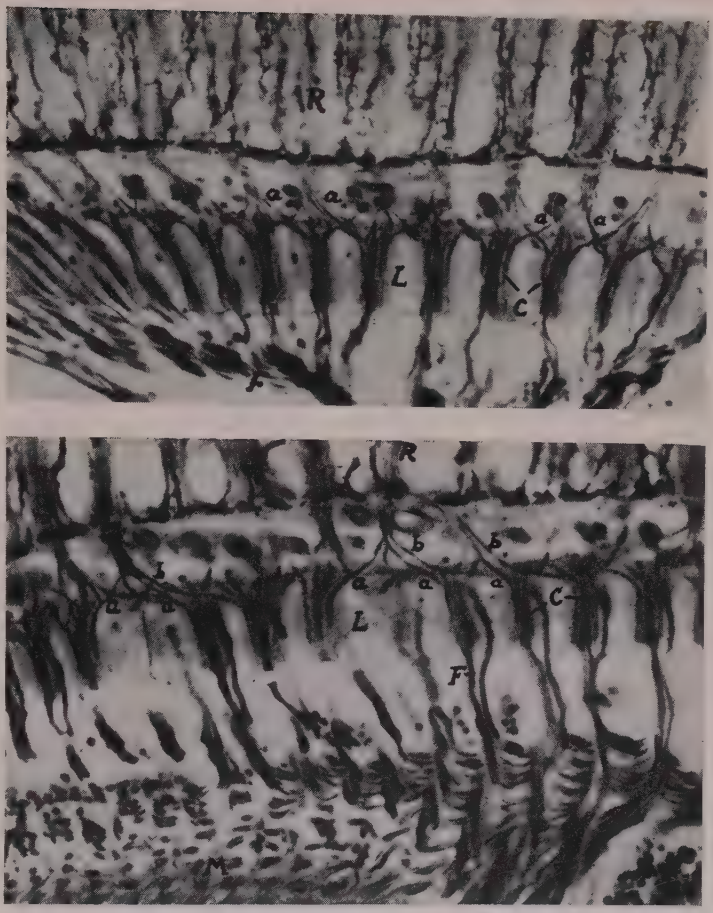


Fig. 4a and b. Sections through the lower part of the retina and the first ganglion, Lamina ganglionaris, of the fly. Silver staining method according to Holmes. Both cuts are parallel to the axes of the ommatidia, but while Fig. 4a is from a sagittal cut, i.e. one along a vertical row of ommatidia, Fig. 4b is from a horizontal cut, along (not precisely) a horizontal row of ommatidia. *R*: retina. *L*: lamina ganglionaris. *C*: cartridges or neuroommatidia of the lamina. *F*: fibres running from the lamina ganglionaris to the second ganglion, the medulia (*M*). a: fibres from an ommatidium to a neighbouring cartridge. b: fibres running to more distant cartridges which are seen only in the horizontal cut, Fig. 4b.

beyond the immediately neighbouring ommatidium. The careful analysis of the pattern of these connections in tangential cuts (Fig. 5) reveals the

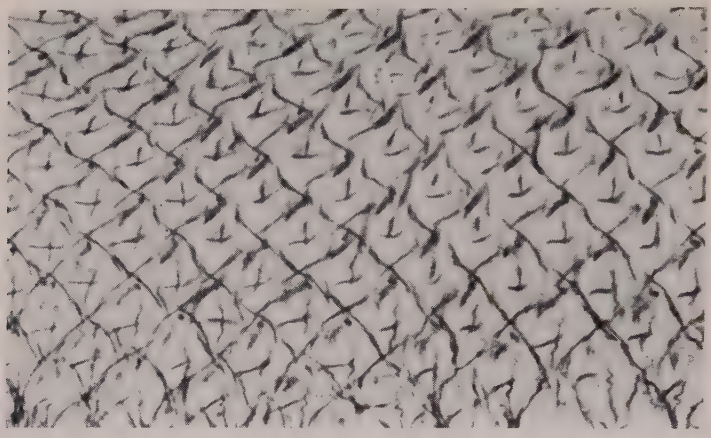


Fig. 5. Tangential cut (perpendicular to the axes of the ommatidia) through the layer of fibres between the retina and the lamina, to show the regular arrangement of the fibres. From a depigmented and silver stained preparation, kindly lent to me by Dr. G. F. Meyer, Max-Planck Inst. f. Biologie, Tübingen.

scheme of Figure 6a which upon stretching in the horizontal direction yields the pattern of Figure 6b. The reason for which this stretching has to be applied resides in the peculiar distortion of the pattern of the lamina with respect to the retina in the anterior region of the eye.

At this point, if one would still maintain that these fibers are identifiable with the lines of Reichardt's diagram, one would have to embark on a series of speculations about the meaning of such asymmetrical wiring in the context of the optomotor experiments which have up to now failed to reveal a similar functional asymmetry. We shall not propose such speculations, since an explanation of this pattern can now be given on the basis of an entirely different set of observations based on the optical properties of the first retinal layers.

If a thin superficial fresh cut of the fly's eye is viewed through a microscope from the inner side, following a technique first proposed by Autrum and Wiedemann (1962), the action of the dioptric apparatus can be observed directly. With diffuse illumination (Fig. 7a) in each ommatidium we observe seven bright spots arranged in a pattern which the

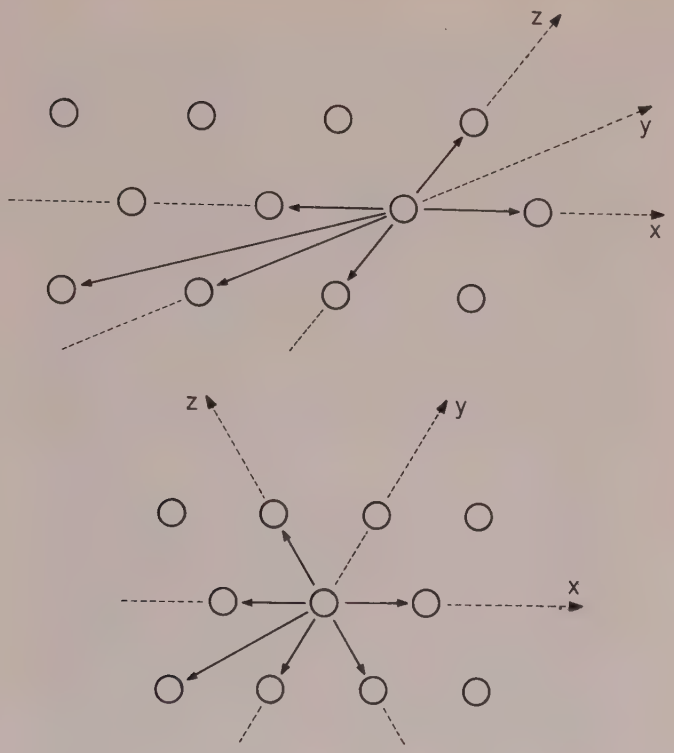


Fig. 6. a (top): Distribution of fibers from one ommatidium onto six cartridges of the lamina. Two more fibers, running straight down, are not shown. b (bottom): Correcting for the compression of the lamina with respect to the retina in the antero-posterior direction, the distribution of fibers matches the layout of retinular cells shown in Figure 7a (except for a 180° rotation).

reader may easily identify as the same, but rotated 180°, as that of Figure 6a. These spots are nothing but the tips of the seven retinular cells of one ommatidium reaching the focal plane of the dioptric system and acting as light pipes. Using a small light, in the same set of ommatidia only a few of these points will light up (Fig. 7b). With different positions of the lamp, it can be seen that different spots of the pattern will be illuminated in different ommatidia by one and the same position of the light. The explanation of the fibre distribution of Figure 6a is given in the

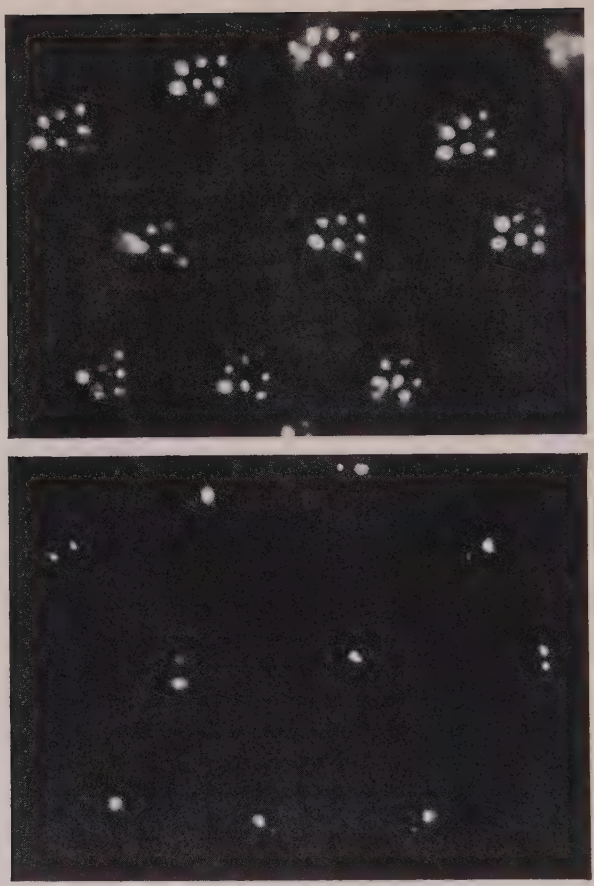


Fig. 7. Microphotographs obtained by the method proposed by Autrum and Wiedemann, 1962. A superficial section of the eye is detached from a live fly by a tangential cut just below the tips of the crystalline cones, i.e. through the uppermost level of the retinular cells. One looks through the dioptric apparatus of the fly from behind. In Figure 7a the preparation is illuminated by diffuse light. In each of a group of ommatidia shown, seven bright dots, corresponding to seven retinular cells light up (with exceptions due to distortions caused by the section). In Figure 7b a point-like source of light is used, which causes only one or at most two retinular cells to light up brightly in each ommatidium. Close inspection reveals that non-homologous retinular cells are illuminated in different ommatidia.

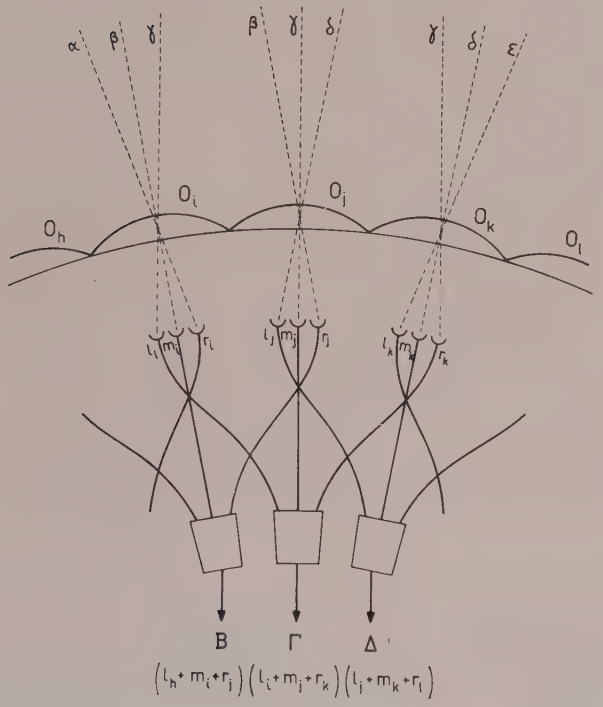


Fig. 8. Diagram to show how the fibre distribution of Figure 6 corrects for the divergent optical axes of the retinular cells of one ommatidium. O_h, O_i, O_j, ... represent a row of ommatidia of the compound eye of the fly. l_i, m_i, r_i are the left, middle and right retinular cell respectively of ommatidium O_i, similarly for O_j etc. The lines of sight of these retinular cells are denoted by α, β, γ, ... If the angle formed by the axes of two neighbouring ommatidia is the same as the angle between the lines of sight of two neighbouring retinular cells, there will be sets of (in our drawing three, in reality seven) non-homologous retinular cells of different ommatidia which look in the same direction (α, β, γ, ...). The twist of the fibre bundles proximal to the retinular cells, actually observed in our preparation, and the distribution of fibres shown in Figure 6 will have the effect of collecting in each cartridge of the lamina (B, Γ, Δ etc.) all the visual input derived from one direction (β, γ, δ etc.). This makes it very likely that the computation of perceived velocity does not occur at the level of the lamina.

diagram of Figure 8. Light reaching the fly's eye at a certain angle will maximally stimulate different retinular cells of a set of seven ommatidia (only three in the two-dimensional diagram of Fig. 8) whose optical axes all point in the same direction. If at the next level an orderly representation of the angular position of outside objects is again desirable, we may twist the fibres arising from each ommatidium around 180°, to compensate for the optical inversion of the image due to the lens, and distribute the fibres according to the same scheme which characterizes the arrangement of the sensory elements in the retinula behind the lens. If the angles between two neighbouring ommatidia match the angles between the optical axes of two neighbouring sense cells of one ommatidium, the neuroommaditia of the lamina are nothing but an orderly representation of points in the (two-dimensional) environment. Measurements by Kirschfeld (1966, private communication) and Kirschfeld (1965) indicate that these angles are the same to an astonishing degree of precision. Moreover, going back to our preparations (Fig. 4b), we will be gratified to discover that the fibre bundles between the retina and the lamina actually reveal the 180° twist which our scheme requires.

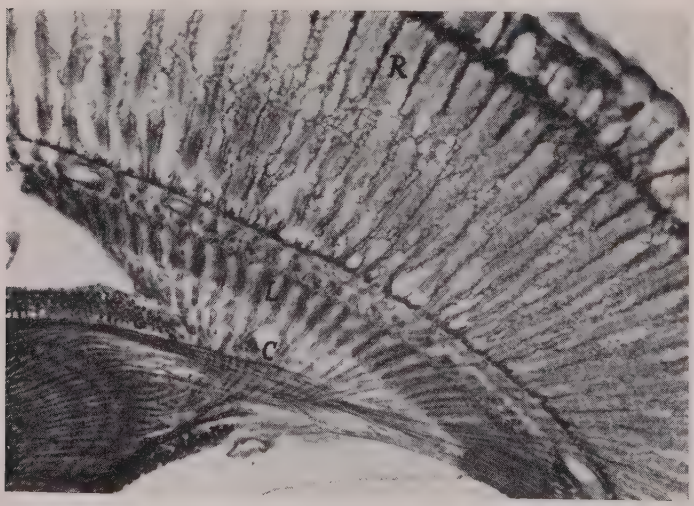


Fig. 9. Horizontal section through retina (*R*), lamina (*L*) and medulla (*M*) to show the chiasm, *C*, which transposes the medial and lateral halves of the visual input from each eye.

This in turn would indicate that the lamina cannot be the site of the computation postulated by Reichardt's model. In fact, although fibres from different ommatidia are brought together in the lamina, according to the scheme developed here information reaching the eye from different angular directions does *not* converge onto a single synaptic junction at this level.

The next candidate for the mechanism underlying the optomotor reactions is the ganglion following the lamina, the medulla. On it a set of fibres leaving the lamina, and therefore the image of the outside world is again projected in an orderly fashion but for one impressive disruption, due to the chiasm which transposes the medial and lateral parts of the visual field (Fig. 9).* This chiasm represents an extremely widespread feature of arthropod nervous systems and is followed in many species, e.g. crabs,

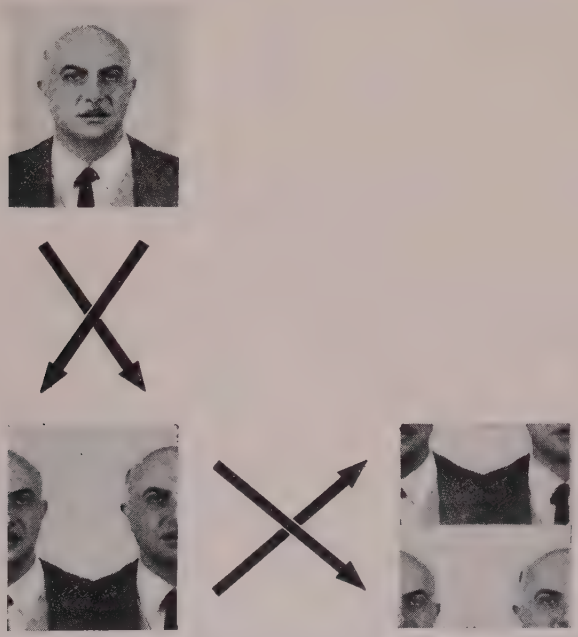


Fig. 10. Schematic illustration of the disrupture of an image due to the two consecutive chiasms, rotated 90° one with respect to the other, which are found in the optic system of many arthropods. The total effect is that of an inside-out transformation of the picture.

* It has since been shown that the disrupture of the image is slightly more complicated than described here, as will be illustrated in a forthcoming publication.

by a similar chiasm between the second and third optic ganglion, only rotated 90° with respect to the first. What this does in terms of neural optics is illustrated by Figure 10. The biological significance of this disrupture of the image is a matter for future research. It is tempting to make some cybernetical speculations for instance in relation to a mechanism made to detect the enlargement of a visual pattern, which has been shown to exist in the fly, an animal in which a characteristic "landing" reaction has as its specific stimulus divergent movement in the visual field, produced normally by an approaching landing surface, and experimentally e.g. by a spiral painted on a rotating disk facing the animal (Braitenberg and Taddei, 1965), It may be easier and perhaps more economical in terms of fibres to detect convergence of excitation toward one point of a planar ganglion, rather than divergence. Since very plausible, the expansion of an image, indicating an approaching object, is a much more important event in the life of a fly than the contraction of an image, the chiasmatic arrangement which transforms expansion into contraction at the ganglionic level may be a useful mechanism.

ACKNOWLEDGEMENT

This research was sponsored in part by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Grant AF EOAR 66-25 through the European Office of Aerospace Research (OAR), U.S. Air Force. Further reproduction is authorized to satisfy needs of the U.S. Government.

I wish to express my gratitude to Prof. Dr. W. Reichardt, Max-Planck-Inst. f. Biologie, Tübingen, for the opportunity to carry out some of this research in the stimulating environment of his department.

REFERENCES

Autrum, H., and Wiedemann, I. Versuche über den Strahlengang im Insektenauge (Appositionsauge). *Z. für Naturforschg.*, **17**b, 480–482 (1962).

Braitenberg, V., and Taddei Ferretti, C. Landing reaction of musca domestica induced by visual stimuli. *Naturwissenschaften*, **53**, 155 (1966).

Fernandez-Moran, H. Fine structure of the light receptors in the compound eyes of insects. *Exp. Cell Res.*, **5**, 586–644 (1958).

Hassenstein, B., and Reichardt, W. Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. *Z. für Naturforschg.*, **11**b, 513–524 (1956).

Kirschfeld, K. Das anatomische und das physiologische Sehfeld der Ommatidien im Komplexauge von Musca. *Kybernetik*, **2**, 249–257 (1965).

Kirschfeld, K. Private communication (1966).

Reichardt, W. Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems. *Z. für Naturforschg.*, **12**b, 448–457 (1957).

Reichardt, W., and Varju, D. Übertragungseigenschaften im Auswertesystem für das Bewegungssehen. *Z. für Naturforschg.*, **14**b, 674–689 (1959).

Trujillo-Cenoz, O. Some aspects of the structural organization of the intermediate retina of dipterans. *J. Ultrastruct. Res.*, **13**, 1–33 (1965).

ALLEN SANDBERG

and LAWRENCE STARK

Presbyterian St. Luke's Hospital,

and College of Engineering. University of Illinois at Chicago Circle,

Chicago, Illinois.

Functional Analysis of Pupil Nonlinearities

INTRODUCTION

In attempting to describe mathematically or to characterize the "black-box" input-output behavior of the human pupil light reflex, we investigated three different basic formulations of the system. The Wiener G-functionals, a recently developed mathematical characterization of systems, will be discussed in detail but we will first briefly present two other, more standard, approaches to the problem, describing functions and heuristic modelling. Neglecting the experimental details—some of which will be presented later— let us assume that the experimenter has

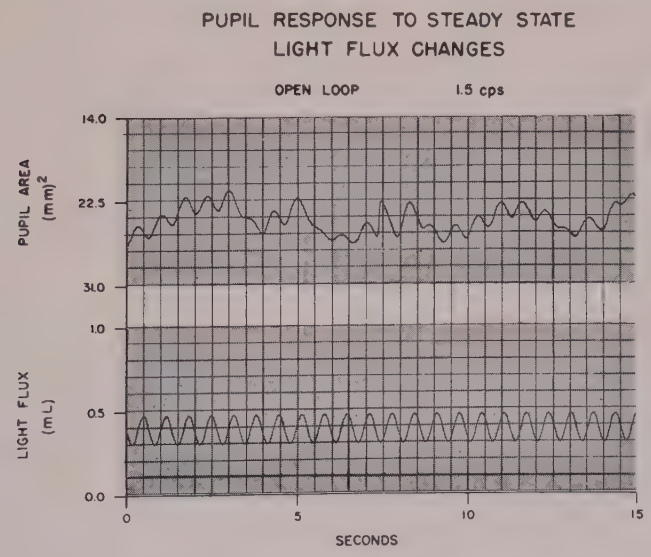


Fig. 1. Pupil response to steady state light flux changes.

available transducers capable of generating input stimuli as waveforms of arbitrary time dependence and monitoring an analog voltage proportional to the output, pupil area of a normal awake human subject.

The linear transfer function description attempts to represent the input-output behavior by a linear system. Stark and Sherman[1] first used this method and were able to predict the frequency of high loop-gain, instability oscillations.[2] These studies have been extended in a quantitative manner and are in part presented here. Figure 1 shows typical input-output records of the human pupil response to sinusoidally varying light; gain and phase *vs* frequency curves for this class of experiments are seen in Figure 2. The asymptotic slope of the gain *vs* frequency curves indicates a system of approximately third order. One is immediately suspicious of any linear model since the gain curves change appreciably with stimulus input amplitude. Note, however, that the phase *vs* frequency curves are almost identical for different stimulus amplitudes. It is this

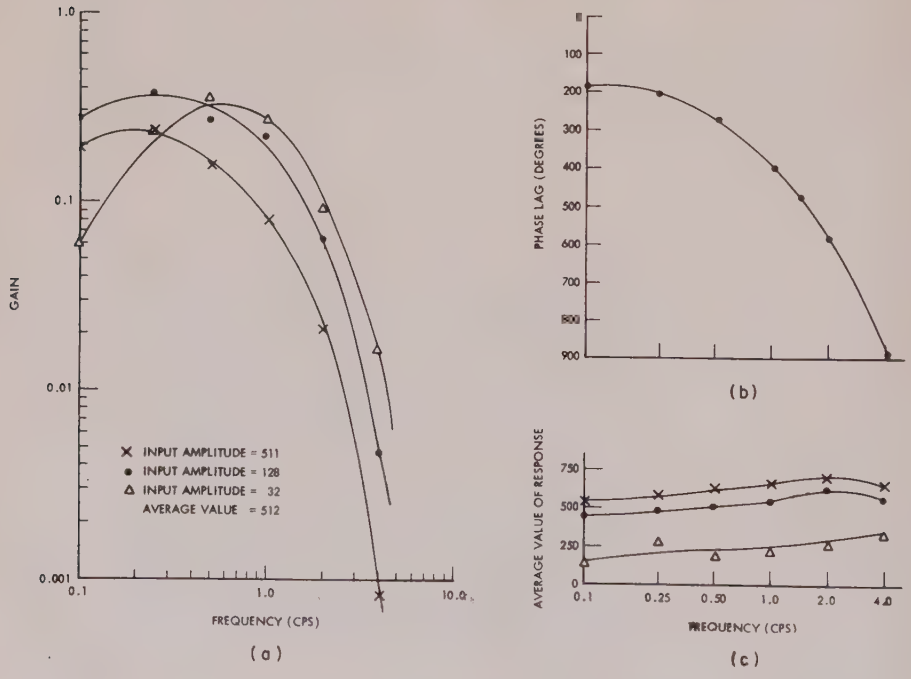


Fig. 2. Sinusoidal steady state response of pupil system to light as a function of frequency with stimulus amplitude as a parameter.

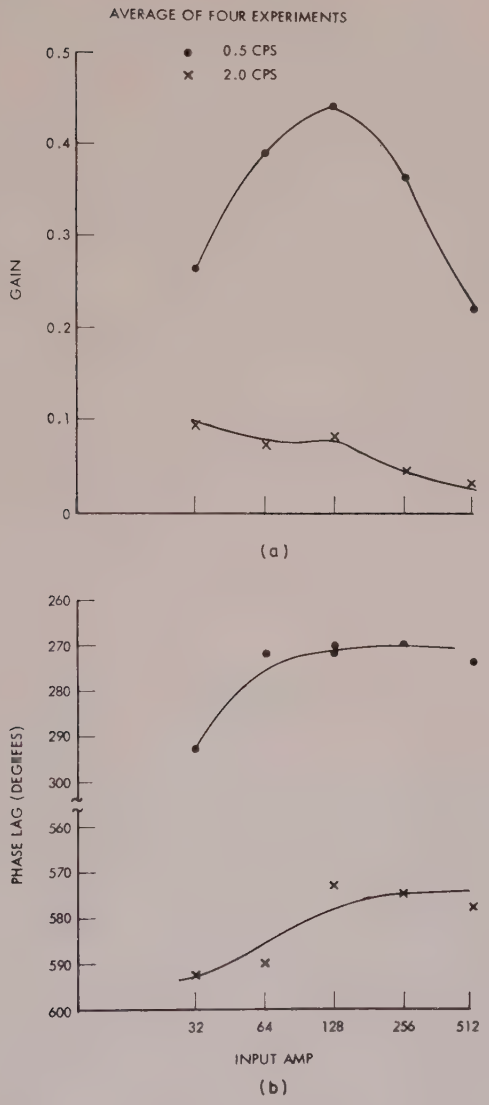


Fig. 3. Sinusoidal steady state response of pupil system to light as a function of stimulus amplitude with frequency as a parameter.

very strong behavioral constraint in the pupil light reflex system which enabled Stark, Cornsweet and Baker[2, 9] to use a linear model to predict correctly the closed loop oscillation frequency in their studies of the pupil.

In order to present more clearly the pupil nonlinearities, we have plotted in Figure 3 the gain and phase *vs* input amplitude with frequency of stimulation as a parameter. Since the gain *vs* amplitude curves at the two stimulation frequencies are appreciably different shapes, the pupil system cannot be represented simply as a no-memory nonlinearity followed by a linear system. Therefore it is impossible to describe the system in simple describing function theory. It can, however, be shown that the previous sinusoidal response data is consistent with a model comprised of some combination of cascaded linear systems separated by no-memory nonlinearities.[7]

An extremely popular and often useful modeling technique is to perform a series of experiments usually with transient stimuli. One acquires experience with the system which coupled with other information one might think pertinent—including preconceived notions—and thereby formulates a time invariant, nonlinear heuristic model. The model's primary purposes are to describe the input-output behavior and approximate, at least in some sense, the topology of the signal flow in the real physical system.

In the case of the pupil light reflex, there are several easily observed nonlinear effects. Typical pulse responses are seen in Figure 4. In addition to the basic wave shape, one immediately observes that a substantial amplitude compression exists fairly early in the signal processing system since the amplitude of response varies over about a factor of three while the input signal changes over a factor of thirty-two and the response shapes are extremely close when all the responses are normalized to the same peak-to-peak height.

Figure 5 illustrates another important nonlinearity-asymmetry. This nonlinearity seems to show its effect especially when the stimuli approach the small signal region.[8] Because of this some investigators have raised serious objections to the use of methods which attempt to linearize the pupil system in order to carry out a quantitative analysis.[7, 10] The model illustrated in Figure 6 represents one attempt to form a heuristic model of the pupil light reflex. In such model making one must choose between making the model more complex to incorporate some experimental data or settling for a simpler model, incomplete but more understandable.

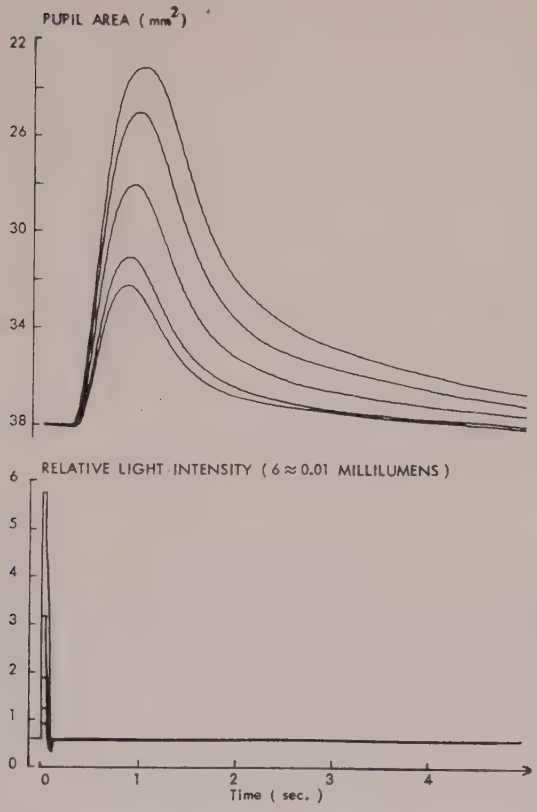


Fig. 4. Human pupil responses to short pulses of light.

We will not stress this heuristic method of characterization, but instead proceed to the Wiener method, which lacks close connection with possible physiological processes, but is a general mathematical description capable of representing an extremely wide class of systems.

This characterization, the main subject of the present paper, is essentially a Volterra functional expansion of the pupil system with kernels $h_0(\,)$, $h_1(\,)$, $h_2(\,)$, etc.* Output and input are related as follows:

$$y(t) = \int h_1(\tau)\, x(t-\tau)\, d\tau + \iint h_2(\tau_1, \tau_2)\, x(t-\tau_1)\, x(t-\tau_2)\, d\tau_1 d\tau_2 + \ldots \tag{1}$$

* $h_0(\,)$ is the zeroth order kernel and represents the constant dc level with zero input and is ignored in the remainder of our paper.

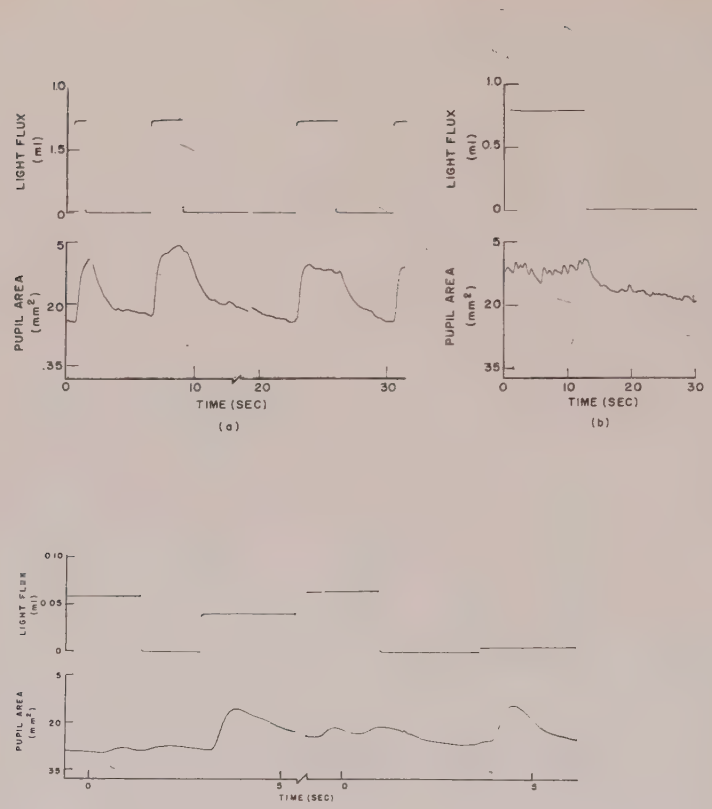


Fig. 5. Pupil light reflex showing asymmetrical nonlinearity effect.

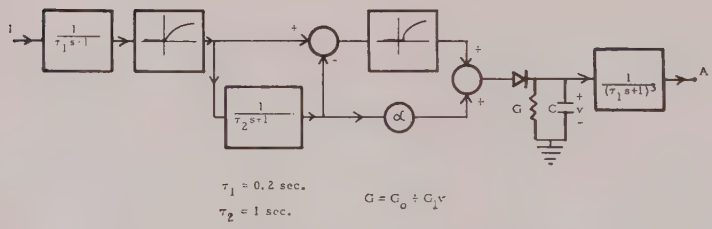


Fig. 6. Heuristic model of human pupil light reflex.

The computationally difficult task in this characterization is the determination of the kernels $h_i(\tau_1, \tau_2, \ldots, \tau_i)$, $i = 1, \ldots, n$. An even more profound question which arises in this method is whether any physical significance or interpretation can be given the kernels—even if they could be found—or are we to pay the price for mathematical elegance by an inability to interpret our results? Do the kernels have any other identifiable significance in addition to their being the magic kernels that simply crank out the correct output when introduced into the Volterra expansion? First, let's describe two analytic procedures for obtaining the kernels and then present the experimental results pertinent to each theoretical method.

METHODS

1. Measurements by use of random signals

A linear system, as shown in Figure 7, is represented in the time domain by its impulse response function $h(\tau)$. The output $y(t)$ is related to the input $x(t)$ by the superposition or convolution integral.

$$y(t) = \int h(\tau)\, x(t - \tau)\, d\tau \tag{2}$$

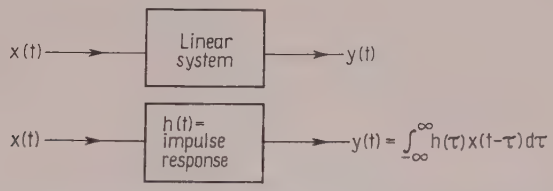


Fig. 7. Representation of linear system by its impulse response $h(t)$.

One method of experimentally computing $h(\tau)$ is to excite the linear system with white noise, that is

$$\varphi_{xx}(\tau) = Au_0(\tau)$$

where u_0 is the Dirac delta function, and to crosscorrelate the output and input.[5] It can be seen that

$$h(\tau) = (1/A)\, \varphi_{yx}(\tau). \tag{3}$$

The representation of a general nonlinear system is shown in Figure 8. If the system under study can be suitably approximated by only one of the Volterra terms of the infinite expansion—in general—then a simple technique exists for determining the kernel of that term.[4] Figure 9

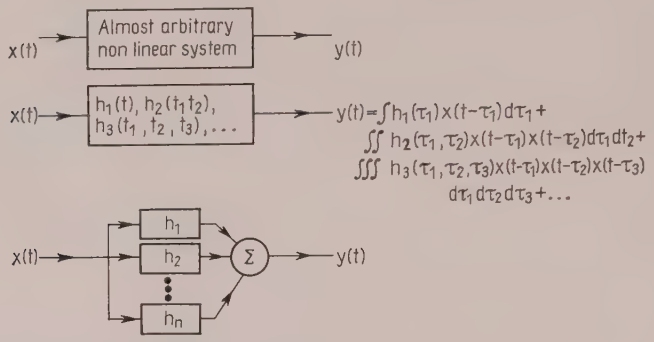


Fig. 8. Volterra representation of nonlinear system. $h_1(\tau_1)$, $h_2(\tau_1, \tau_2)$, etc. are known as Volterra kernels of system.

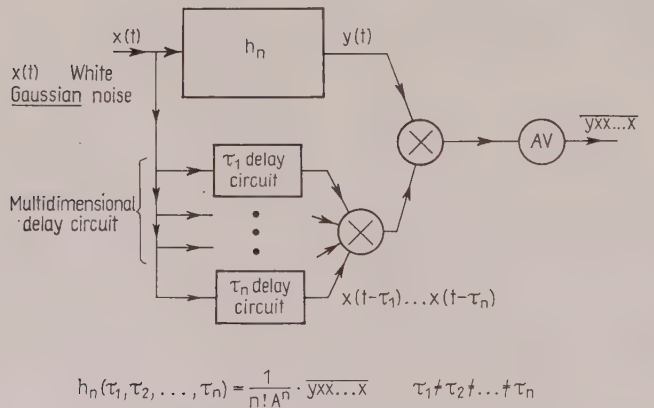


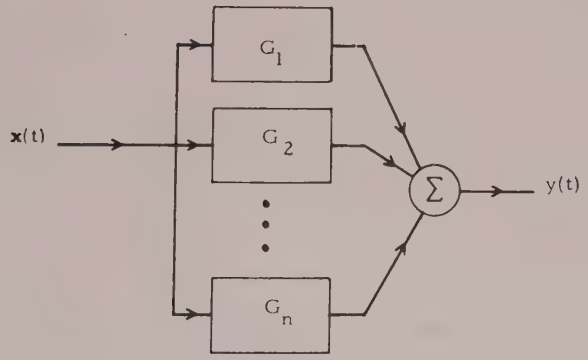
Fig. 9. Measurement of isolated high order kernel.

illustrates the computation of h_n by multi-dimensional crosscorrelation between output and input when the input is white Gaussian noise. The computation of the isolated high order Volterra kernel is a direct extension of the case of a linear system. It should be noted that although we cannot determine the values of $h_n(\tau_1, \ldots, \tau_n)$ along any curve where any

two τ's are equal with this method, one can practically assume continuity of h_n to find such values if needed.

An extremely serious drawback to the Volterra series is that it is not an *orthogonal* functional expansion. Thus if the first n terms have already been computed, then the entire computation must be repeated if $n + 1$ terms are desired. Another important defect of the Volterra method is that computation of the non-isolated kernels is no longer a simple task but involves the solution of a set of simultaneous integral equations.[3]

In much the same way that Legendre polynomials are formed to make an orthogonal function set useful for curve fitting, so can a set of orthogonal functionals for nonlinear system characterization be formed. This was first done by Norbert Wiener, whose work was further simplified by Y. W. Lee and his co-workers.[5] Omitting details of this work, we can say that the general measuring scheme in the determination of Wiener G-functionals is identical to that of determining isolated Volterra kernels. The Wiener representation is illustrated in Figure 10 where the experimentally determined g_1 are used for forming the Wiener G-functional expansion.



$$G_1\ g_1, x(t) = \int g_1(\tau)\, x(t-\tau)\, d\tau$$

$$G_2\ g_2, x(t) = \iint g_2(\tau_1, \tau_2)\, x(t-\tau_1)\, x(t-\tau_2)\, d\tau_1\, d\tau_2 - A\iint g_2(\tau, \tau)\, d\tau$$

$$G_3\ g_3, x(t) = \iiint g_3(\tau_1, \tau_2, \tau_3)\, x(t-\tau_1)\, x(t-\tau_2)\, x(t-\tau_3)\, d\tau_1\, d\tau_2\, d\tau_3 - 3A\int \left[\iint g_3(\tau_2, \tau_2, \tau_1)\, d\tau_2\right] x(t-\tau_1)\, d\tau_1$$

Fig. 10. Orthogonal Wiener *G*-functionals.

After using a Legendre series for minimal square curve fitting, one may go back and, by merely collecting appropriate terms, form the algebraic polynomial of the form:

$$y = a_0 + a_1 x + a_2 x^2 + \dots . \tag{4}$$

In a similar manner, after one has first determined the G-functionals—because of their computational convenience—it is a relatively simple matter to perform some simple integration and collection of terms in order to form the corresponding Volterra expansion. Since a finite G-functional expansion represents the best minimal squares approximation to the system, the derived Volterra expansion must, for the constrained maximum order of the expansion, also represent the best system characterization in the mean square sense.

Continuing to relate ordinary power series expansions to functional expansions we see that just as the polynomial series is more amenable to interpretation than the corresponding orthogonal expansions, so the individual terms of a Volterra expansion are more easily interpreted than the G-functionals themselves. We will examine some of the properties of individual Volterra kernels by applying special transient inputs to the system.

It should be noted that the condition for use of white noise is incompatible with actually performing the necessary measurements experimentally since such an input has infinite variance and it can be shown that as a consequence all of the needed averages (high order correlation functions) will also have undefined variances, not to mention the infinite power content of such a signal. The effect of band-limited noise on the measurement will be seen later.

A Volterra or Wiener G-functional characterization can be used to represent a wide class of nonlinear time invariant finite memory systems. The more theoretical aspects such as the exact, necessary and/or sufficient conditions for this type of characterization to converge will not, and in some cases, cannot be derived.

2. Measurements in the Time Domain

Whereas the previous analysis with random inputs gave rise to a nice mathematical framework, the work now to be presented lacks that foundation but obtains for use some badly needed insight into high order

kernels. If a linear system is excited by an impulse, the response is by definition the kernel or impulse response function of the linear system. The generalization of this type of input in order to determine higher order kernels than the first consists of presenting multi-pulse stimuli where time t and the relative spacing of the pulses are the independent variables tracing out the high order kernels. The method is not orthogonal in the sense that the entire procedure must be repeated if we wish to extend the maximum order of the approximation. A more serious objection to the method is that we will form the Volterra characterization to exactly agree with the real system for one very specialized input and if the system contains more kernels than we assume in the analysis, the derived characterization might prove very poor for other classes of input. The advantage of a random input to characterize the system is that one essentially obtains the best characterization over a wide class of inputs although the representation might not be very good for a particular input of the Volterra series that has been truncated. We will treat in detail the case of double pulse stimuli.

In this case the system is to be approximated by the first two terms of a Volterra series. The word "approximate" could be misleading since we are not using any criterion of goodness de facto because the method is going to make $h_1(\tau)$ and $h_2(\tau_1, \tau_2)$ be such that they completely describe the double pulse experiment. The approximation is that the system is being represented for all inputs as the first two terms of a Volterra expansion. The three pertinent experiments necessary for this method are tabulated below

Experiment 1: $x(t) = Au_0(t)$

Therefore: $y_1(t) = Ah_1(t) + A^2h_2(t, t)$

Experiment 2: $x(t) = Au_0(t - T)$

Therefore: $y_2(t) = Ah_1(t - T) + A^2h_2(t - T, t - T)$

Experiment 3: $x(t) = Au_0(t) + Au_0(t - T)$

Therefore: $y_3(t) = Ah_1(t) + Ah_1(t - T) + A^2h_2(t, t) + A^2h_2(t - T, t - T) + 2A^2h_2(t, t - T)$

where $h_2(\tau_1, \tau_2) = h_2(\tau_2, \tau_1)$ has been used.

It is easily seen that

$$y_3 - y_1 - y_2 = 2A^2 h_2(t, t - T)$$

and that

$$-y_3 \mid_{T=0} + 4y_1 = 2Ah_1(t).$$

The term in the output of experiment 3, $h_2(t, t - T)$, represents a nonlinear interaction between the two stimuli pulses. Note that if $h_2(t, t - T) = 0$ for $T \neq 0$, then for $T \neq 0$ the system would obey what we might call time superposition in the sense that we could add the responses due to each stimulus pulse presented separately and correctly predict the output when both were presented together. Thus $h_2(\tau_1 = \tau_2)$ being nonzero off its main diagonal $(\tau_1 = \tau_2)$ line indicates the degree of time nonlinear interaction between different portions of the input signal. A no-memory squarer followed by an arbitrary linear system and possibly shunted by another linear system has this exact property. Any no-memory nonlinearity followed by a linear system obeys the principle of time superposition. In this general case the order of the required Volterra expansion is determined by the order of the nonlinearity and every kernel is zero everywhere except where $\tau_i = \tau_j$; along these lines the kernel is a wall of impulses. Thus the magnitude of the kernels off their diagonals indicate the extent of the nonlinear interaction in time.

3. Further Topics in Quantitative Measurement: Cascading and Bandwidth Effects

If an isolated Volterra kernel $h_2(\tau_1, \tau_2)$ is followed by a linear system $h(\tau)$ the following relation between input $x(t)$ and output $y(t)$ is readily derived:

$$y(t) = \iint [h_2(\lambda_1 - \xi, \lambda_2 - \xi)\, h(\xi)\, d\xi]\, x(t - \lambda_1) \times x(t - \lambda_2)\, d\lambda_1 d\lambda_2. \tag{5}$$

The new kernel is therefore

$$h_2(\lambda_1, \lambda_2) = \int h_2(\lambda_1 - \xi, \lambda_2 - \xi)\, h(\xi)\, d\xi. \tag{6}$$

For any given (λ_1, λ_2) this can be seen to be just a weighted average of the original h_2 along lines parallel to the main diagonal $\lambda_1 = \lambda_2$; thus we are in effect smearing out h_2 along lines parallel to $\lambda_1 = \lambda_2$.

If an isolated Volterra kernel h_2 is preceded by a linear system h, the following relation between input and output is true:

$$y(t) = \iint [\iint h_2(\lambda_1, \lambda_2)\, h(\tau_1 - \lambda_1)\, h(\tau_2 - \lambda_2)\, d\lambda_1 d\lambda_2]\, x(t - \tau_1)\, x(t - \tau_2) \times d\tau_1 d\tau_2. \tag{7}$$

The new second order kernel is therefore

$$\iint h_2(\lambda_1, \lambda_2)\, h(\tau_1 - \lambda_1)\, h(\tau_2 - \lambda_2)\, d\lambda_1 d\lambda_2 \tag{8}$$

which is simply a two dimensional convolution integral. This will clearly smear the original h_2 both parallel and perpendicular to the $\lambda_1 = \lambda_2$ diagonal line.

In conclusion we see that a linear system preceding a nonlinear one will generally increase the nonlinear interaction time whereas the linear system following the nonlinear one will only increase the overall memory of the system.

Using some of the above results it is interesting to ask: is it possible to find two linear systems, one of which will precede an h_2 and the other which will follow the same h_2 so that the overall behavior of the two resulting systems are identical? The easiest route to a solution is to first define the multi-dimensional Laplace transform as

$$F(s_1, \ldots, s_n) = \int \cdots \int f(t_1, \ldots, t_n)\, e^{-s_1 t_1} \cdots e^{-s t_n} dt_1 \cdots dt_n. \tag{9}$$

Let $h_2(\tau_1, \tau_2) \leftrightarrow H_2(s_1, s_2)$ and $h(\tau) \leftrightarrow H(s)$, $h'(\tau) \leftrightarrow H'(s)$ define the nonlinear system and the two linear systems respectively with h' denoting the one preceding h_2. That is, is it possible for us to find an $h(\tau)$ and $h'(\tau)$ such that the two configurations of Figure 11 are equivalent. Using the

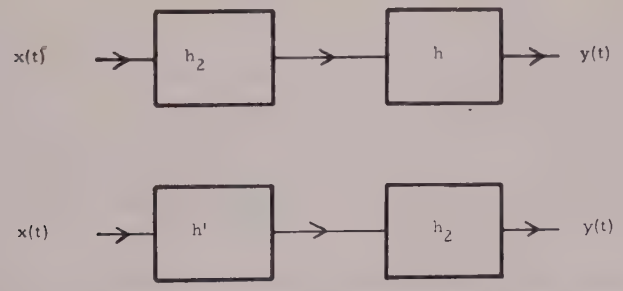


Fig. 11. Permutations of a second order kernel and a linear system.

previously derived material it is not difficult to show that the basic condition for the systems of Figure 11 to be equivalent is that

$$H_2(s_1, s_2)\, H(s_1)\, H(s_2) = H_2(s_1, s_2)\, H'(s_1 + s_2)$$

or

$$H(s_1)\, H(s_2) = H'(s_1 + s_2) \tag{10}$$

for all values of s_1 and s_2. The only non-trivial $H(s)$ and $H'(s)$ for which this is true is $H(s) = H'(s) = e^{-Ts}$ which represents a time delay of T seconds. This result should not be surprising given the previously developed interpretations concerning the difference between linear systems preceding and following nonlinear ones.

We now turn our attention to a seemingly more pressing problem of measurement. Functional analysis theory requires white noise whereas statistical theory tells us that we have an impossible task in obtaining reliable correlation estimates not to mention the impossibility of generating white noise from energy considerations. It will turn out, as expected, that we will only have to use of a bandwidth of an order of magnitude greater than the system's bandwidth in order to make the previously stated equations for the determination of Wiener G-functionals valid for all practical purposes. We call noise of this bandwidth adequate bandwidth noise. Our basic problem is illustrated in Figure 12 in which

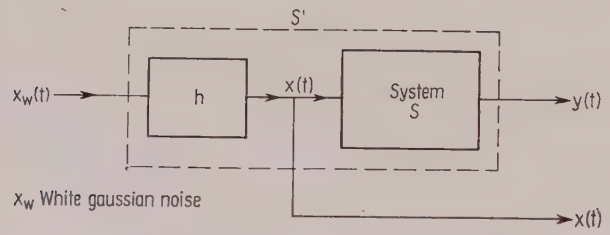


Fig. 12. Measurement of high order kernels with non-white Gaussian noise.

the system we are trying to characterize is denoted by S and only signals $y(t)$ and $x(t)$ are available.[6] $h(t)$ is the impulse response of a lowpass filter which filters the imaginary white Gaussian noise $x_w(t)$ to $x(t)$ the real input into the system. Each kernel of S' and S are related by

$$h_n'(\tau_1, \ldots, \tau_n) = \int \ldots \int h_n(\lambda_1, \ldots \lambda_n)\, h(\tau_1 - \lambda_1) \ldots h(\tau_n - \lambda_n)\, d\lambda_1 \ldots d\lambda_n \tag{11}$$

which is merely the generalization of Equation 8. By taking the multi-dimensional Laplace transform of both sides and dividing through $H(s_1) \ldots H(s_n)$ and then taking the inverse transform it is, at least in principle, possible to get h_n thus characterizing S if we knew h_n. Characterization of S' by h_n however involves the computation of

$$\varphi_{yx_w x_w \quad x_w}(\tau_1, \ldots, \tau_n) \tag{12}$$

which cannot be done since $s_w(t)$ is unavailable in any physical sense.

However $\varphi_{yx_w \ldots x}$ is related to $\varphi_{yx \ldots x}$ by

$$\varphi_{xy_w \ldots x}(\tau_1, \ldots, \tau_n) = \int \ldots \int \varphi_{yx_w \ldots x_w}(\xi_1 + \tau_1, \ldots, \xi_n + \tau_n) \times h(\xi_1) \ldots h(\xi_n)\, d\xi_1 \ldots d\xi_n. \tag{13}$$

The above becomes obvious after we consider the case for a second order kernel as follows:

by definition $\varphi_{yxx}(\tau_1, \tau_2) = \overline{y(t)\, x(t - \tau_1)\, x(t - \tau_2)}^{t}$,

and $x(t - \tau) = \int h(\xi)\, x_w(t - \tau - \xi)\, d\xi$,

therefore:

$$\varphi_{yxx}(\tau_1, \tau_2) = \iint h(\xi_1)\, h(\xi_2)\, \overline{y(t)\, x_w(t - \tau_1 - \xi_1)\, x_w(t - \tau_2 - \xi_2)}^{t} \cdot d\xi_1 d\xi_2 \tag{14}$$

and hence, the result previously stated is proved.

We can again use multi-dimensional Laplace transform techniques to find $\varphi_{yx_w \ldots x_w}(\tau_1, \ldots, \tau_n)$ if desired.

In practical measurements, we use adequate bandwidth noise so that multi-dimensional Laplace transforms and their inversions prove to be unnecessary. The effect of lesser bandwidth of the input can, however, be quantitatively estimated by reference to Equations 11 and 13—i.e., Equation 11 shows that the obtained correlation functions are really the result of convolving the desired correlation functions with the low pass filters' impulse response. Equation 13 shows the corresponding property for the kernel of interest.

4. Experimental Methods

An infrared reflecting pupillometer was used to continuously record the pupil area. A glow modulator, voltage-to-light transducer, was used as stimulus generator in an optical arrangement which prohibited the pupil from affecting the light reaching the retina as it normally does.

An on-line GE-225 digital computer with integral analog-to-digital and digital-to-analog conversion equipment was used in the experimental phase of the study. The on-line computer was used to generate the Gaussian noise stimulus and to input the corresponding response. Stimulus and response were punched on cards in a highly compressed binary format for further data processing. The necessary high-order correlation and averaging programs were written in Fortran and run on IBM 7094 at the Massachusetts Institute of Technology Computation Center. Computation time for evaluation of the first and second kernels seen in Figure 15 was approximately ten minutes.

The on-line computer was also used in the double pulse experiment to generate the required stimuli in a random order to minimize long-term trends and to average the corresponding responses. At the termination of the experiment, the average response data was punched on cards. Another relatively simple program performed the necessary algebraic manipulations on the data and plotted out the second order kernel.

EXPERIMENTAL RESULTS

1. Noise Excitation

Figure 13 illustrates typical input-output data for pupil system's response to random light excitation. From sinusoidal experiments we have seen that the pupil system has negligible response beyond 4 cps. As a consequence, the noise bandwidth was made about 10 cps. The rectification-like nonlinear behavior is clearly evident by noting the sharp difference between the steady state pupil area before and during excitation. The computer was also programmed in such a way as to be able to output the same pseudo-random stimulus many times in order to observe the effect of noise unrelated to the excitation. Input-output data for two identical pseudo-random stimuli functions is seen in Figure 14. Finite

Fig. 14. Pupil response for two identical pseudo-random stimuli functions.

Fig. 13. Effect of asymmetry on pupil response to random excitation.

data length and random pupil variation unrelated to the input were the main sources of error. In the taking of experimental data, the subject is instructed not to blink for periods up to one minute. Inasmuch as such a strain on the subject often introduces unwanted drift as the run proceeds, it was found very useful to substract from the data the best – in the least square sense – second order equation as an attempt to cope with this problem. If this is not done, one may obtain experimental auto- and cross-correlations which have large variation from run to run.

Figure 15 shows the first and second order kernels which are basically of opposite sign. The shape of the first order kernel and that of the main

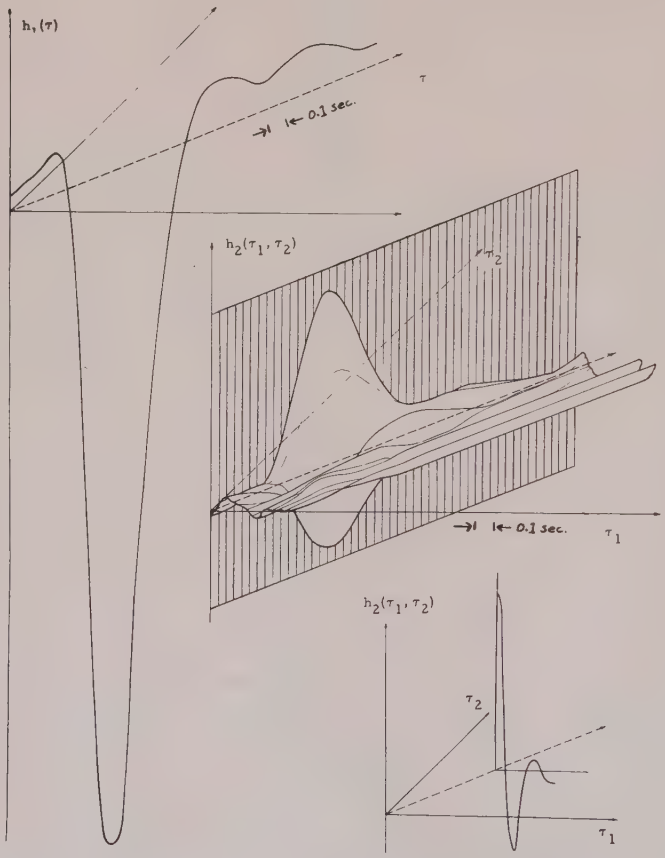


Fig. 15. Wiener kernels $h_1(\tau)$, $h_2(\tau_1, \tau_2)$, and cross-section of h_2.

diagonal slice of the second closely resemble a pupil pulse response. The width of the second order kernel off the main diagonal (see right bottom of Figure 15 for cross-section of h_2) is only about one-half second indicating that there is no second order nonlinear interaction longer than this time.

2. Double Pulse Excitation

Stimulus response records for the double pulse experiment are seen in Figure 16.[11] Basically, the relatively noise-free nature of the data is due to two factors: the pulse experiments were performed at relatively low light with the subject fixating at optical infinity thereby eliminating the large amount of noise related to high levels of illumination and near fixation, and extensive computational averaging of the data.

If one examines the data closely it would be observed that the latent period is the same for the second as for the first pulse response. This implies that the higher order kernels all have the same inherent time

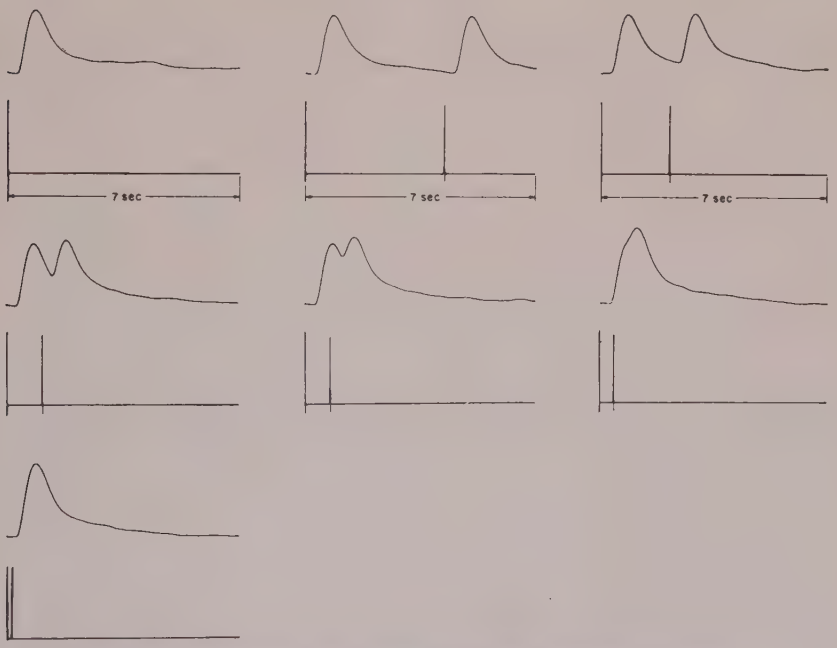


Fig. 16. Double pulse light stimulation of the human pupil system.

delay which would seem to indicate that the time delay mechanism is not intimately connected with the nonlinear interaction process.

Figure 17 shows the computed second order kernel, under the assumption that the system can be completely described by just the first two terms in a Volterra expansion. That the shapes of the slices shown in

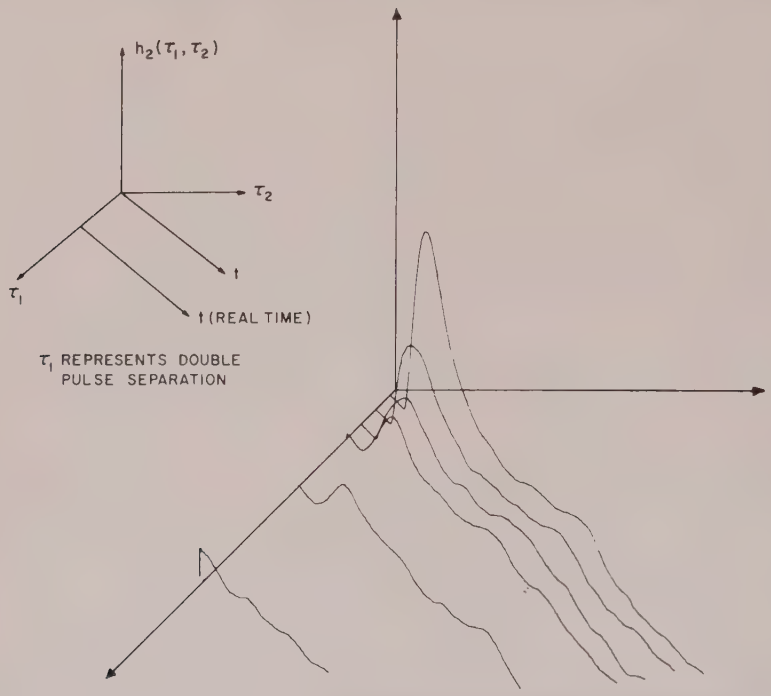


Fig. 17. Second order Volterra kernel as derived from double pulse experiment.

Figure 17 closely resemble the elementary pupil pulse responses again emphasizes the importance of nonlinear interaction occurring before the basic time shaping mechanism occurs in the system. Note also that the spread off the main diagonal is appreciably greater in the double pulse experiment than in the noise experiments.

DISCUSSION

The difference in spread of the two kernels found is surprising. One can hypothesize that the relatively narrow width of the random kernel is due to the pupil being at a small average area because other experiments

indicate a larger bandwidth of the pupil system when the pupil is small. Another possibility is that the continued excitation puts the neurological portion of the system into a state where nonlinear interactions are greatly reduced.

The basic sign difference between the first and second order kernels represents the scale compression nonlinearity mentioned earlier. This is analogous to the first two nonconstant terms of a Taylor series expansion for a logarithmic type curve having opposite signs.

For some values of (τ_1, τ_2), h_2 obtained by random stimulation is of the same sign as the first order kernel. One might therefore conclude that there exists some nonlinear interaction whose effect is to enhance or facilitate the response at a particular time due to the stimulus some time before. This is in contradiction to the double pulse experiment described here. It would indeed be interesting to see the results of experiments suitably designed so as to emphasize this phenomena if it exists. A more probable, though less provoking, interpretation is that the negative going portion of h_2 will have no effect on the output if we were to carry out the function expansion further.

In attempting to characterize a nonlinear system such as the pupil light reflex, it is good to have as many alternative methods of investigation as possible.

Unfortunately, it is often difficult to quantitatively relate the results of the different methods described here. The first two methods of characterization have been used with considerable success in describing biological systems: sinusoidal analysis is a very familiar approach in linear system analysis and people have developed a good feeling for it and its application; heuristic modeling guides one, hopefully, to the threshold of understanding or—which is even more important—into asking questions one would never have thought of without the model.

The functional analysis approach to the human pupil light reflex was attempted for two reasons. First, we wished to develop a canonical representation of the system which would make evident in a mathematical manner certain characteristic nonlinearities of the system. Secondly, we wanted to explore the usefulness of the Wiener-Lee theory of nonlinear systems on a real problem of characterization where a comparison with other methods could be made. The main application of the theory to date has been to solve problems which would be less elegantly solved by other techniques.

SUMMARY

Three general approaches to nonlinear input-output systems are contrasted.

(1) Describing function based mainly on sinusoidal excitation functions.
(2) Heuristic experimental analysis coupled with model building based on intuitive guesses as to component connections, component characteristics, and efficient particular transient inputs.
(3) Wiener G-functional expansions based upon stochastic driving functions.

The mathematical methods for experimental measurement of the kernels of a system are described for random signals and multi-pulse inputs.

The effect of topological arrangement of component boxes and of input noise bandwidth on the methods above are discussed. Apparatus and experimental design using an on-line digital computer is briefly described.

The experimental results are separately displayed for random noise and multi-pulse excitation functions.

The first (h_1) and second (h_2) order kernels are of opposite sign and thus represent the first terms of a saturation nonlinearity. (h_1) and the main diagonal slide of (h_2) resemble the pupil pulse response and thus suggest that the nonlinear interacting occurs before the basic quasi-linear time shaping mechanism. The width of (h_2) off the main diagonal is less than one-half second, indication that there is no second order nonlinear interaction longer than this time. Similarly the time delay mechanism is shown to be independent of the nonlinear interaction.

The differences noted between the kernels obtained from the random and multi-pulse experiments include greater (h_2) width off the main diagonal in pulse experiments and (h_2) valleys only in the noise experiments.

SUPPORT ACKNOWLEDGEMENT STATEMENT

We wish to acknowledge partial support from the following grants and contracts: National Institutes of Health grants NB-3055, NB-3090, MH-06175; Office of Naval Research (Nonr-609(39)), Nonr-1841(70); Air Force (AF-33(616)-7282 and 7588), AFOSR 49(638)1313; and Army Chemical Corps (DA-18-108-405-Cml-942) at the Massachusetts Institute

of Technology. More recently at the Presbyterian St. Luke's Hospital and the University of Illinois in Chicago, Illinois, we have been supported in part by special grants awarded the Biomedical Engineering Department from the W. Clement and Jessie V. Stone Foundation, the Smith Kline and French Foundation and Public Health Research grant FR-05477 from the General Research Support Branch, Division of Research Facilities and Resources. Also National Institutes of Health grants 7-RO 1-MH-11907-01 PMY, 1-RO 1 NB-06197-01, 1-RO 1-NB-06487-01, and Office of Naval Research (Nonr-609(39)).

This paper is reprinted with permission of the Presbyterian St. Luke's Hospital Medical Bulletin from Vol. 5, No. 2, pp. 89–105 (April 1966).

REFERENCES

1. Stark, L., and Sherman, P. M. A servoanalytic study of the consensual pupil reflex to light. *J. Neurophysiol.* **20**, 17–26 (1957).
2. Stark, L., and Cornsweet, T. N. Testing a servoanalytic hypothesis for pupil oscillations. *Science* **127**, 588 (1958).
3. Katzelson, J. Synthesis of nonlinear filters. Sc. D. Thesis, Elect. Eng. Dept., Massachusetts Institute of Technology, Sept. 1963.
4. Lee, Y. W. Statistical theory of nonlinear systems. Class Notes for Massachusetts Institute of Technology Elect. Eng. Course 6.572.
5. Lee, Y. W. Statistical Theory of Communication. John Wiley, New York, N. Y.
6. Schetzen, M. Measurement of the kernels of a nonlinear system by crosscorrelation with Gaussian non-white inputs. Quarterly Progress Report No. 63, R.L.E., Massachusetts Institute of Technology, Oct. 15, 1961, pp. 113–117.
7. Stark, L.: Stability, oscillations, and noise in the human pupil servomechanism. *Proceedings of the Institute of Radio Engineers*, **47**, 1925–1939 (Nov. 1959).
8. Stark, L. Biological rhythms, noise and asymmetry in the pupil-retinal control system. *Annals of the New York Academy of Sciences*, **98**, 1096–1108 (October 30, 1962), Article 4.
9. Stark, L., and Baker, F.: Stability and oscillations in a neurological servomechanism. *J. Neurophysiol.* 22, 156–164 (1959).
10. Clynes, M.: The non-linear biological dynamics of unidirectional rate sensitivity illustrated by analog computer analysis, pupillary reflex to light and sound, and heart rate behavior. *Annals of the New York Academy of Sciences*, **98**, 806–845 (Oct. 30, 1962), Article 4.
11. Baker, F. H.: Pupillary response to double pulse stimulation: A study of nonlinearity in the human pupil system. *J. Opt. Soc. Am.* **53**, 1430 (1963).

W. GREY WALTER

Burden Neurological Institute,

Bristol, England

Electric Signs of Expectancy and Decision in the Human Brain

The aim of this paper is to present to you an outline of some discoveries we have made during the last few years relating to brain function. As Professor McCulloch has said, these are concerned both with the establishment of general theories of the brain and also with the specifications of mechanisms which may help brain modelers to mimic or replicate or surpass the functions of the brain. The chief phenomenon that I am going to describe is what we called, when we first corroborated its reality in the brain, the Contingent Negative Variation or CNV. This term is self-explanatory and the interest of it to this audience is mainly that the effect is *contingent*. It is an electro-negative potential change on the surface of the brain in the frontal lobes in man, related not to the intensity or modality of signals, but to their significance and their contingent association with other signals and actions. The other pet name for this effect is the "Expectancy Wave". The reason for using this term is that Contingent Negative Variation doesn't go very well into French or German or Russian, whereas expectancy seems to be a term which is represented in other languages. In fact, this effect seems to reflect the degree of expectancy felt by an individual that something interesting and important is going to happen to him. First I am going to assert what I could have proved in detail, that this is a true brain effect.

Figure 1 shows an X-ray of electrodes implanted in the brain of a patient for therapeutic purposes; although most of the effects I am going to describe are derived from normal people with electrodes on their scalps, the confirmation of this effect has been obtained by recording from inside the brain in conscious, normally moving patients, so that we know that these effects are actually due to brain activity. They are not electrode artifacts and they are not due to movements of the eyes or scalp. As you see in this X-ray, we have had many patients with electrodes scattered all over the brain in the cortex and on the cortex, and we've studied

Fig 1

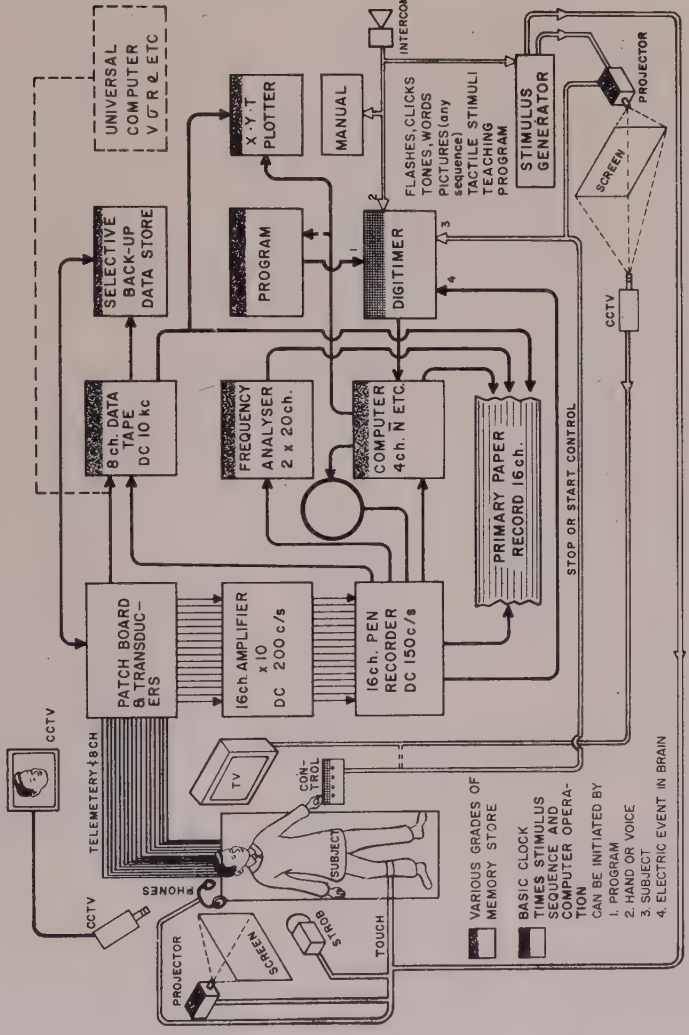

BURDEN NEUROLOGICAL INSTITUTE SCHEMATIC OF EXPERIMENTAL RIG
UNIVERSAL COMPUTER V G R & ETC
SELECTIVE BACK-UP DATA STORE
8 ch. DATA TAPE DC 10 kc
FREQUENCY ANALYSER 2 x 20 ch.
COMPUTER 4 ch. N ETC.
PROGRAM
X Y T PLOTTER
MANUAL
DIGITIMER
INTERCOM
FLASHES, CLICKS TONES, WORDS PICTURES (any sequence) TACTILE STIMULI TEACHING PROGRAM
STIMULUS GENERATOR
PROJECTOR
SCREEN
CCTV
PRIMARY PAPER RECORD 16 ch.
STOP OR START CONTROL
PATCH BOARD & TRANSDUCERS
16 ch. AMPLIFIER x 10 DC 200 c/s
16 ch. PEN RECORDER DC 150 c/s
TELEMETERY 8 CH
TV
CONTROL
SUBJECT
PHONES
TOUCH
BOOTS
VARIOUS GRADES OF MEMORY STORE
BASIC CLOCK TIMES STIMULUS SEQUENCE AND COMPUTER OPERATION
CAN BE INITIATED BY
1. PROGRAM
2. HAND OR VOICE
3. SUBJECT
4. ELECTRIC EVENT IN BRAIN

Fig. 2

altogether, I suppose, about five or six thousand electrodes, so we know a good deal about the details of brain function. Figure 2 shows the rig we use. On the left-hand side, the subject is surrounded by a conglomeration of recording and stimulating equipment. The point I would like to draw your attention to particularly is that we are able to use what we call "semantic stimuli". We can use flashes, clicks, tactile stimulators to the skin and so forth, but as well as this, there are closed circuit television systems whereby we can present pictures or we can use auditory verbal stimuli. On the right are various types of computer and recording equipment and you will notice the basic fact that, as in all computing equipment, whether in the flesh or the metal, there are a large number of storage systems and several clocks. Clocks and stores are the essence of a computing system and we use our clocks and stores to study the clocks and stores in the brain.

Figure 3 shows the standard procedures we use for investigating contingent responses in the brain. The top traces are the raw brain record with the signals indicated below it—the flash followed by a click and the response which are hard to see in the primary record. Below that is one set of presentations which is used for averaging. There are twelve presentations each and the average is then written out directly on the original record, also, of course, stored on tape and computed later by various means. The bottom line indicates the conjugation of this paradigm and I should perhaps say at this point that what we claim to be studying is the grammar of the brain. The way in which the brain conjugates and declines the verbs and substantives of its essential transactions with environment. The bottom circles are the procedures we use to study first of all habituation (which, as Ashby showed many years ago, is one of the basic properties of this system)—that is, the very subtle way in which a system will cease to respond to insignificant monotonous stimuli and then respond again to them if they become significant. After this we have the association of signals in the strictly Pavlovian sense. And then the association with an operant response and this is the crucial transition in our experiments.

Figure 4 shows the complete polygraphic record with the EEG, its averages, and also—and this is a feature I shall not have time to describe in detail, but it is a very important one—the continuous recording of the autonomic variables—what you might call the housekeeping of the body. In order to appraise the importance of these effects, one must make sure that the housekeeping of the body is either normal or is adequately

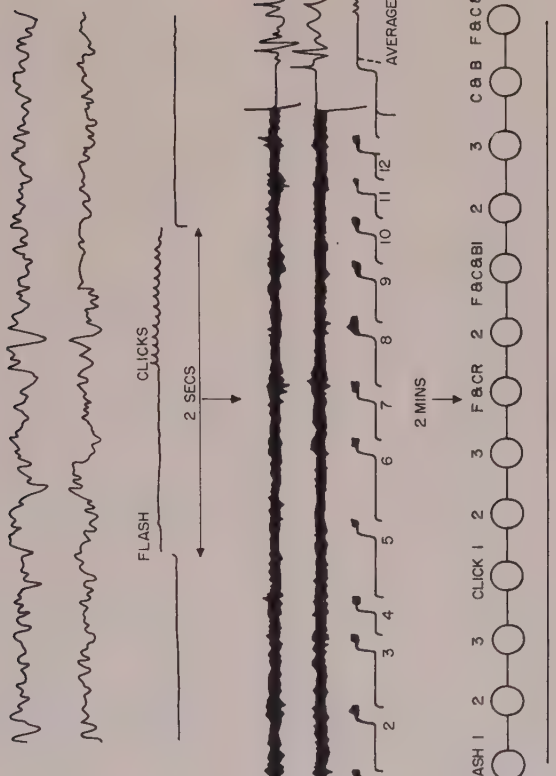

FLASH
CLICKS
2 SECS
AVERAGE
1 2 3 4 5 6 7 8 9 10 11 12
2 MINS
FLASH I 2 3 CLICK I 2 3 FACR 2 FACABI 2 3 CAB FACAB
1/2 HOUR

Fig. 3

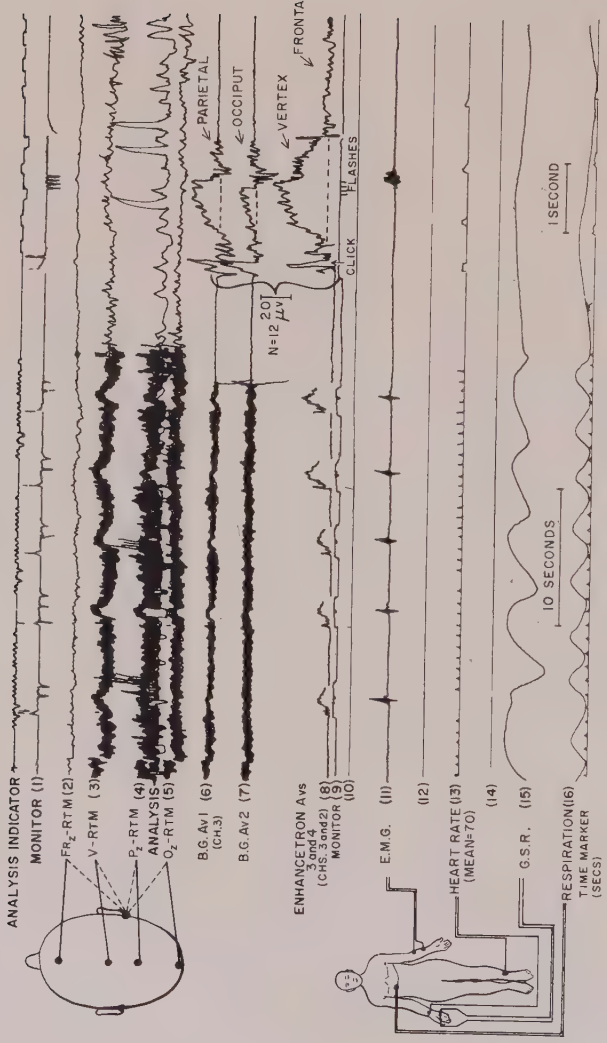

ANALYSIS INDICATOR
MONITOR (1)
FR$_z$-RTM (2)
V-RTM (3)
P$_z$-RTM (4)
ANALYSIS
O$_z$-RTM (5)
B.G. Av1 (6)
(CH.3)
B.G. Av2 (7)
ENHANCETRON Avs
3 and 4
(CHS. 3 and 2) (8)
MONITOR (9)
(10)
E.M.G. (11)
(12)
HEART RATE (13)
(MEAN=70)
(14)
G.S.R. (15)
RESPIRATION (16)
TIME MARKER
(SECS)
PARIETAL
OCCIPUT
VERTEX
FRONTAL
N=12 20 μv
CLICK
FLASHES
10 SECONDS
1 SECOND

Fig. 4

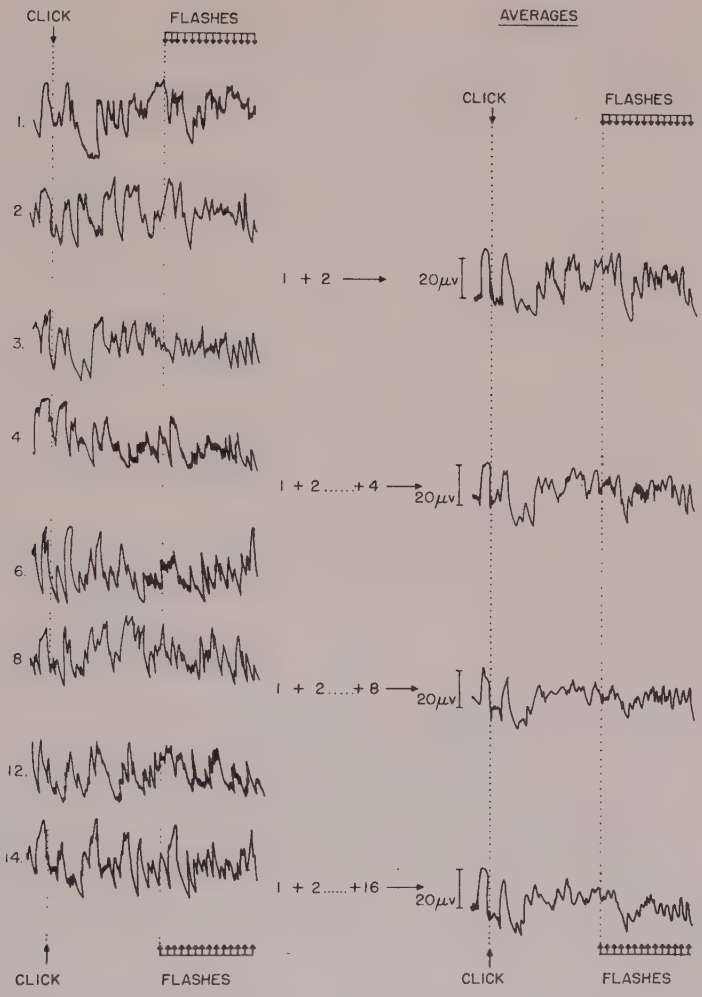

CLICK
FLASHES
AVERAGES
1.
2
3.
4
6.
8
12.
14.
1 + 2
20μv
1 + 2.....+4
20μv
1 + 2.....+8
20μv
1 + 2.....+16
20μv
CLICK
FLASHES
CLICK
FLASHES

Fig. 5

monitored, so we record pulse rate, respiration, muscle activity, GSR, and so forth, continuously in 16 channels. This represents the more or less raw record.

Figure 5 presents an example of what happens with averaging. I put this in just to emphasize that we are well aware of the dangers and difficulties of averaging and autocorrelation procedures. These eliminate certain features and emphasize others. They emphasize the features we are interested in at the moment—as you will see, the expectancy wave or CNV is admirably emphasized by these procedures – but what averaging does, of course, is progressively to reduce the noise to signal ratio in the system according to the square root of the number of samples, so that components which are phase or time locked to the computer are progressively clarified. This is just an example of averaging and it shows the degree of resolution we get with the rather small number of samples we use. We have to use a small number of samples because if we go on presenting exactly the same situation indefinitely with a human subject he gets bored and goes to sleep or goes home or at any rate packs up and ceases to take an interest.

In Figure 6 an example of a record taken from a subject with implanted electrodes giving a direct comparison with scalp recording is shown. The response to a series of clicks on the left-hand side and the same clicks later on, on the right-hand side, and below these are the responses to clicks and flashes. From each of these figures I shall try to extract one particular point, and this one illustrates an extremely important truth about the frontal lobes of man, which is that they are an intrinsic part of the sensory system. All information in all modalities, whether visual, auditory, tactile, visceral, is relayed in great profusion to the frontal lobes through the reticular formation that Professor McCulloch and his colleagues described in model form at this meeting. The record you see in this figure indicates the response of the same brain region to these two modalities, and if they had been tactile and visceral or what you will, the same regions of the brain would respond in this sort of way. So this is the first important fact—our frontal lobes respond to all sensory stimuli. The silent areas are silent only because they are listening very intently to all that goes on in the outside world and in the body.

Figure 7, from the same subject a few seconds later, shows a later stage in the paradigm I spoke of, where the signals are associated with an operant response and you can see the CNV has appeared because now

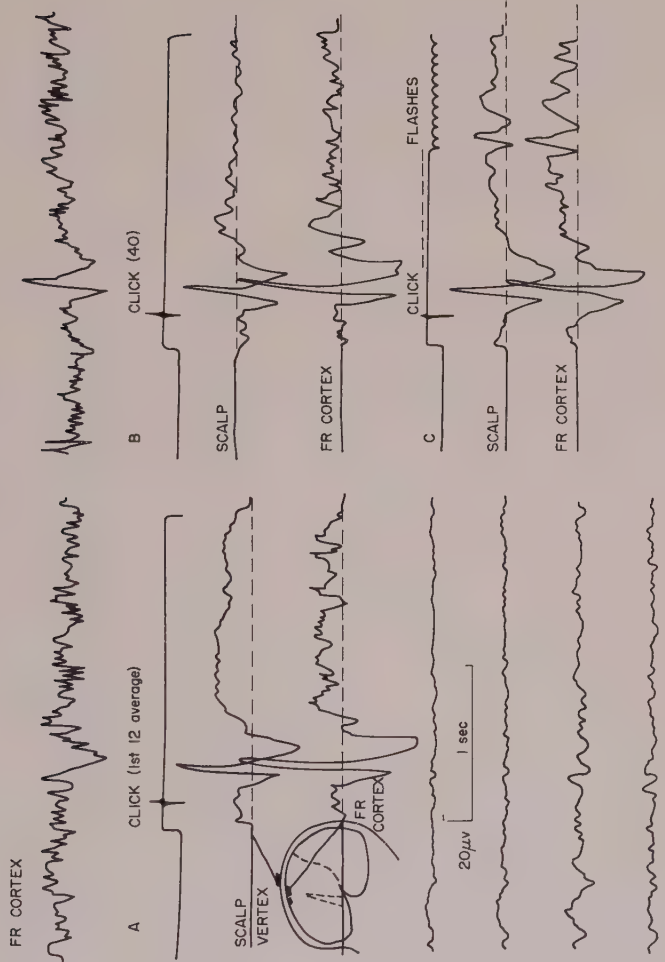

FR CORTEX
A
CLICK (1st 12 average)
SCALP
VERTEX
FR CORTEX
B
CLICK (40)
SCALP
FR CORTEX
C
CLICK
FLASHES
SCALP
FR CORTEX
12
13
14
15
20μv
1 sec

Fig. 6

the subject has been asked to press a button as soon as he sees the second stimulus, which in this case is flashes. Since we consider we are talking about the grammar of the brain, we call the first stimulus "conditional" and the second one "imperative". We are saying to the subject in effect through the stimuli: "*If* there are clicks (conditional), there *will* be (future indicative) flashes so (imperative) *press* the button!" When we introduce the imperative clause, "press the button", we see developing between

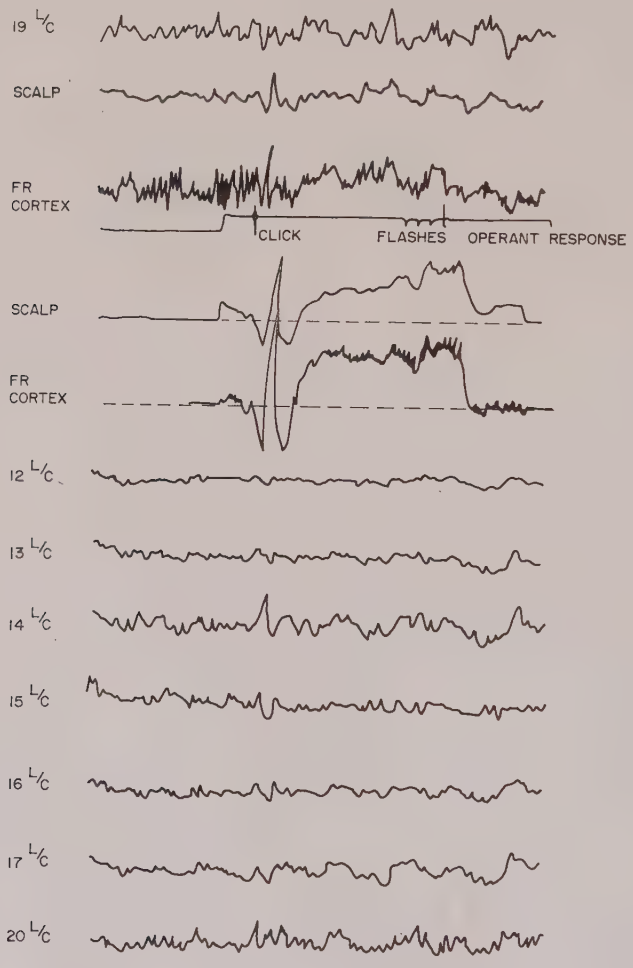


Fig. 7

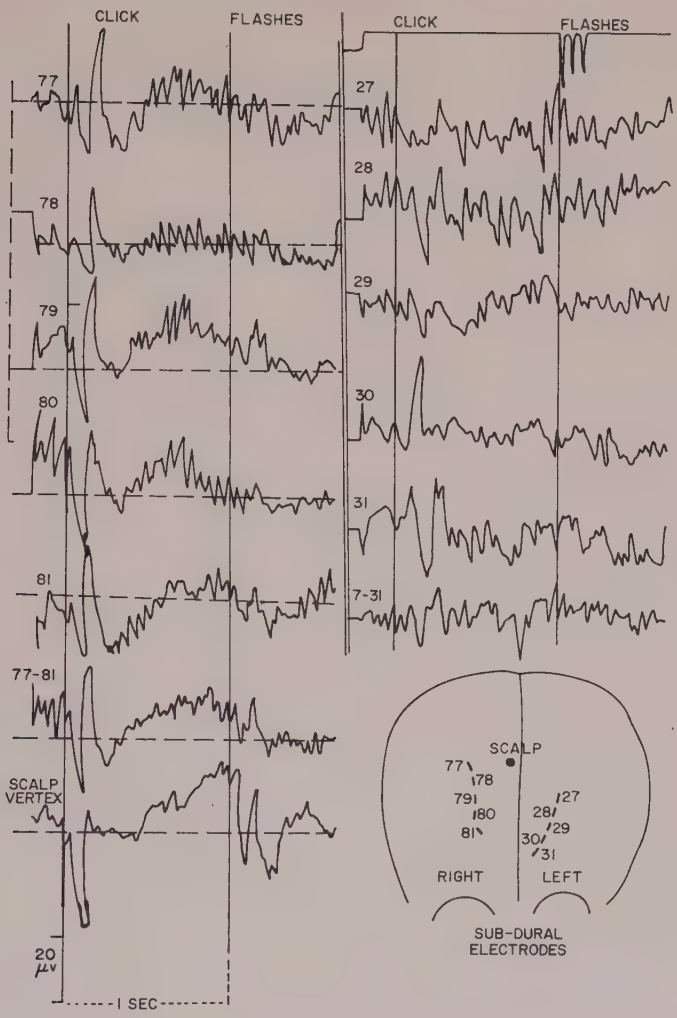


Fig. 8

the conditional stimulus and the imperative one this slow negative wave. This is the Contingent Negative Variation or CNV which starts about 200 or 300 milliseconds after the conditional response, lasts until the imperative stimulus and terminates very abruptly, as you can see, both on the cortex with the implanted electrodes and on the scalp with the action by the subject—with his decisive action. You can also see in the

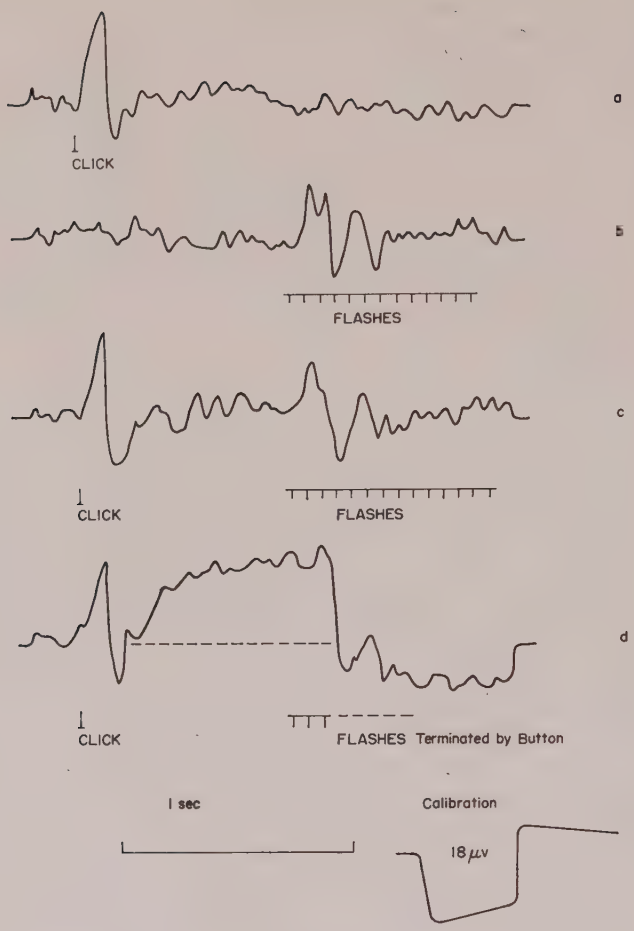


Fig. 9

other records further down all of the parts of the brain responding in their own particular versions to the conditional stimulus. The distribution of this effect is peculiar. Figure 8 shows a subject, again with implanted electrodes. The numbers on the X-ray figure at the bottom are electrodes on the cortex, subdurally on the left and right side, which are numbered and the traces are numbered accordingly. Both the evoked responses proper and the CNV or expectancy wave are distributed in a very patchy way over the brain. The electrodes on the left-hand side, that is, your

right as you look at it, electrodes 27 through 31, show evoked potentials but almost no sign of an expectancy wave, while those on the opposite side show in patches, signs of this expectancy wave. We have confirmed this in many cases and the effect develops in patches all over the frontal cortex, both on the external convexity cortex, the medial cortex, and the orbital cortex—but only in patches. They are not always the same patches on each occasion and they are not always the same patches for different stimuli. So here we have an example of what I call "dispersive convergence". The information reaches the cortex through the reticular system by "idiodromic projection", which means private lines, and when it gets there, it is distributed in little patches or bundles in the cortex.

Figure 9 shows a record from a normal subject with no implanted electrodes, but with conventional scalp electrodes and this is the canonical form of the response development. As it is literally canonical, we have made templates from these records from about 150 normal subjects. And, to our astonishment, this particular feature is by far the most constant and consistent feature in the whole repertoire of brain activity. Some people have α-rhythm, some people have not—I have none, for example. Some may have θ-rhythm, some of you at the back at the moment may have δ-rhythm, but whoever you are, everybody of your age and intelligence and competence I can guarantee absolutely will show this pattern. At the top of the figure we have the response to an isolated stimulus, the click stimulus, which would habituate if we went on repeating it. The next one down, the responses to flashes, which again would die away if we went on repeating it. The next down is the association of those signals which restores them, but again, this would habituate if nothing were done about it; but again, when we asked the subject to perform an action—in this case, to press a button to the imperative flashes, we get the CNV or expectancy wave rising slowly very much like the time-base of an oscilloscope with a saw-tooth waveform starting after the conditional response, building up to the moment of action and then going back to zero very abruptly.

In Figure 10 the effect of withdrawing the imperative stimuli without warning is shown. The second line down shows a true conditional response, probably the only example of a really authentic conditional response. That is the little hump on the top of the second line where the subject would have expected the imperative clicks to appear; they didn't appear, so there could be no action, but his expectancy was still there and

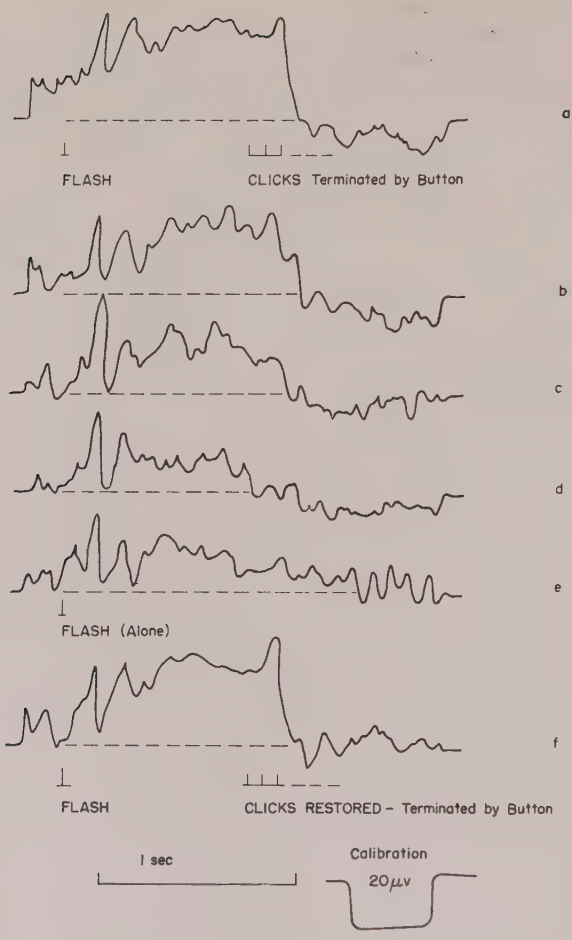


Fig. 10

his brain responded as if there were going to be something interesting to respond to. And as we go on diluting the association by withholding imperative stimuli, the expectancy wave declines steadily. It takes about 24 trials to disappear almost completely. In the bottom trace we have restored the association and the expectancy wave reappears quite quickly. This again is an absolutely canonical form that we have repeated with many hundreds of normal subjects and patients. With all normal subjects this is the effect we get. The clinical story is different, and it is not relevant

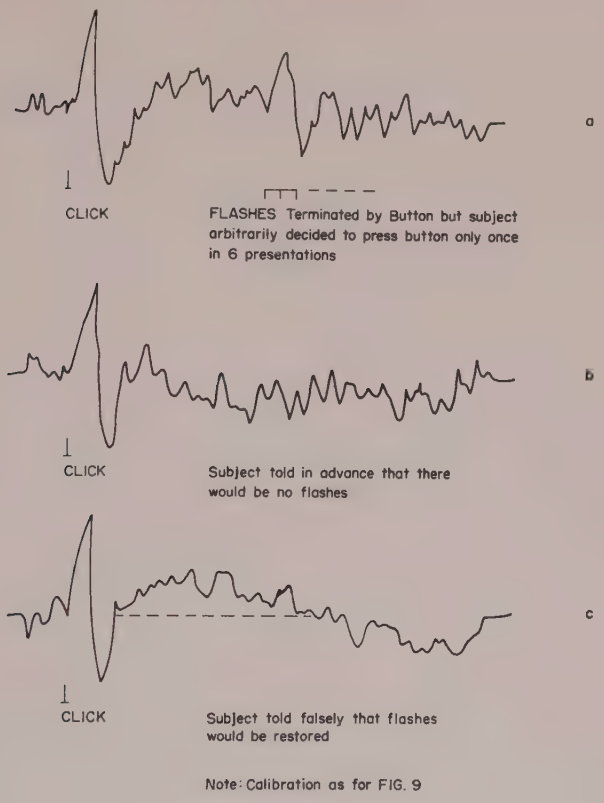


Fig. 11

to this audience to discuss the clinical applications, but I may say that the differences between normal people and clinical conditions in this respect are extremely striking and diagnostically important.

Figure 11 is in contrast to this. It is the effect of social instruction about extinction. In the previous figure it took 24 trials for the expectancy to die down to something approaching zero following direct experience of nonreinforcement. In this figure there are represented three ways in which the expectancy wave can be suppressed or abolished. The first is by telling the subject not to bother about the operant response. After you have established the expectancy wave, you say: "Don't bother to press the button for a while". This figure shows that without any warning to me, the subject decided not to cooperate on a particular occasion.

The top trace shows the disappearance of the expectancy waves; this was the same subject as was in the previous two figures.

The second trace is one of the most interesting of all the records I am going to show you, because, on this occasion, I told the subject beforehand that I was not going to provide him with the imperative stimulus. Instead of withdrawing without warning, I gave an explicit warning. And the effect of this was to abolish the expectancy wave at once. In other words, his set having been changed by my warning, a single phrase from me was equivalent to 24 or 30 direct experiences. One of the facilities we have now is to be able, as it were, to titrate the strength of social instruction against the power of direct experience. We can say that a particular word or phrase from the experimenter to the subject, whoever it may be, is equivalent to so many direct experiences. In the case of children, for example, or people faced with learning difficult tasks, one word of warning or advice from the instructor is worth say, 30 direct trials of untutored experience; we can measure the actual power of instruction by the changes in the expectancy wave.

The third trace was an accident. I intended to restore the imperative stimuli, but I didn't quite push the plug home in the jack; the subject didn't know this and in the first few trials—these are all averages remember—he expected to have reinforcement and produced an expectancy wave. But after three mistrials he decided that something had gone wrong – he was being fooled, or the apparatus had broken down, so there was no expectancy wave. The average shows the compound of the three when he did and the three when he didn't expect reinforcement. So again we can actually measure the power of a lie—we can bluff the brain system and we can see the extent to which it was bluffed, how long it was bluffed for, and whether it is more easily fooled by one experimenter or by another and so forth. So here we have something in the brain which reflects quite accurately the effects of social factors as well as of physiological ones.

Figure 12 shows something a little more subtle which is what we call "equivocation" – what psychologists call partial reinforcement. In the figures so far, the reinforcements have been absolute, all conditional stimuli were always followed by imperative stimuli. But on this occasion we began to dilute the probability of association by interspersing unreinforced conditional stimuli. The contingency of association began to fall, and in the end, as seen in the fourth trace, the association has been

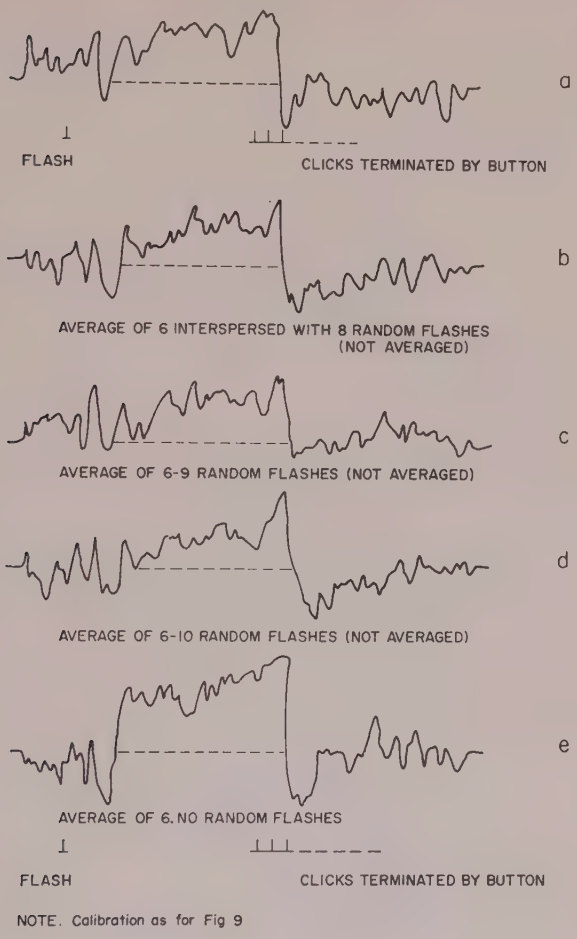


Fig. 12

diluted to a point at which the probability was only about a half, and the contingency having fallen to this level the expectancy wave also falls to the same level. Restoration of the association over six trials restored the expectancy wave. This again is an absolutely invariable factor, although the criteria of probability vary from person to person. This is a very personal factor and one which, again, is very important in the clinic. Figure 13 shows the detail of this effect; the second trace is the analysis of the trials with unreinforced conditional stimuli and the bottom one

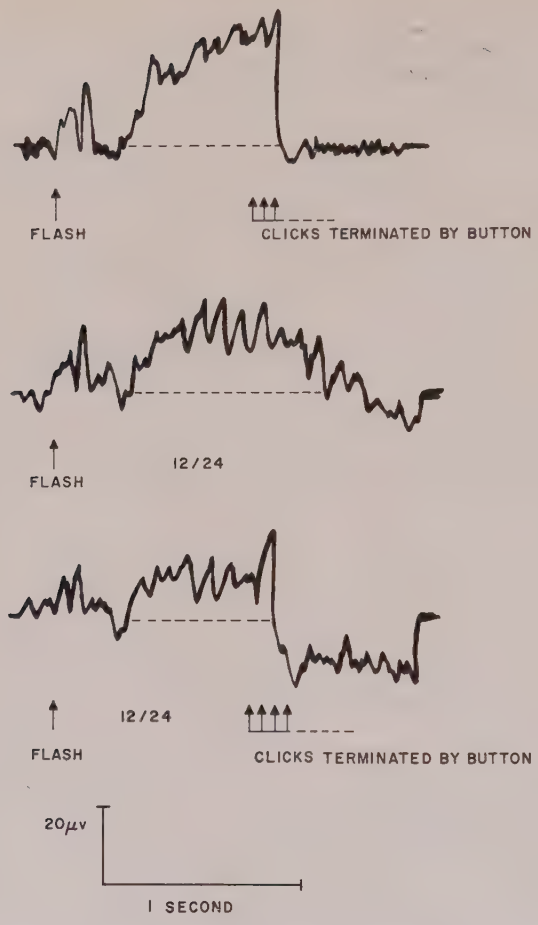


Fig. 13

of the reinforced conditional stimuli, showing that both the reinforced and the unreinforced produce some degree of expectancy because the subject could not know which stimuli were going to be reinforced, but these added up to a smaller figure than the top one where the association was absolute.

Figure 14 is a histogram in which the height of the bars is the height of the CNV or expectancy wave in microvolts as seen in averages of six trials in each case. The figures below are the ratios of the reinforcement by unconditional flashes of conditional clicks and below that are the

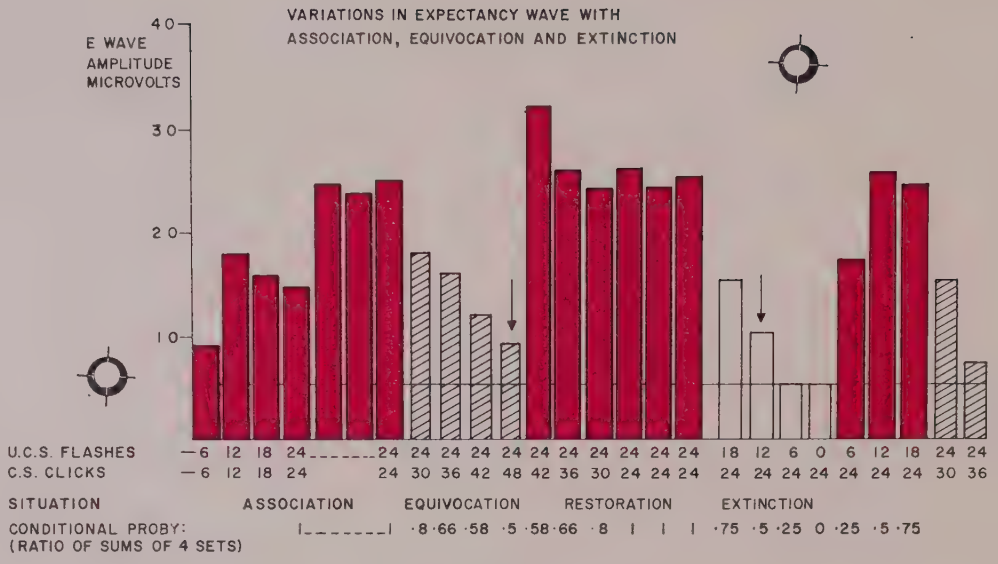

VARIATIONS IN EXPECTANCY WAVE WITH
ASSOCIATION, EQUIVOCATION AND EXTINCTION
E WAVE
AMPLITUDE
MICROVOLTS
40
30
20
10
U.C.S. FLASHES —6 12 18 24 24 24 30 36 42 48 42 36 30 24 24 24 18 12 6 0 6 12 18 24 24
C.S. CLICKS —6 12 18 24 24 24 30 36 42 48 42 36 30 24 24 24 24 24 24 24 24 24 24 30 36
SITUATION ASSOCIATION EQUIVOCATION RESTORATION EXTINCTION
CONDITIONAL PROBY: (RATIO OF SUMS OF 4 SETS) ·8 ·66 ·58 ·5 ·58 ·66 ·8 1 1 1 ·75 ·5 ·25 0 ·25 ·5 ·75

Fig. 14

probabilities calculated on the basis of 24 trials. It took 24 trials on the average in this particular subject (and this is a normal figure for the average normal person) for the expectancy wave to build up to its plateau of about 25 microvolts. You see the first bars on the left rise slowly to reach the plateau figure and the cross-hatched bars are where we introduced equivocation. The probability falls again taking a sample of 24 trials until at the fourth bar the probability has fallen to 0.5 and the expectancy wave has also fallen to about half its value. Now this is a sane, reasonable, balanced sort of individual. His expectancy wave, the electrical potential of his brain, mirrors quite precisely the objective probability of association. Restoration again produces a slight overshoot and it regains its original 25 microvolts. The open bars indicate the extinction trials, where the imperative stimulus was withdrawn altogether. Here again, when the probability, averaged over 24 trials, falls to half, the expectancy wave has also fallen to half. And the same process can be repeated again and again in a cooperative subject. The interesting feature of this is that you can go on doing this for thousands of trials; this effect does not habituate—it can persist for weeks, months, or years. We have studied the same subject, one of my colleagues, now for nearly four years since we first discovered this and he showed exactly the same pattern last week as he did four years ago.

Figure 15 gives you an idea of what this is doing in the brain; it seems to be acting as a sort of cortical primer, shortening reaction time and economizing effort. The top trace shows the height of the expectancy wave in microvolts and the very bottom one is the histogram of the distribution of muscle activity as the subject presses the button. In A the probability was one, the expectancy wave was 16 microvolts and the mode of his reaction time was 200 milliseconds with some responses as short as 100 milliseconds. I then reduced the probability, again by equivocation. When the probability in this subject was 0.8, the expectancy wave was down to 13 microvolts, the EMG latency mode was 280 and the shortest latency 140 milliseconds. On the extreme right I had reduced the probability to 0.7; in this case the person was slightly less sanguine than the previous subject (and this is a very precise measure of the optimism of the subject) so when the probability was 0.7, with the odds still in favor of reinforcement, his expectancy wave had subsided absolutely to zero. The EMG latency mode was now 360 milliseconds and the shortest was 200. This can be represented again in a graph. Figure 16

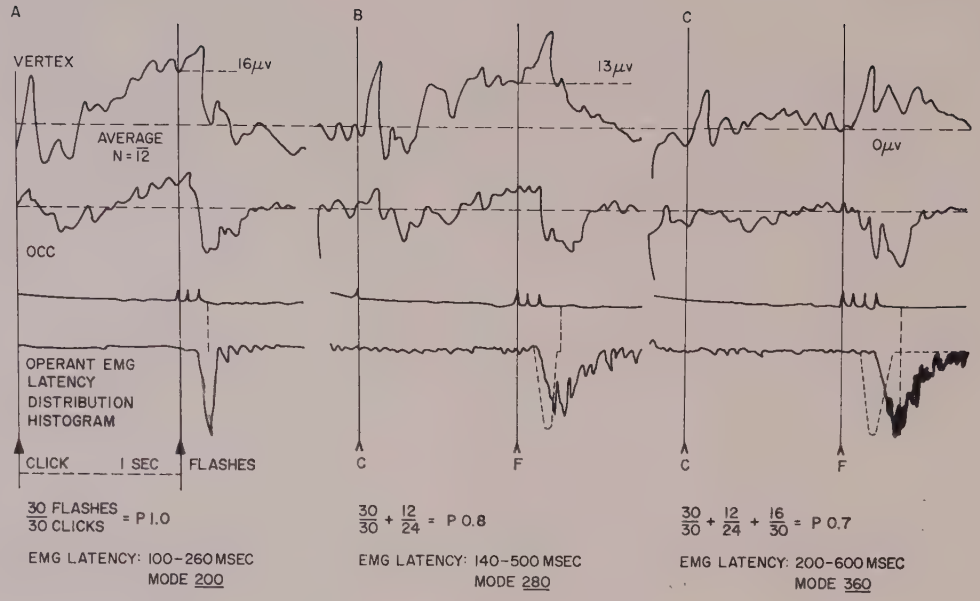

A
B
C
VERTEX
16μv
13μv
0μv
AVERAGE
N = 12
OCC
OPERANT EMG
LATENCY
DISTRIBUTION
HISTOGRAM
CLICK
1 SEC
FLASHES
C
F
C
F
30 FLASHES / 30 CLICKS = P 1.0
EMG LATENCY: 100–260 MSEC
MODE 200
30/30 + 12/24 = P 0.8
EMG LATENCY: 140–500 MSEC
MODE 280
30/30 + 12/24 + 16/30 = P 0.7
EMG LATENCY: 200–600 MSEC
MODE 360

Fig. 15

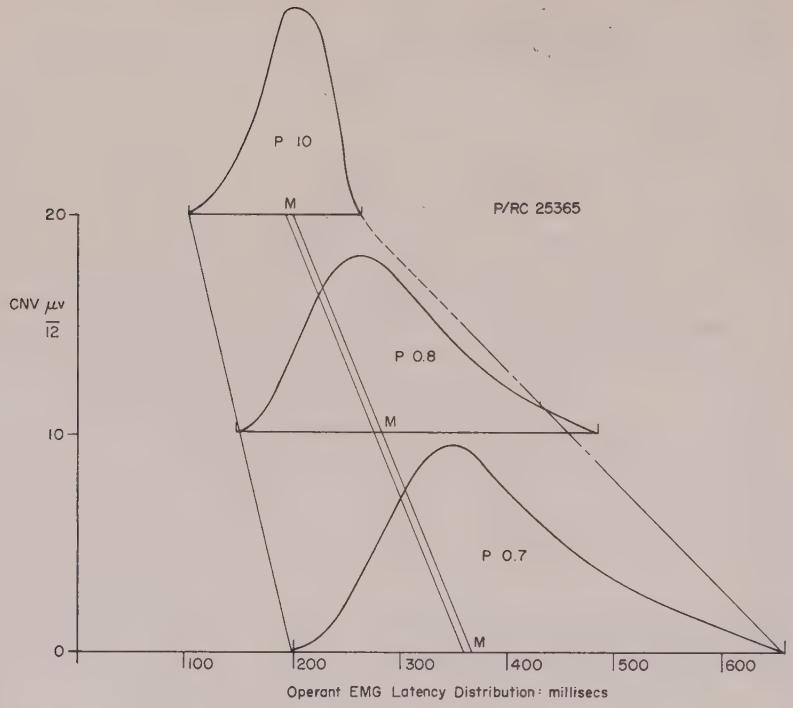


Fig. 16

shows a graph plotting the size of the CNV on the ordinate against the EMG latency distribution in milliseconds along the abscissa. At the top with a probability one, the CNV was just over 20 microvolts and there was a tight, neat distribution of response latencies with the mode at about 200 milliseconds. With the probability reduced to 0.8, the CNV is reduced to about 10 microvolts, the distribution of muscle activity is wider with a longer latency and so on further down to a probability 0.7 with a much wider distribution. If you join the modes of the reaction time EMG latency, you get something approximating a linear relationship between this and the size of the frontal CNV. So here we have an indication of what has happened in the brain—there is a priming process triggering the brain, insuring synchronous, economical responses of the motor system.

Figure 17 gives a very important aspect of this particular phenomenon—that it is not intensity sensitive, that it doesn't depend upon the brilliance

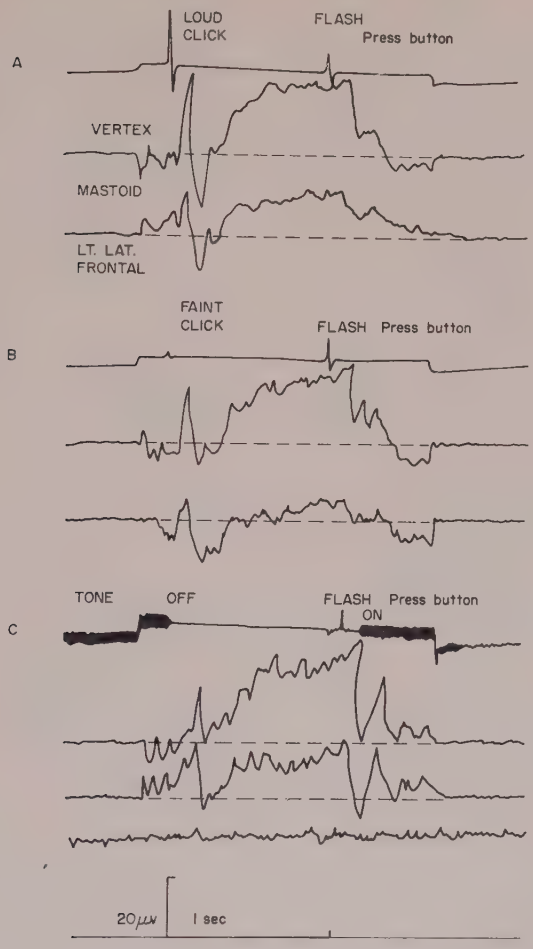


Fig. 17

of the light or the loudness of a sound or the degree of pain. The top trace shows the response to very loud clicks of about 110 decibels followed by flashes, to which the subject responded as before. The next one is the response to threshold clicks so faint that the subject was often unable to hear some of them; the response is smaller but, as reported in other papers at this meeting, not all that much smaller. The CNV is exactly the same because the *information* content in the faint conditional stimulus is the same as when it was loud, and the bottom trace is the reductio ad

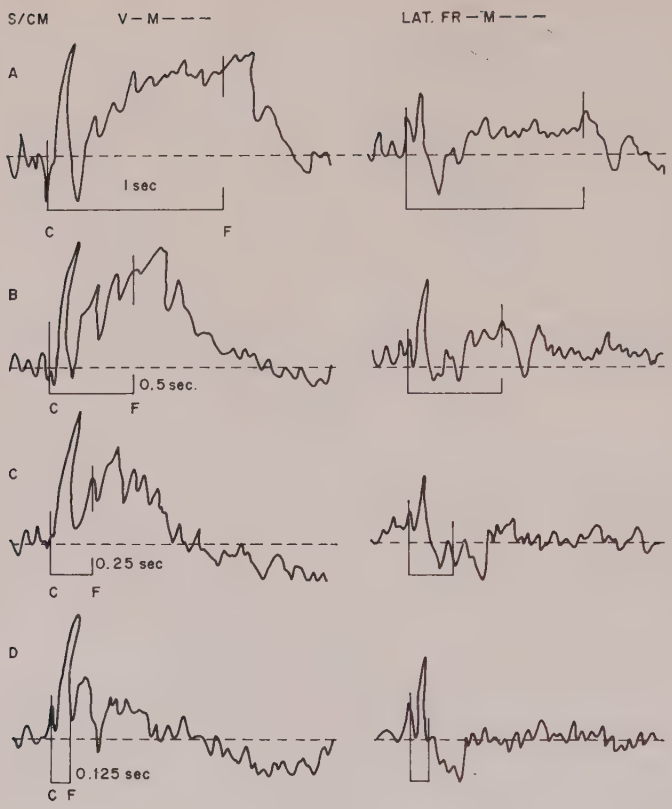


Fig. 18

absurdum where a continuous tone is interrupted as a stimulus and acts as a "negative" stimulus; the response is, of course, an "off response" which is longer delayed than the "on response", but the CNV is exactly the same. This, with many other experiments that we have done, shows that the CNV responds to the information in the conditional stimulus not in any way to its intensity or its modality.

In Figure 18 the relation of the CNV to the interval between stimuli is shown. In the top trace on the left, the duration is one second between the conditional and imperative stimulus, the next is 0.5 and then a quarter and an eighth of a second. If the interval between the conditional and imperative stimulus is less than about half a second the CNV can't

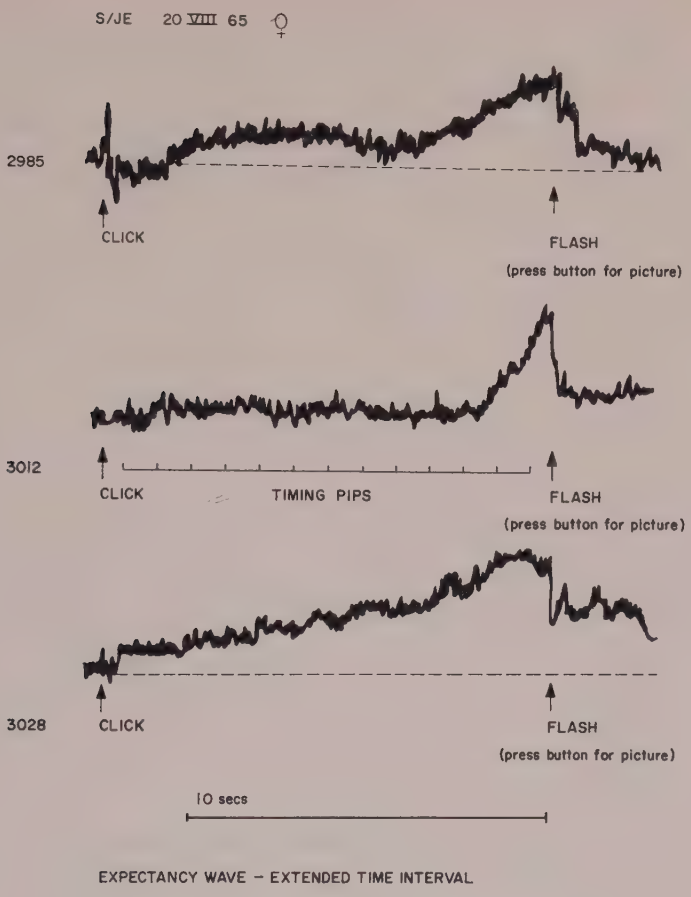


Fig. 19

develop and this relates quite well to psychological data on reaction times, that if the fore-period between a warning stimulus and the imperative stimulus is less than half a second, the reaction times are not very much shortened. If the fore-period is longer than about half to one second, then the reaction times are shorter.

Figure 19 shows the other extreme of extending this period. The interval between the conditional and warning stimuli is about 15 seconds—with a 16 second average. The top one is the average of the first trials and in this experiment the second stimulus was the opening of an elec-

tronic gate and if the subject pressed the button at that time, she could see an interesting picture. The pictures included Paris fashions with very short skirts and so on, in which she was interested, but she could see them only if she pressed the button at the right time. She was not told when the gate was opened; she had to find out. In these trials you can see the breaking up of the CNV into two waves—the first wave lasting about 5 seconds, followed by a sudden rise when it comes to the crucial period. The next set of trials were made easier by providing a countdown in the form of timing pips every second between the conditional stimulus and the opening of the gate. You can see the value of the countdown; the expectancy doesn't start to rise until about the 13th second, then it rises very abruptly and cuts off when subject performs her action.

The bottom trace is after the subject had been trained with the countdown, but the timing pips had been withdrawn, and she had to estimate the time herself. Now the expectancy rises gradually over the whole period of about 13 or 14 seconds with a slow potential rise in the brain terminating abruptly with the operant response. What is this long term process? I suggest that this may be an outward and visible sign of short term memory. This is the process whereby the brain stores information, such as that a warning click will be followed a certain time later by something of interest. This is the sort of memory you use when you dial a telephone number that you have just looked up in the book and having dialed it and got a busy tone or wrong number, as we often do in England, you find that in dialing you have forgotten the number. The ten digit number which you have taken the time to memorize, has been erased in the process of dialing—this is destructive write-out. I suggest that this potential change reflects the operation of a short term memory with a destructive write-out—again resembling a time-base waveform.

Figure 20 shows the effect when the subject is expected to do something more subtle than pressing a button. In the top trace we have the association of flashes and clicks, the subject doing nothing at all as in the very first figure of all. The second trace shows the response when all the subject has to do is to say "now" when the flashes occur; this doesn't stop the flash—he just says "now". But in the bottom trace, all he has to do is to *think* "now" and this has exactly the same effect; it is not necessary that the subject should perform any motor action at all. All that is necessary for the development of a CNV is that he should change his mind, in effect make a decision of some sort.

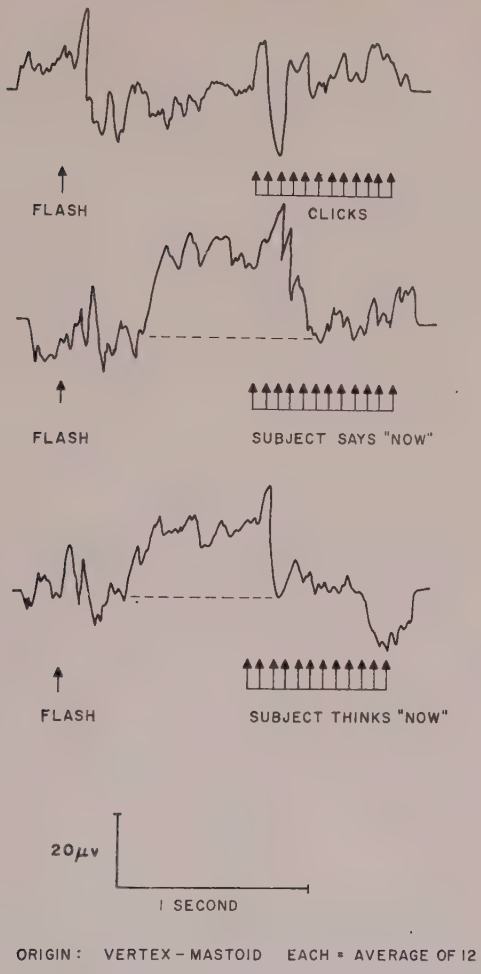


Fig. 20

Figure 21 shows the effect of distraction in a normal subject, the interspersion of tones when a click-flash association is being established. This reduces intramodality interference, the early components of the evoked response, but it does *not* reduce the CNV. On the other hand, in a patient with a chronic intractable anxiety neurosis, the effect of intramodality distraction is to abolish the expectancy wave as well as to reduce the initial response. This is what the patients say: "I can't bear distraction, I can't bear noise, things keep rushing in on me, everything is important."

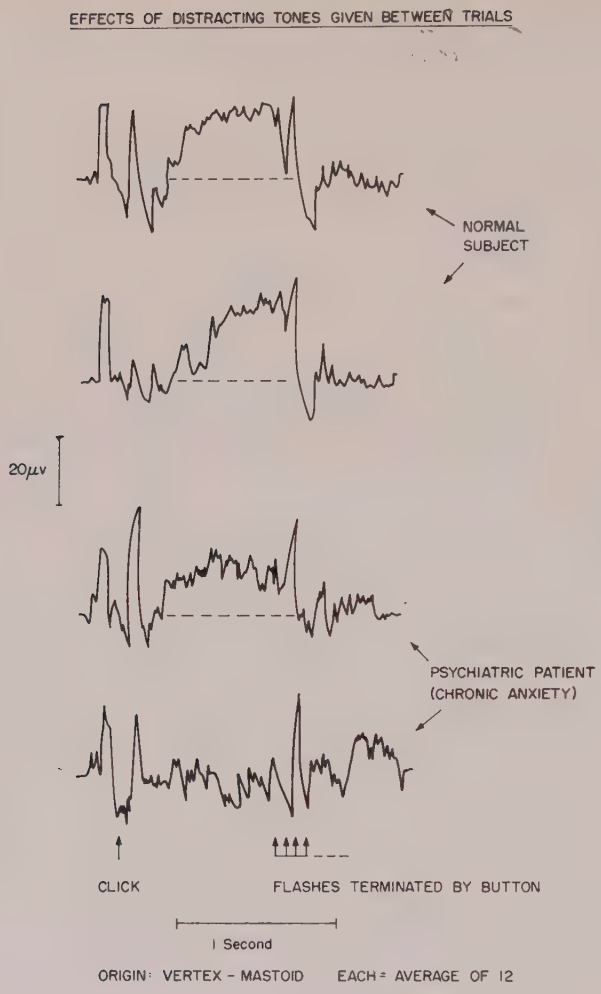


Fig. 21

And in the end, the effect of everything being important is, of course, that nothing becomes important and the subject becomes suicidal.

Figure 22 leads up to an almost science fictional aspect of this effect. Kornhuber in Germany, working in Jung's department in Freiburg, described a couple of years ago what he called a "Bereitschaftswelle", a readiness wave in the brain preceding voluntary actions. We have confirmed this and by playing tricks with our tape recorder and computer

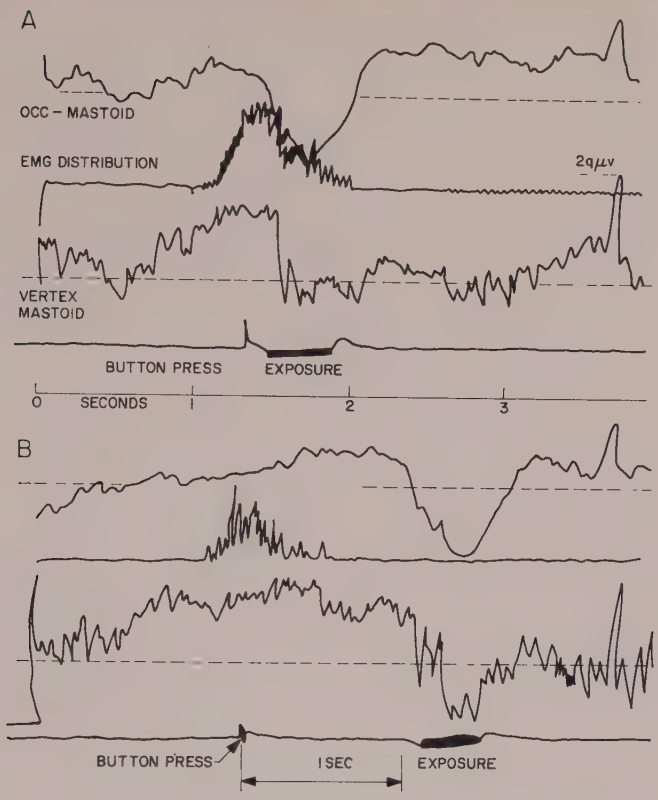


Fig. 22

we can take on-line averages of events in the brain which precede voluntary spontaneous actions. In the top trace the subject was told that he could press a button anytime he liked, of his own free will and when he pressed it he would see an interesting picture at once. For about one second before his action there is a rise in potential, a negative wave, with very much the same topography as the expectancy wave. What we suggest is that the impulse to press the button at any particular time is not a metaphorical, but a real impulse. The subject feels impelled to press the button at some particular time and not at other times. And for a second or two before his decision, the potential of his frontal cortex becomes negative and cuts off again quite sharply when his action is consumated. Now, in the bottom set of traces the exposure of the picture

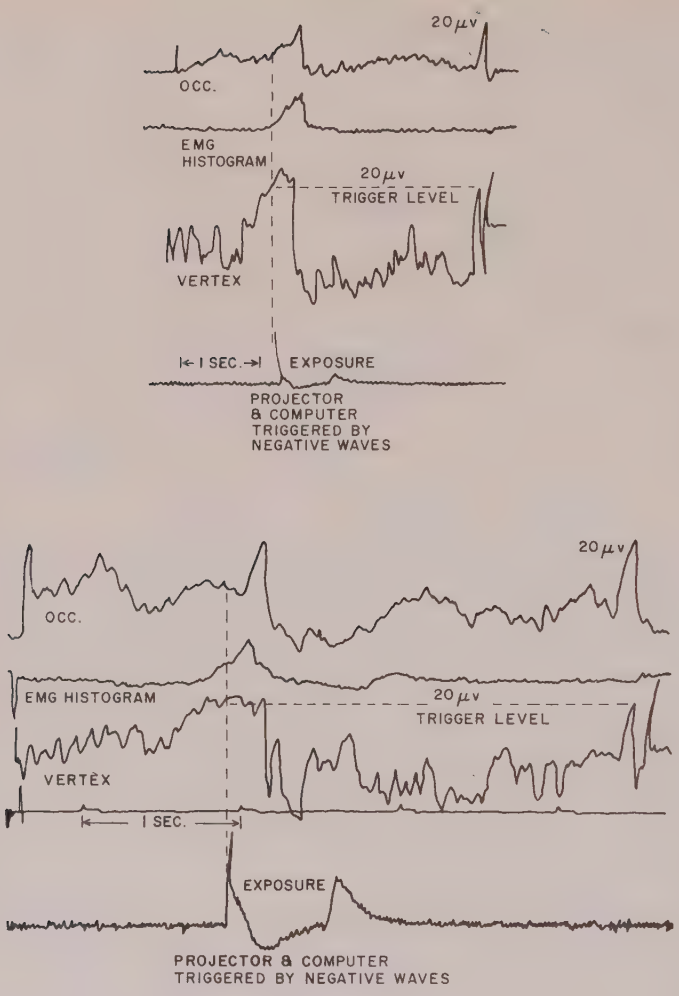


Fig. 23

was delayed by one second, but the same instructions were given: "press the button when you like and you'll see a picture". As before, the intention wave builds up for about a second before he actually presses the button, but now sustaining itself until the actual consumation of the action—until the picture appeared. So here we have the merging of the intention wave, the "Bereitschaftswelle", into the expectancy wave.

He presses the button after one or two seconds of potential rise, but nothing happens at once, and until the expected event does happen, his expectancy potential is sustained and is discharged only when the picture appears.

Figure 23 illustrates a development which we have started only recently. It is perhaps only a gimmick now, but I think it might have interesting applications. By suitable filtering and a bit of switchcraft you can make these intention and expectancy waves do something, for example, operate machinery. In the top trace, taken at a slow scan speed, the negative wave in the subject was made to initiate the computer and the exposure of the picture directly. The electromyogram is of the forearm and the subject was asked to move a finger if she had wanted to see the picture. The electromyogram response was after the expectancy wave. The procedure for getting this to work is to start off by letting the subject press a button as before to see the picture. Then, without telling them, we transfer the computer control from the button they have been pressing to the output of their own expectancy waves. Now the culmination of the expectancy wave produces a picture automatically whenever the subject "wants" to see it. To obtain regular results one has to learn to concentrate in a peculiar way on the specific experience one wants to have and not on concentration itself. This is a very odd feeling and when I had been through it I felt a sense of self-revelation, as when one masters a previously mysterious skill such as flying or hitting a ball where one wants it to go. It is as though my private thought and desire to have a new slide actually resulted in the movement of the slide itself. The lower trace shows the same thing taken at higher speed and you can see the computer trigger level set at 20 microvolts through the filters and the production of the picture is achieved by an act of the will. Here we are bypassing the effector pathways and allowing the brain to do something all by itself. Of course, you might get all sorts of chance electrical events triggering the picture, but in fact, after some practice, all the pictures that are projected are the ones which the subject wanted to see, but they appear a fraction of a second before the subject thought she wanted to see them.

The next few figures show the relation of the CNV to everyday actions as recorded by telemetry in collaboration with Dr. Storm van Leeuwen and Mr. Kamp from Utrecht in Holland. On Dr. van Leeuwen's head is strapped a little 8-channel radio telemeter set and my colleague is

Fig. 24

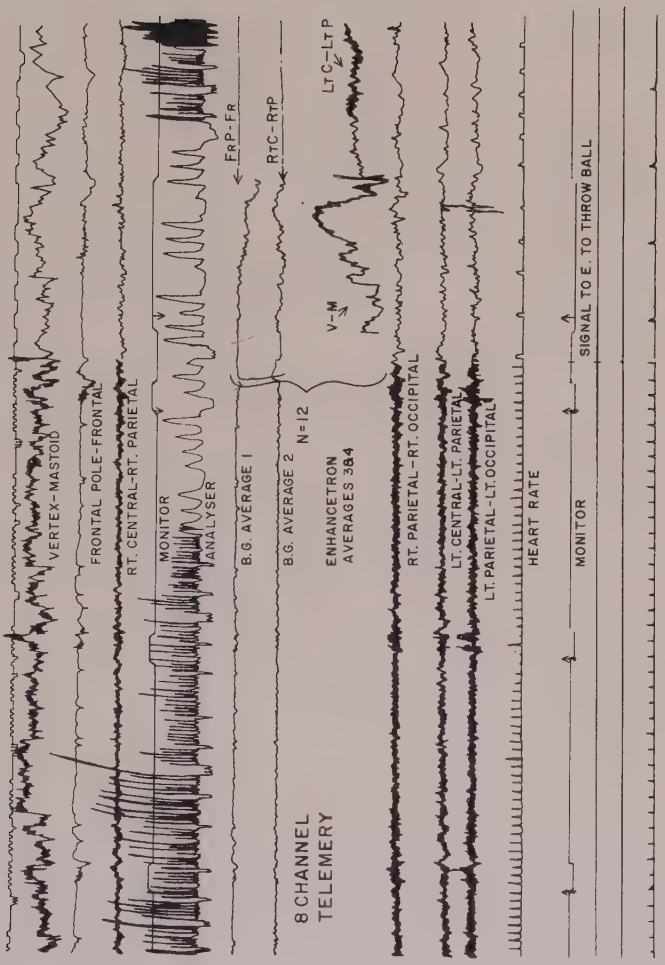

8 CHANNEL
TELEMERY
VERTEX-MASTOID
FRONTAL POLE-FRONTAL
RT. CENTRAL-RT. PARIETAL
MONITOR
ANALYSER
B.G. AVERAGE 1
B.G. AVERAGE 2
N=12
ENHANCETRON
AVERAGES 384
FRP-FR
RTC-RTP
LTC-LTP
V-M
RT. PARIETAL-RT. OCCIPITAL
LT. CENTRAL-LT. PARIETAL
LT. PARIETAL-LT. OCCIPITAL
HEART RATE
MONITOR
SIGNAL TO E. TO THROW BALL

Fig. 25

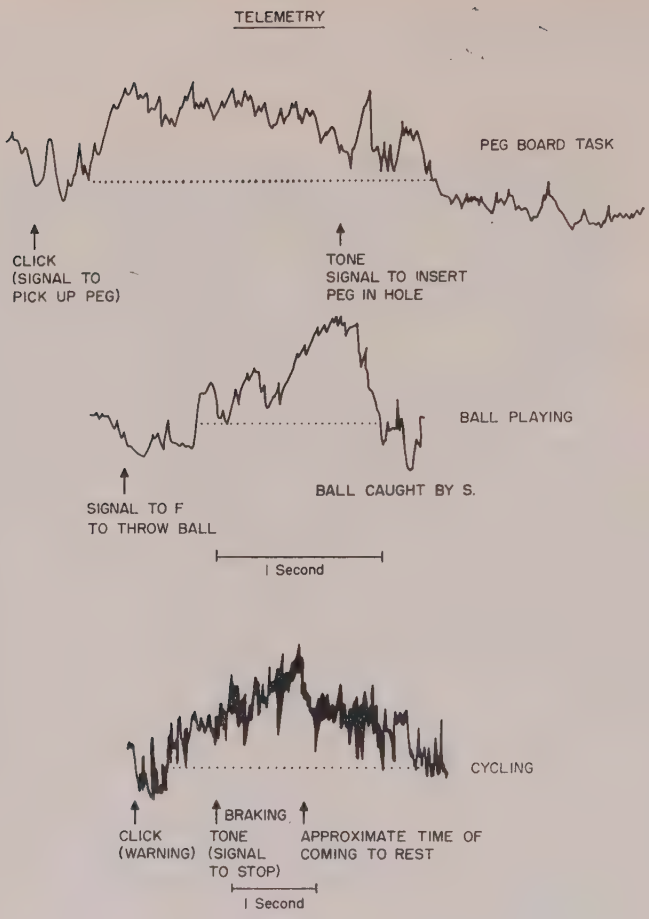


Fig. 26

throwing a ball to him following instructions obtained through a radio receiver (Fig. 24). Figure 25 shows the response obtained by telemetry of this 8-channel record with the build-up of expectancy as the ball was thrown. The signal to the experimenter to throw the ball is indicated and as the ball appeared in the air the subject decided to catch it; in this condition you get an expectancy wave exactly as in the laboratory. The purpose of this is to demonstrate that this is a real-life effect and not a laboratory artifact. Figures 26 and 27 show similar effects.

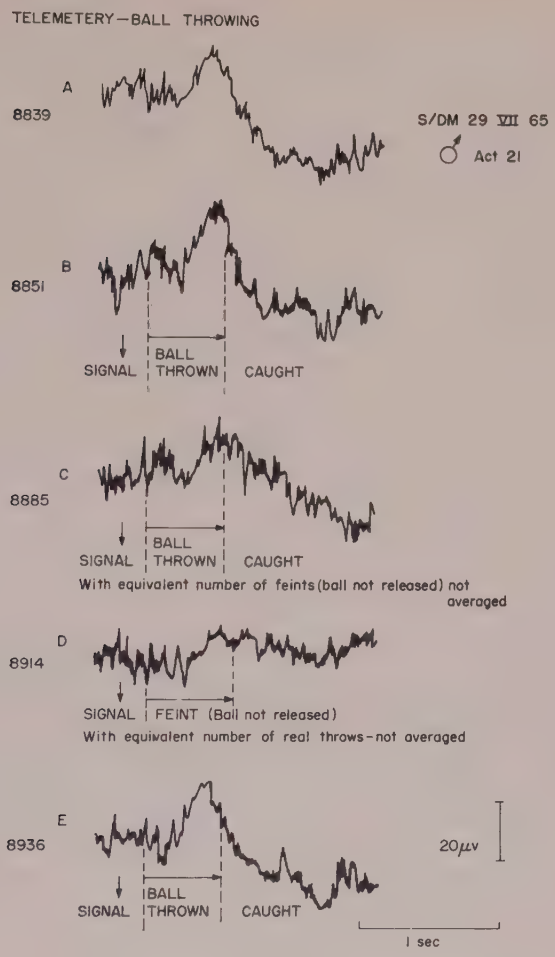


Fig. 27

In Figure 26 the telemetered response of a subject doing a pegboard task is shown. For every click received by radio, an expectancy wave appeared as the subject was prepared to do the task. The second trace is again ball throwing in the courtyard, and the third one is a child cycling. He was told he would get a click every now and then followed by a tone. that a click meant—prepare to stop, as for a red light on a crossing and the tone meant stop *now*, as at an intersection. Here again you see the expectancy wave building up and subsiding at the moment when the subject had completed his task, that is, when the bicycle came to rest.

Figure 27 shows the expectancy wave with a number of different situations. The middle three are interesting because the subject whom you saw in the photograph learned after a while, of course, to watch the experimenter's hand movement to catch the ball, rather than watch for the ball in the air. But then we asked the experimenter to feint occasionally. Instead of getting a click on his receiver he would get a tone. Then he would make a feint movement and not throw the ball. The subject at first began to develop an expectancy wave for the movement of the arm from the other experimenter, but after a few feints, the expectancy wave to the arm movement subsided and it appeared only when the ball was actually in the air. So again this shows that the real-life situation such as fooling someone while playing a game, is associated with exactly the same effects as in our laboratory situation.

This is only an outline of some of our observations just to establish the general relations of this effect in the brain, but let me now say one word about what we think it is: as far as we can tell from the studies with intracerebral electrodes, the CNV seems to be depolarization of apical dendrites in patches over the whole of the frontal cortex, possibly with participation of the neuroglia which is fashionable at the present time. Often the wave seems to spread from front to back at about 15 centimeters a second in normal people and we surmise that it controls the threshold of the pyramidal cells in the frontal and prefrontal region. We suggest that this is a challenge to brain modelers. Here is an analog change, a slow rise in voltage, like the time-base of an oscilloscope acting as a cortical primer and trigger, which seems to be an essential part of coherent, intelligent, human activity.

J. STAGGE

and A. SCHIEF

Institut für Schwingungsforschung

der Fraunhofer-Gesellschaft,

Tübingen, Federal Republic of Germany

The Detection of Signals in Electric Fish (Gnathonemus petersii) in the Presence of Noise

ABSTRACT

The transmitter of Gnathonemus petersii produces short pulses of varying pulse rate. The electric current field around the fish is similar to the field of a dipole. It is used for communication and position finding. The receptors of the fish perceive voltages fed into the aquarium from the outside and variations of their own transmitter field. The reception of these signals is disturbed by a superimposed electric noise field. Position finding is less affected by this noise field than communication; in position finding the fish can apply synchronous detection.

INTRODUCTION

Several investigators carried out investigations on the properties of the electric apparatus of weakly electric fish. The performance of this device has been found by training the animals or by electrophysiological studies in the receiving organs.[1, 2, 3] In Gnathonemus petersii, a weakly electric fish living in the rivers of Africa, quantitative measurements of the electric apparatus as a system for communication and active position finding were carried out.[4, 5] Starting from this work in Gnathonemus petersii, the susceptibility to jamming by Gaussian noise is investigated from the standpoint of communication theory.

TRANSMITTER

The transmitter of Gnathonemus petersii is placed in the tail of the animal (Fig. 1). It generates short pulses of a duration of about 200 μsec (Fig. 2); essential components of the spectrum are in the frequency range between 1 and 10 kcps. The pulse rate depends on the excitement of the animal; it varies between 0 and 50 pps. In this respect there is a significant difference compared with the results got with other kinds of electric

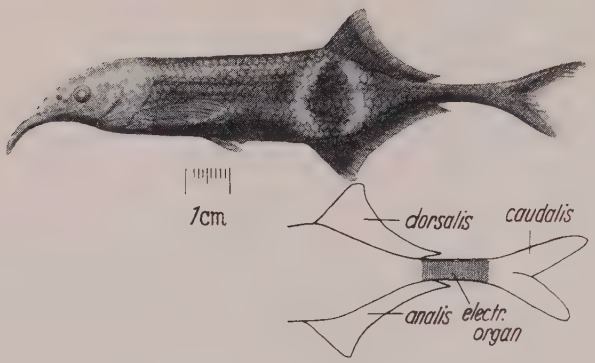


Fig. 1. Gnathonemus petersii. Position of the transmitter.

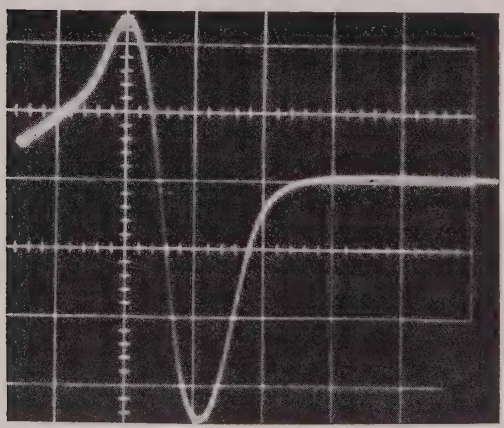


Fig. 2. Pulse of Gnathonemus petersii. Scale mark = 100 μsec.

fish, e.g. Gymnarchus niloticus; this fish delivers a highly constant frequency of about 300 pps which is practically not affected by external stimulation. The transmitter of Gnathonemus petersii produces an electric current field around the animal which is similar to the field of a dipole (Fig. 3). A voltage of about 6 V_{p-p} is measured between those

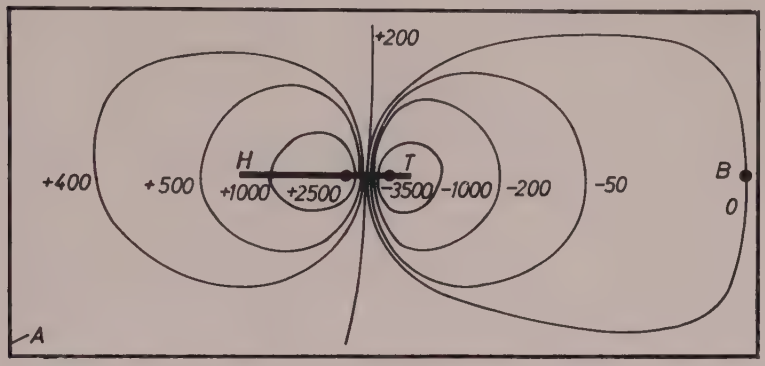


Fig. 3. Equipotential lines of the current field of Gnathonemus petersii. The numbers are the potential differences (mV_{p-p}) referred to the point *B*. Numbers are positive, when the pulses start with positive part.
A: Aquarium, H: Head, T: Tail of the fish.

points of the skin of the fish at which the electric current is leaving and re-entering the body. The transmitter is well adapted to the electric resistance of the surrounding water. The range of the electric current field of physiological significance is about 1 m; it results from this fact that the travelling time of the pulses due to the finite propagation velocity of the electric current field is negligible.

COMMUNICATION AND POSITION FINDING

The pulses generated in the transmitter are used by the fish in two different ways. Firstly they are a system for communication among different animals; this communication can consist in a very simple kind of marking the hunting ground of an animal. Secondly they are a tool for active position finding; the animal detects and locates obstacles in the surrounding water by observing the distortions of its own current field caused by differences in conductivity and dielectric constant. In both cases there must be receiving organs.

The method of investigating these problems makes use of the fact that Gnathonemus petersii changes its transmitter frequency remarkably when it detects an electric signal fed into the water or when an obstacle appears in its neighborhood. To fix the animal within an aquarium without impairing its mobility by a crude encroachment, a tube of colourless material or of a net is offered to it. A part of this tube is covered by a thin layer of transparent red colour that darkens the corresponding part of the interiour of the tube. When no other hiding place is within the aquarium, the fish will place itself preferably in the coloured part of the tube. In this way the electric phenomena around the animal are conveniently observed by small gold electrodes fixed in the side of the tube without mechanically disturbing the animal and without time-consuming training experiments.[5]

To measure the sensitivity of the fish to electric fields fed into the water from the outside, a variety of forms of voltages have been applied. The results described in this paper were gained with pulses very similar to those produced by the fish itself. They were periodically repeated at a rate of 30 pps. The pulses of the fish under observation are received by a pair of electrodes and fed to a frequency-to-voltage converter and a recorder. The stimulating pulses are suppressed by a gate at the input of the converter. Different sets of electrodes have been used to produce a stimulating electric field around the fish. Figure 4 shows a schematic view of the tube with several electrodes arranged in a row at the top. The distribution of the sensitivity to electric fields on the surface of the fish is measured by applying stimulating voltages to pairs of electrodes to generate locally limited dipole fields, and calculating the corresponding threshold strength of the electric field at the skin of the fish. The results for a typical specimen are given in Figure 5, curve a; the sensitivity is defined as the reciprocal of the threshold field strength (cm/mV_{p-p}). Curve b represents the distribution of the mormyromasts, the supposed receiving organs of Gnathonemus petersii. Both curves have a maximum at the front part of the fish. At this maximum the threshold field strength is 0.7 mV_{p-p}/cm. The value of the threshold field strength when the fish is exposed to a homogeneous field parallel to its axis is about 0.2 mV_{p-p}/cm.

To prove the reaction of the fish to changes of conductivity and dielectric constant all other kinds of stimulation, e.g. mechanical, optical, olfactory, must be eliminated. The changes of the electric properties are achieved by connecting a pair of electrodes (Fig. 4) to resistors or capa-

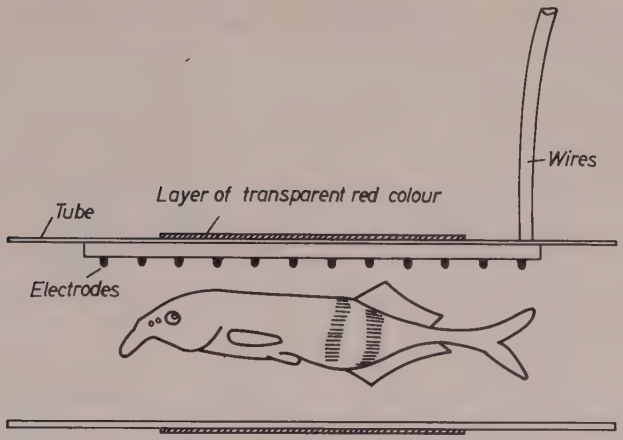


Fig. 4. Schematic view of a tube of colourless material or of a net.

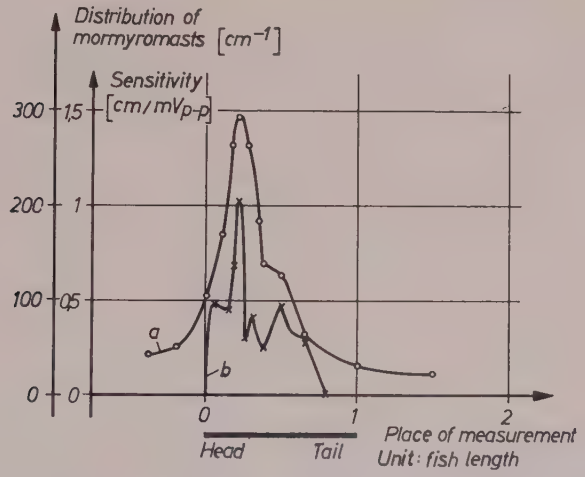


Fig. 5. Distribution of sensitivity to external electric fields, measured with pulses similar to Fig. 2 (curve a).
Distribution of mormyromasts along the body of the animal (mormyromasts per cm) (curve b).

citors by wires outside the aquarium. The transmitter frequency of the fish is observed as indicator. The sensitivity to this kind of stimulus was measured along the body of the animal. The reaction is very small when the electrodes are on an equipotential surface of its transmitter field

(Fig. 3). Again the highest sensitivity is in the front part of the fish. When inserting the resistors or capacitors, contact voltages must be carefully avoided by properly choosing contact materials; an inductor of high inductance and low resistance is permanently connected to the electrodes to suppress dc contact potentials between them without influencing the impedance in the frequency range of the pulses produced by the fish. When no obstacle is in the environment of the fish, its transmitter produces a primary electric field strength of about 100 mV_{p-p}/cm at the front part of the fish. The highest resistor between two electrodes still perceived by the fish changes this strength by an amount of 0.5 mV_{p-p}/cm, as measured in a model experiment by replacing the transmitter of the fish by an artificial dipole. The pair of electrodes loaded by the resistor can be replaced by an auxiliary dipole that causes a field strength at the front part of the fish corresponding to the change of its own sending field. This little change carries the information on the obstacle. The maximum sensitivity of the receiving organ with respect to obstacles thus corresponds to the sensitivity gained with external voltages.[5]

The reaction of the fish to these two kinds of stimulus is quite different. When external voltages are fed to the electrodes the fish reduces its transmitter frequency or stops transmitting for a moment; in this way the signal to be received is less masked by its own pulses. When obstacles are simulated to the fish by connecting resistors, its frequency increases to about 50 pps and thus provides a higher rate of information to the fish.

DETECTION OF SIGNALS IN THE PRESENCE OF NOISE

Applying the methods described, the influence of noise has been investigated. Communication and position finding are disturbed by an electric noise field added to the electric fields the fish is to perceive. There is another kind of disturbance that is of significance in the case of position finding. The electric field produced by the fish itself may be modulated by more or less random alterations of the electric properties of its environment, caused by surface waves, air bubbles or when the fish is swimming. This disturbance is superimposed to the signal in a nonlinear manner. The experiments described below are related to an additive superposition of Gaussian noise in the frequency range 1 to

10 kcps corresponding to the spectrum of the pulses of the fish. These experiments yield preliminary information on the influence of noise.

The fish is exposed to a stationary homogeneous noise field of known intensity. When it has got accustomed to this process an additional electric field or a simulated obstacle is produced as a signal by a pair of electrodes near to the front part of the fish. To a given noise intensity the smallest voltage and the highest resistor between the electrodes causing a reaction of the fish have been measured. By means of calibrating curves the corresponding threshold field strength and the corresponding threshold variation of the transmitter field strength at the front part of the fish have been studied. For one typical specimen the results are shown in Figure 6. The threshold field strength (measured in mV_{p-p}/cm) is depicted as a function of the noise field strength (measured in $m\,V_{rms}/cm$);

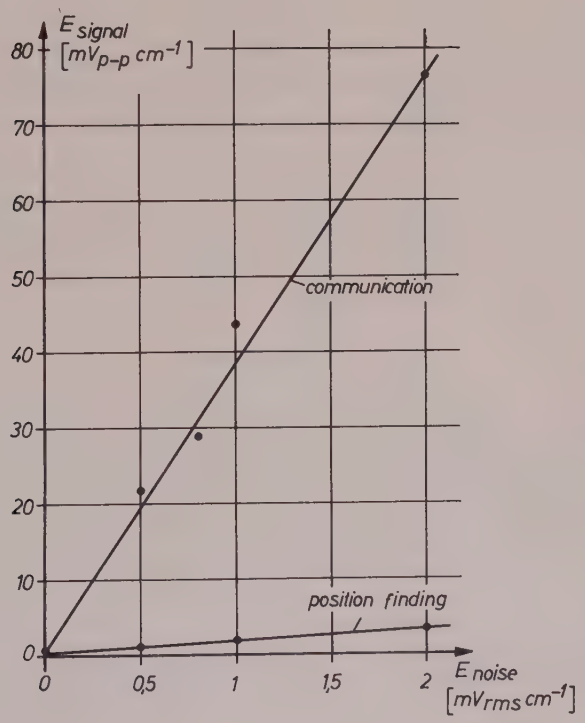


Fig. 6. Threshold field strength as function of noise field strength.

the upper curve holds for communication, the lower one for position finding. Starting from the same value with no noise the slope of the upper curve is about 20 times the slope of the lower curve. This means that position finding is remarkably less affected by additional Gaussian noise than communication.

DISCUSSION

When taking into account the conceptions of information theory these results do not surprise. In the case of position finding the fish can apply synchronous detection; it "knows" the time of arrival and the shape of the pulses transmitted by itself, as there is no significant travelling time. For this reason the receivers can be switched on only during the time the pulses are present; then only a small fraction of the noise is efficient, because the ratio of pulse length and pulse distance is very small (200 μsec/20 msec). On the other hand, for communication the animal has to keep its receivers sensitive all the time it is watching signals transmitted by other animals; thus all noise is efficient in communication.

There is another explanation of these results, making use of the pulse shape of the signals. It was stated that only small changes of the

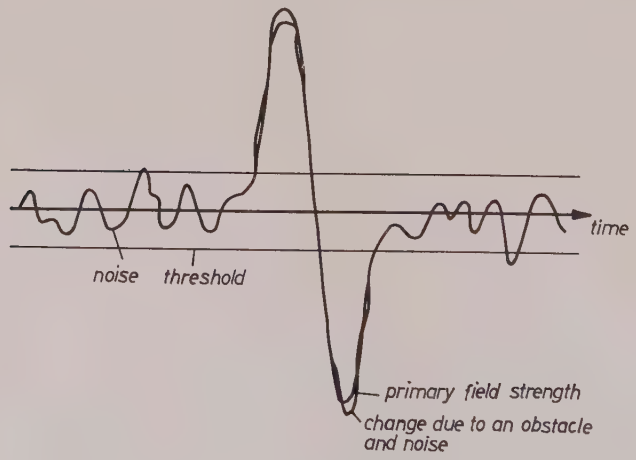


Fig. 7. Detection of obstacles in the presence of noise.

primary transmitter field strength at the skin of the fish carry information on the electric properties of its environment. When the reception is to be performed in the presence of noise the high informationsless primary field at the receivers is a means for a kind of detection very similar to the synchronous one. As shown in Figure 7 a threshold detector might be assumed that suppresses nearly all noise that occurs between the fish pulses. During the time the pulses are present approximately the same relations as in synchronous detection take place. If it is assumed that the threshold of the receiver is controlled by the noise level, there is also an explanation of the high sensitivity in communication when no noise is present. When the intensity of the primary field is high compared with the noise field strength a square law detector will operate in a similar sense.

The concept of a threshold organ is more satisfying as a model for the performance of a receiving organ in biological objects than that of a synchronous detector controlled by the voltage of the transmitter.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Prof. Dr. Ing. habil. H. Tischner for his encouragement and valuable discussions. This work was supported by the Ministry of Defense of the F. R. of Germany.

REFERENCES

1. Lissmann, H. W., and Machin, K. E. The mechanism of object location in Gymnarchus niloticus and similar fish. *J. Experim. Biol.* **35** (1958), 451–486.
2. Fessard, A., and Szabo, T. Mise en évidence d'un recepteur sensible a l'electricité dans la peau des Mormyres. *C. R. Acad. Sc. Paris*, **253** (1961), 1859–1860.
3. Hagiwara, S., Kusano, K., and Negeshi, K. Physiological properties of electroreceptors of some Gymnotids. *J. Neurophysiol.* **25** (1962), 430–449.
4. Harder, W., Schief, A., and Uhlemann, H. Zur Funktion des elektrischen Organs von Gnathonemus petersii. *Z. für vergl. Physiol.* **48** (1964), 302–331.
5. Harder, W., Schief, A., and Uhlemann, H. Zur Empfindlichkeit des schwachelektrischen Fisches Gnathonemus petersii gegenüber elektrischen Feldern. Z. für Vergl. Physiol. **54** (1967), 89–108.

GEORGE BIERNSON

Applied Research Laboratory, Sylvania Electronic Systems

A Division of Sylvania Electric Products Inc.

Waltham, Massachusetts

Evaluation of Physiological Evidence for Trichromatic Theory

INTRODUCTION

The popular theory of color vision is that the eye uses the same three-color principle employed in color photography and color television. It is assumed that the retina has three types of receptors with different spectral responses that effectively enable it to sense three color-separation "photographs" of the scene. This is called the Trichromatic theory and was proposed by Thomas Young in 1801. Recent spectrophotometer measurements on the human retina[1, 2] have detected what appear to be three types of cones with different spectral responses, and it has been assumed that this has verified the Trichromatic theory.

From a bionics point of view, a verification of the Trichromatic theory raises far more questions than it answers. Psychological evidence shows that color vision is a remarkably effective process, vastly superior in many respects to man-made color systems in terms of operating range (the eye detects color over an intensity range of one million), accuracy of spectral discrimination (the eye can distinguish 10 million different shades of color), strong insensitivity to changes of illuminant spectrum, and many other characteristics. We would like to understand how the eye operates so we can apply the same principles in equipment design. However, if we assume that color vision uses the same three-color principle employed in present color equipment, we are unable to explain its high performance.

Mathematically, the crucial assumption of the Trichromatic theory is that the retina takes three weighted spectral averages of the light, and so detects only crude information concerning the light spectra. Far better performance is theoretically possible if the retina sensed much more of the spectral information, and employed this "full" spectral information

in the data processing associated with adaptation, contrast enhancement, etc., before reducing it to three-color information. Therefore, we should carefully evaluate the evidence for the Trichromatic theory to determine whether the theory has truly been verified, or, alternatively, whether the question is still open and we have the freedom to consider, as a possible hypothesis for exploring the mystery of color vision, that the retina may detect much more than three-color spectral information. This paper will evaluate the physiological evidence for the Trichromatic theory. The major psychological evidence was considered in reference 3.

PHYSIOLOGICAL EVIDENCE

The primary physiological evidence supporting the Trichromatic theory consists of spectrophotometer measurements on individual human retinal cones by Marks, Dobelle, and MacNichol[1] and by Brown and Wald.[2] The spectral transmission of a cone is measured before and after bleaching its photopigment with light, and these spectra, which differ by a few per cent, are subtracted to form the difference spectrum. They found that the difference spectra fall into three classes, with response peaks in the blue, green, and yellow regions of the spectrum. Although there is a good deal of variation from cone to cone of the difference spectra within each class, the classes are clearly distinguishable. They have assumed that this proves that the retina has three types of cones containing three different photopigments.

These experiments were performed on dead retinas in which the photopigment does not regenerate. Consequently, only one measurement can be made on a single cone. Also, only one cone in a given region of the retina can be measured, because the stray light bleaches the neighboring cones. Hence the experiment provides no indication of the retinal organization of the assumed three cone types.

A strong physiological argument against the three-cone hypothesis is that no three-way classification of cones has yet been observed by electron or optical microscopes, either in terms of cone structure or in terms of the neurological interconnections among the cones. To explain this, it is postulated that the difference among cones occurs on a molecular, submicroscopic level. However this postulate has the serious physiological deficiency that it is difficult to explain how the cones could have evolved

into a three-way organization without exhibiting any observable structure associated with the embryological development.

Although it has been assumed that the three classes of cone spectral response are the result of three different photopigments, only one cone-photopigment, iodopsin, has been detected in solution, and then only in the bird retina.[4, 5] Therefore, one could assume a single cone-photopigment, and that the spectral response differences are produced by the electromagnetic properties of the cones rather than the chemical properties.

Figure 1 is a simplified sketch showing the major electromagnetic features of a human retinal cone. Light impinges on the broad end and

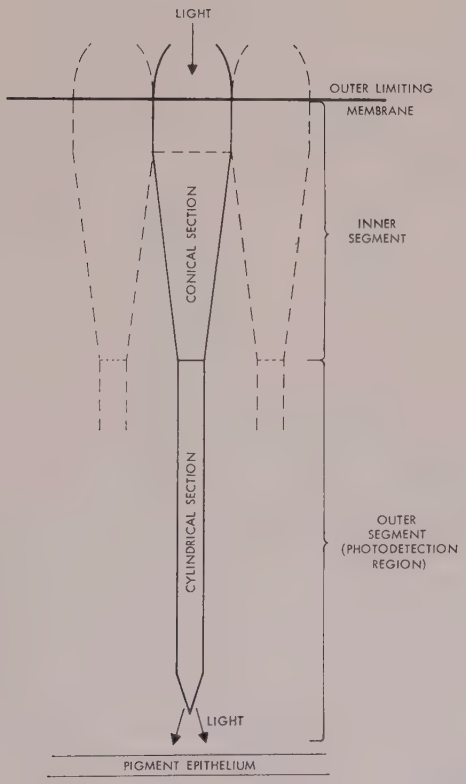


Fig. 1. Sketch of Human Retinal Cone Illustrating Major Electromagnetic Features.

is concentrated by the tapered conical section into the narrow cylindrical outer segment where photodetection takes place. The unabsorbed light radiates out the tip of the cone and is absorbed in the black layer called the pigment epithelium. Since the outer segment is only about two wavelengths in diameter, it acts as a dielectric waveguide, and light can propagate down it only as a summation of a few waveguide modes.

Enoch[6, 7, 8] has found from microscopic observations that the light radiating from the tips of the receptors exhibits mode patterns. By cataloguing these patterns he identified at least nine of the twelve modes having the lowest cutoff frequencies. These are as follows in order of increasing cutoff frequency: (HE_{11}), (TM_{01}, TE_{01}, HE_{21}), (EH_{11}, HE_{31}), (HE_{12}), (EH_{21}, HE_{41}), (TM_{02}, TE_{02}, HE_{22}). The modes combined within parentheses have almost the same mode pattern and cutoff frequency, and so it is difficult to distinguish one from another. Enoch observed mode patterns for each of the six groups of modes, and also more complex patterns corresponding to combinations of modes. From the complex patterns he established the presence of at least two modes in the second, third, and sixth groups. It is known that the TE_{01} and TE_{02} modes are very difficult to excite, and so it is reasonable to assume that they were not present. Hence we conclude that Enoch probably observed the following nine or ten modes: (HE_{11}), (TM_{01}, HE_{21}), (EH_{11}, HE_{31}), (HE_{12}), (EH_{21} and/or HE_{41}), (TM_{02}, HE_{22}). He was not able to tell from his experiment which modes were from rods and which were from cones.

Enoch found that the mode patterns for individual receptors were remarkably stable with changes of wavelength. As the wavelength of the light was varied across the visible range, Enoch[18] reported that "In a large number of cases the modal pattern did not seem to vary at all." Thus Enoch's observations have shown that the individual receptors are accentuating, or locking onto, particular modes, or in some cases combinations of modes.

Why do the cones lock onto different modes? If we assume it is because the cones are intrinsically different, we are forced to conclude that there are at least nine or ten different types of cones (one for each mode), which is difficult to accept. Besides, the spatial arrangements of the mode patterns observed across the retina are too chaotic to indicate that any reasonable basis for color discrimination could be derived from a comparison of the signals from cones propagating different patterns.

Therefore, we are forced to conclude that there must be a process within the retina which causes identical cones to lock onto different modes. If there is such a process, we would also expect it would cause identical cones to exhibit different spectral responses. This could well be the reason that three classes of spectral response have been found in the spectrophotometer measurements on individual cones.

An explanation of what might cause a receptor to lock onto a particular mode pattern can be found by considering what happens as the photopigment molecules in the receptor are bleached. The uneven light distribution in the receptor produced by the waveguide modes causes the photopigment to be bleached unevenly in the same spatial pattern as the energy distribution of the modes. Since bleaching changes the spectral absorption of the photopigment molecule, it alters its electromagnetic properties. Therefore, photopigment bleaching produces an electromagnetic inhomogeneity within the receptor having the spatial distribution of the mode pattern. This electromagnetic inhomogeneity should alter, at least to some extent, the excitation of the modes. Hence it very well could cause the receptor to excite more strongly those modes that correspond to the bleaching pattern, and thereby could make the receptor lock onto a particular mode pattern as the photopigment is bleached.

Thus we are led to suspect that the three classes of spectral response obtained in the spectrophotometer experiments may be caused by identical cones locking onto different mode patterns. However, if this is true, why are there only three spectral responses. whereas Enoch has observed nine or ten different modes? Let us see why this might be so.

Enoch[8] reported that the mode patterns for the last two groups of modes were very rarely seen. Hence, it is reasonable to neglect them when we consider the spectrophotometer experiments, because cones accentuating these modes would very likely not be detected by the limited sampling procedure used in those experiments, or if detected would probably be rejected as an already bleached cone or other anomaly. This leaves us only with the following six modes to consider, which we will call the frequent modes: (HE_{11}), (TM_{01}, HE_{21}), (EH_{11}, HE_{31}), (HE_{12}). Because the modes in each group have almost identical mode patterns at all wavelengths (although the field vectors are different) we would expect them to exhibit similar spectral responses. Therefore, it is reasonable to expect no more than four classes of spectral responses in the spectrophotometer experiments.

To calculate the spectral response of a mode is a very difficult problem, even if one assumes a homogeneous electromagnetic cone structure. Because of uneven photopigment bleaching, the cone is certainly not homogeneous, and this makes the problem even more complex. Nevertheless, we can obtain some important constraints on the mode spectral responses quite simply by considering the efficiencies of the modes. As a mode propagates down the cylindrical outer segment of a receptor (which is where photodetection takes place), part of the energy propagates outside the receptor. We define the efficiency of a mode as the percentage of the total mode power that propagates inside the receptor. The solid curves of Figure 2 are the spectral plots of mode efficiencies, over the

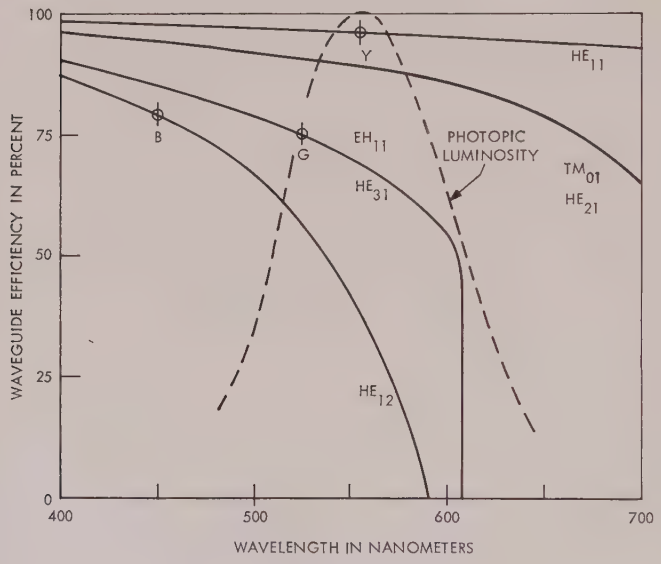


Fig. 2. Waveguide Mode Efficiencies for 6 Frequent Modes in Retinal Receptors

visible wavelength band, for the six frequent modes.[9] Since there are four mode groups, there are four mode-efficiency plots for these six modes.

It is known that the conical section of the cone couples the light energy efficiently into the outer segment. Therefore, we conclude that a mode

cannot be excited strongly in the outer segment unless its mode efficiency is reasonably high, say 50 per cent. Otherwise, a large proportion of the light power would be lost outside the receptor, and the coupling action of the conical section would be inefficient. This allows us to conclude that the (EH_{11}, HE_{31}) modes cannot be excited strongly at wavelengths greater than 600 nm (nanometers), and the HE_{12} mode cannot be excited strongly above 540 nm. The HE_{11}, TM_{01} and HE_{21} modes could all be excited strongly throughout the visible wavelength range.

Since the spectral absorption of the rod photopigment, rhodopsin matches the luminosity curve for scotopic (dark-adapted) vision, it is reasonable to assume that, if there is only a single cone-photopigment, its spectral absorption should approximately match the luminosity curve for photopic (daylight) vision. Besides, the one cone-photopigment detected in solution, iodopsin, has a difference spectrum that approximates the photopic luminosity curve.[4, 5] The dashed curve in Figure 2 is the photopic luminosity curve, and we will assume that this represents the spectral absorption of the cone-photopigment.

In the spectrophotometer measurements on individual cones by Brown and Wald,[2] the spectra measured on the "blue," "green" and "yellow" cones have response peaks at 450 nm, 525 nm, and 555 nm. These wavelength values are indicated in Figure 2 by three circles labeled (B), (G) and (Y), respectively.

Let us estimate the spectral absorption curves for a cone that has locked onto the EH_{11} or the HE_{31} mode. The mode efficiency plot shows that at wavelengths greater than 610 nm there is no power carried in the mode. The mode should begin to carry significant power at 50 per cent efficiency (600 nm) and its power should increase rapidly at shorter wavelengths. However, the spectral response of the light power absorbed by the cone is the product of the mode response multiplied by the response of the photopigment (the photopic luminosity curve). Since the photopigment response drops rapidly with decreasing wavelength below 550 nm, its effect predominates at short wavelengths. Hence the spectral response of the cone must rise to a peak and then fall with decreasing wavelength, regardless of the actual shape of the mode spectral response. As shown by the (G) circle, the "green" cones measured by Brown and Wald had response peaks at the point of 75 percent efficiency for the (EH_{11}, HE_{31}) modes. This is about where we would expect a peak for a cone that has locked onto one of these modes. Therefore, we are led to

suspect that a "green" cone is a cone that has locked onto either the EH_{11} or HE_{31} mode.

It should be noted that we cannot tell from the mode efficiency plot the actual shape of the mode response. For example, the spectral response of the EH_{11} mode might be monotonically increasing with decreasing wavelength, or it might rise to a peak and then fall. However, the cone response would be peaked in either case because the photopigment response dominates at short wavelengths.

Similar reasoning suggests that a "blue" cone is actually a cone that has locked onto the HE_{12} mode. The mode efficiency curve for the HE_{12} mode is zero above 590 nm, but it rises so slowly with decreasing wavelength it does not reach 50 per cent efficiency until 540 nm. Hence, a cone that locks onto the HE_{12} mode should have small response at wavelengths greater than 540 nm; and as wavelength is decreased below 540 nm the cone response should rise to a peak, and then it should fall because of the dropping response curve of the photopigment. The "blue" cones have a spectral response peak at the wavelength of 85 per cent efficiency for the HE_{12} mode, which is about where one would expect the response to peak for a cone that has locked onto the HE_{12} mode.

Since the mode efficiencies of the HE_{11}, TM_{01} and HE_{21} modes are high throughout the visible range, one would expect the spectral responses of these modes to vary only slowly with wavelength. Hence, the spectral response of a cone that locks onto one of these modes should be determined primarily by the response of the cone-photopigment, assumed to be the dashed photopic luminosity curve. As shown by the (Y) circle, the "yellow" cones have a response peak at the same wavelength as the peak of the photopigment response. This suggests that a "yellow" cone is actually a cone that has locked onto the HE_{11}, TM_{01}, or HE_{21} mode.

One would expect a significant difference in the spectral response of a cone that has locked onto the HE_{11} mode from one that has locked onto the TM_{01} or HE_{21} mode, because the mode efficiency plots are quite different. Nevertheless, one would not expect these differences to be any larger than the variation in spectral response of "yellow" cones that has actually been measured in the spectrophotometer experiments. As was stated previously there are rather large variations of measured spectral response within the "blue," "green" and "yellow" classifications.

Another effect that would help to produce three classes of cone spectral response in the spectrophotometer experiments is that the HE_{11}, EH_{11},

and HE_{12} modes are much easier to excite than any of the other modes that Enoch observed. If the receptor has axial symmetry, these are the only modes that can be excited, and their field patterns much more closely match that of the incident light than do those of the other modes. Therefore we will call these the dominant modes, because under normal conditions we would expect them to predominate.[10]

Snyder[11, 12, 13] has shown that, if a cone or rod has axial symmetry, the HE_{11} mode should be strongly excited when a light wave impinges on its end, and the HE_{12} mode should be weakly excited. In a cone, mode-coupling occurs as the light propagates down the tapered conical section, which couples energy from the HE_{11} mode into the EH_{11} and HE_{12} modes. Hence, in the outer segment of a cone all three dominant modes can be strongly excited. However, in a rod there is no mode coupling, and so only the HE_{11} mode should be strongly excited in its outer segment.

Thus, we would expect that the HE_{11}, EH_{11}, and HE_{12} modes would usually be accentuated in the cones, giving yellow, green, and blue responses. However, a rod should generally accentuate only the HE_{11} mode, and so should exhibit the response of its photopigment, rhodopsin, which is what occurs. In a spectrophotometer measurement, a rod does not exhibit appreciable mode effects; and this has caused physiologists (who have not understood the mode-coupling properties of the cone's tapered section) to assume erroneously that it is legitimate to ignore mode effects in the spectrophotometer measurements made on cones.

Since a spectrophotometer experiment measures the difference spectrum of the receptor (the change in spectral transmission as the receptor is bleached), it should be compared with the change in receptor mode pattern during bleaching. To make this comparison precisely is extremely difficult because the cone is not completely locked onto a single mode as it is being bleached. On the other hand, the preceding discussion has been able to show with simple reasoning that there are strong correlations between the properties of the waveguide modes and the difference spectra measured on the cones. Although these correlations do not provide conclusive evidence, they do show (1) that the spectrophotometer experiments have definitely not proven the existence of three types of cones or three cone-photopigments, and (2) that the electromagnetic properties of the receptors should be subject to much more intensive research.

CONCLUSIONS

The physiological evidence assumed to prove that the retina has three types of cones containing different photopigments is also consistent with waveguide mode effects within a single type of cone containing a single photopigment. Therefore the physiological evidence has not established the validity of the Trichromatic theory.

The mode patterns within the cones contain considerable spectral information, and so could theoretically provide a color discrimination process far superior to a three-color system. A means by which the cones might use this mode pattern information is described in references 14 and 15. Much more research is needed to explore the relationship between mode patterns in the retinal receptors and color discrimination.

ACKNOWLEDGEMENT

This research was sponsored by the United States Air Force and the independent research program of Sylvania Electronic Systems.

REFERENCES

1. Marks, W. B., Dobelle, W. H., and MacNichol, E. F. Jr. Visual pigments in single primate cones, *Science* **143**, 1181–1183; 13 March 1964.
2. Brown, P. K., and Wald, G. Visual pigments in single rods and cones of the human retina, *Science* **144**, 45–52; 3 April 1964.
3. Biernson, G. A feedback-control model of human vision, *Proc. IEEE* (Letters), **54**, 1226–1229; Sept. 1966.
4. Wald, G. Retinal chemistry and the physiology of vision, *Visual Problems of Colour* (Her Majesty's Stationery Office, London, 1958) **1**, 9–61, (see page 18).
5. Dartnall, H. J. A. Identity and distribution of visual pigments, "The Eye, vol. 2: The Visual Process," H. Davson, Editor (Academic Press, New York, 1962) chapter 18, pp. 367–426 (see pp. 396–397).
6. Enoch, J. M. Nature of transmission of energy in retinal receptors, *J. Opt. Soc. Am.*, **51**, 1122–1126; Oct. 1961.
7. Enoch, J. M. Optical properties of the retinal receptors, *J. Opt. Soc. Am.*, **53**, 71–85; January 1963.

8. Enoch, J. M. Visualization of waveguide modes in retinal receptors, *Am. J. Ophthalmology*, part 2, **5**, 1107–1118; May 1961.
9. Biernson, G., and Kinsley, D. J. Generalized plots of mode patterns in a cylindrical dielectric waveguide applied to retinal cones, *IEEE Trans. on Microwave Theory and Techniques*, MTT-**13**, 345–356, May 1965.
10. Biernson, G., and Snyder, A. W. Electromagnetic effects in the cones of the human retina, *IEE Electronics Lett.* (London), **1**, 90–91; June 1965.
11. Snyder, A. W. Excitation of surface modes along a semi-infinite dielectric cylinder, *IEE Electronics Lett.* (London), **1**, 208–209; Sept. 1965.
12. Snyder, A. W. Surface waveguide modes along a semi-infinite dielectric fiber excited by a plane wave, *J. Opt. Soc. Am.* **56**, 601–606; May 1966.
13. Snyder, A. W. Excitation of waveguide modes in retinal receptors, *J. Opt. Soc. Am.* **56**, 706–706; May 1966.
14. Biernson G., and Snyder, A. W. A theoretical model for color vision, AMRL-TR-65-193, Wright-Patterson Air Force Base, Ohio; Dec. 1965.
15. Biernson, G. A feedback-control model of human vision, *Proc. IEEE*, **54**, 858–872; June 1966.

SECTION III

E. R. CAIANIELLO

Gruppo di Cibernetica del CCNR,

Istituto di Fisica Teorica—Universita di Napoli

Naples, Italy

A Study of Neural Networks and Reverberations†

ABSTRACT

The simulations of neural networks leads naturally to the study of functions and principles which only in part fall within the scope of extant automata theory. Systems of decision equations must be studied with a view especially to obtain practical means for the prevision and computation of diffuse reverberations of wanted general characteristics and for the exclusion of all others. This amounts to deriving sets of constraints on the allowed variability of the couplings among elements in learning processes, failing which the behavior of the simulator would become uncontrollable for practical purposes. A simple mathematical treatment is presented, which exploits the extreme non-linearity of Heaviside functions to re-introduce matrix algebra in the context, permitting to study in a straightforward manner the wanted necessary or sufficient conditions.

INTRODUCTION

In a previous article [1] we outlined a theory of thought processes and thinking machines, which stemmed from a schematization of anatomical and physiological evidence. Forgoing any further reference to biology, we develop here the mathematical theory of the "thinking device" which is qualitatively described in [1]. From now on, we shall refer to it as

† The research reported in this document has been sponsored in part by the Air Force Office of Scientific Research under Contract No. AF EOAR 65-44 through the European Office of Aerospace Research, OAR, United States Air Force.

"Educanda" (name in our Laboratory for some hardware). Our discussion will touch only the most relevant points; a complete account will be given elsewhere. Educanda (ref. [1], [2]) is based on three fundamental principles:

a) Its elements, which we call "neurons", perform binary decisions; their instantaneous behaviour is described by means of *Neuronic Equations* (N.E.) which we may also call *Decision Equations* (D.E.);

b) The memory processes take place according to *Mnemonic Equations* (M.E.) (which we may also call *Evolution Equations*) expressing the changes of the coupling coefficients of the N.E. in time; the N.E. need not at all have the very special form hypothesized in [1] for a model which aimed at some respondence to biological reality;

c) The *Adiabatic Learning Hypothesis* (A.L.H.), which serves to decouple Educanda's instantaneous behaviour from its much slower learning processes. By virtue of the A.L.H. we can consider all coupling coefficients as constant for "short" durations of time.

We remind from [1] that our model makes essential use of *reverberations*, and that any neuron behaves at the same time as input, output and nodal element in the network of connections wherein the memory lies. Reverberations are instrumental in changing the coupling coefficients and giving thereby new logical possibilities to the machine. These changes will depend upon the "history" of the machine, in such a way that learning ensues spontaneously, as consequence solely of the mnemonic laws, without any interventions from the outside.

We emphasize that no probabilistic elements are considered at this stage in our model. We reserve their introduction for a future improvement of it, with the role of improving performances which may be marred by failures or imperfections, as is certainly the case with biological organisms; the machine considered here is an ideal one, free from any such flaws.

After excitation, Educanda's activity will consist of transients and reverberations, the period of which is not, in general, known *a priori*. This lack of information would bar definitely the way to any concrete non-probabilistic use of the machine: too long reverberations would be useless, nor could we read out the output without knowing exactly when and how to do it.

Our main mathematical problem is therefore to develop a model of Educanda such that only reverberations with known periods can take

place, no other restrictions being imposed on them; this is the specific aim of the present research. It is necessary for this:

1) To give the N.E. a suitable form, amenable to algebraic treatment; this task was already accomplished in [2] with the introduction of a formalism quite unrelated to the standard ones of Automata Theory.

2) To determine the constraints to be satisfied at all times by the parameters of the machine (coupling coefficients, thresholds, etc.), as well as to specify which controlling systems are to be added to it, in order that only reverberations of preassigned period be allowed whatever learning processes may have taken place.

The second point requires, again, only the study of the N.E.; we show here, without going into inessential details, how this problem, the basic one in our theory, can be handled and completely solved, by giving *sufficient* conditions* and exact prescriptions on the *modus operandi* of Educanda, which guarantee that its performance will be such as wanted, whatever changes (consistently with our conditions) learning induces in it. Further developments of the present research will be matter for future publications.

The next step in our program will require the specific formulation of mnemonic laws such as to permit a learning of the type already discussed in [1], while fulfilling at all times the constraints determined here. This part of the work is closer to engineering than to mathematics, both because of technical reasons (the actual construction of machines depends on the state of technology), and because any machine, Educanda in particular, can only be a device to perform specific tasks. Therefore, learning must be such as to permit the wanted performance, the knowledge of which is necessary when assigning the mnemonic laws. This study will naturally lead to a quantitative specification of the A.L.H., that is to the specification of the duration of the transition phases in the behaviour of Educanda and of the speed of the learning processes. Explorative work already done in this direction seems to justify some optimism; it is also encouraging to consider the beautiful work done at Stanford University with Adaline, which shows, among other things, the remarkable learning possibilities already of a single neuron.

* In problems of this nature, our concern is with whole behavioural classes; the search for necessary *and* sufficient conditions is in general not possible, and not useful. Meaningful and useful are instead mostly conditions which are *either* necessary *or* sufficient.

NEURONIC EQUATIONS AND REVERBERATIONS

Introduce the Heaviside step function:

$$1[x] = \begin{cases} 1, & x > 0 \\ 0, & x \leqq 0; \end{cases} \tag{1}$$

the equations ruling the instantaneous behaviour of an N neuron's network are written in the following way [1]:

$$U_h(t + \tau) = 1\left[\sum_{k=1}^{N} \sum_{r=0}^{n(h)} a_{nk}^{(r)}\, u_k\,(t - r\tau) - S_h\right] \qquad (h = 1, 2 \ldots N) \tag{2}$$

where τ is a delay in the response of the neuron h, S_h its threshold, and the $a_{hk}^{(r)}$ denote the coupling coefficients. In a previous article [2] we indicated some transformations of Eq. (2) which are particularly useful in the study of periodic solutions or reverberations of preassigned period $R\tau$. A first transformation allows to write Eq. (2) in the form (after quantization of the time variable: $V_h(t + m\tau) = v_{hm}$):

$$V_{m+1} = \sum_r A^{(r)}\, 1\,[V_{m-r}] - \boldsymbol{S} \tag{3}$$

where $\mathbf{V}_m$ and $\mathbf{S}$ represent the vectors:

$$V_m \equiv \begin{bmatrix} V_{1,m} \\ V_{2,m} \\ \vdots \\ V_{N,m} \end{bmatrix} \qquad \boldsymbol{S} \equiv \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix}$$

$V_{h,m}$ is defined as:

$$V_{h,m} = \sum_{k,r} a_{hk}^{(r)} u_{k,\,m-r} - S_h$$

and $A^{(r)}$ is the matrix:

$$A^{(r)} \equiv \left\| a_{hk}^{(r)} \right\|$$

In this way the search for reverberations of period $R\tau$ is equivalent to determining solutions of Eq. (3) with the condition:

$$V_{m+R} = V_m \tag{4}$$

Another remarkable transformation is obtained by introducing the function sgn (x):

$$\operatorname{sgn}(\tau) \equiv \sigma(x) = \begin{cases} +1, & x < 0 \\ 0, & x = 0 \\ -1, & x > 0; \end{cases}$$

from the identity:

$$1[x] = \frac{1 + \sigma(x)}{2} \tag{5}$$

and imposing conditions such that $x \neq 0$ always. If we write (3) and (4) in the form:

$$x_\alpha = \sum_\beta A_{\alpha\beta} 1[x_\beta] - S_\alpha \quad (\alpha, \beta = 1, 2 \ldots M = N \cdot R)$$

we obtain

$$\boldsymbol{X} = \tfrac{1}{2} A\sigma(\boldsymbol{X}) + \tfrac{1}{2} A\,\mathbf{1} - \boldsymbol{S},$$

which for normal systems (defined in [2] as those for which $A\,\mathbf{1} = 2\,\mathbf{S}$) becomes:

$$\boldsymbol{X} = \tfrac{1}{2} A\sigma(\boldsymbol{X}) \tag{6}$$

For normal systems the following problems were solved [2]:

a) Given a network of N neurons, to change the solution of the equations for reverberations of preassigned period into a standard algebraic problem.

b) Given an arbitrary reverberation, to find coupling coefficients and thresholds which guarantee that a network thus built will admit of that reverberation.

c) To determine conditions sufficient for a system (6) to admit of no solution at all.

CONSTRAINTS SECURING WANTED REVERBERATIONS

In this section, we determine the constraints to be satisfied by the parameters of Educanda and the controlling systems to be added to it, which will secure that only reverberations of a preassigned period be performed; we begin our treatment with normal systems, and restrict our study to systems (2) with only $a_{hk}^{(0)} = a_{hk} \neq 0$.

Let us suppose first that the coupling coefficient matrix be decomposable in the form:

$$\tfrac{1}{2} a_{hk} = a_h b_k \quad (h, k = 1, 2 \ldots N). \tag{7}$$

In this case only $2n$ parameters (instead of n^2) are required to determine the connections in our machine.

The equations of interest are now:

$$V_{h,m+1} = a_h \sum_k b_k \sigma(v_{k,m}) \quad (h = 1, 2 \ldots N); \tag{8}$$

let us set:

$$q_0 \equiv \sum_k b_k \sigma(V_{k,0})$$

$$q \equiv \sum_k b_k \sigma(a_k);$$

we have then:

$$v_{h1} = a_h q_0$$

and, from Eq. (8):

$$V_{h,m+1} = a_n \sum_k b_k \sigma\Big[a_k \sum_s b_s \sigma(V_{s,m-1})\Big] \quad (h = 1, 2 \ldots N)$$

From the identity

$$\sigma(xy) = \sigma(x)\,\sigma(y)$$

it follows:

$$V_{h,m+1} = a_h q \sum_k b_k \sigma(V_{k,m-1}) \quad (h = 1, 2 \ldots N).$$

By iteration, we finally find:

$$V_{h,m+1} = a_h q[\sigma(q)]^{m-1}\,\sigma(q_0),\ m > 1 \quad (h = 1, 2 \ldots, N). \tag{9}$$

From Eq. (9) we see that:

$$\sigma(V_{h,m+1}) = \sigma(V_{h,m-1}),\ m > 1 \quad (h = 1, 2 \ldots N).$$

Let us now distinguish two cases:

a) $q > 0$. The machine then, exceptioning possibly an initial transient, does not change any more its state (reverberation of period 1);

b) $q < 0$. Excepting again an initial transient, the machine performs a reverberation of period 2. Reverberations are thus, at most, of period 2.

Let us now take into account the time-dependence of the coupling coefficients (due to some unspecified learning mechanism), by requiring only that the separability property (7) be maintained at all times:

$$\tfrac{1}{2}\, a_{hk}^{(m)} = a_h^{(m)} b_k^{(m)}$$

In this case, Eqs. (8) and (9) become respectively

$$V_{h,m+1} = a_h^{(m)} \sum_k b_h^{(m)} \sigma(V_{k,m}) \quad (h = 1, 2 \dots N). \tag{8'}$$

$$V_{h,m+1} = a_h^{(m)} \prod_{i=1}^{m} q^{(i)} \sigma(q_0) \quad (h = 1, 2 \dots N) \tag{9'}$$

where

$$q^{(i)} = \sum_k b_k^{(i)} \sigma(a_k^{(i-1)})$$

We find that, whatever learning may have taken place subject to these conditions, the machine will perform reverberations of period 2 at most.

Let us now consider the normal system

$$V_{m+1} = B\sigma(V_m) \quad [B \equiv (b_{hk} = \tfrac{1}{2} a_{hk})]$$

and introduce the matrix $\tilde{B}$ defined in the following way:

$$\tilde{B} \equiv \begin{pmatrix} b_{11} b_{12} \dots b_{1N} \cdots \beta_{11} \beta_{12} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ b_{N1} b_{N2} \dots b_{NN} \cdots \beta_{N1} \beta_{N2} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ \alpha_{11} \alpha_{12} \dots \alpha_{1N} \cdots \alpha \quad 0 \\ \alpha_{21} \alpha_{22} \dots \alpha_{2N} \cdots 0 \quad \beta \end{pmatrix}$$

where

$$\alpha_{ij}, \beta_{ji} \quad (i = 1, 2), (j = 1, 2 \dots N)$$

$$\alpha, \beta \quad (\alpha, \beta \neq 0)$$

are arbitrary numbers to begin with.

Let us introduce the matrix

$$\tilde{\sigma} \equiv \begin{vmatrix} \sigma(x_{11}) & \sigma(x_{12}) & \sigma(x_{13}) \dots \\ \sigma(x_{21}) & \sigma(x_{22}) & \sigma(x_{23}) \dots \\ \vdots & \vdots & \vdots \\ \sigma(x_{N1}) & \sigma(x_{N2}) & \sigma(x_{N3}) \\ 1 - \frac{\boldsymbol{\alpha}_1}{\alpha} \cdot \sigma(X_1) & -\frac{\boldsymbol{\alpha}_1}{\alpha}, \sigma(X_2) & 1 - \frac{\boldsymbol{\alpha}_1}{\alpha}, \sigma(X_3) \dots \\ -\frac{\boldsymbol{\alpha}_2}{\beta} \cdot \sigma(X_1) & 1 - \frac{\boldsymbol{\alpha}_2}{\beta} \cdot \sigma X(_2) & -\frac{\boldsymbol{\alpha}_2}{\beta} \cdot \sigma(X_3) \dots \end{vmatrix}$$

where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are the vectors

$$\boldsymbol{\alpha}_i \equiv (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iN}) \quad (i = 1, 2),$$

in which the elements $x_{h,j}$ are recursively defined in the following way:

$$x_{h1} = V_{h1} \quad (h = 1, 2, \ldots, N), (j = 2, 3, \ldots)$$

$$x_{h,j} = \sum_k b_{hk}\sigma(x_{k,j-1}) + \beta_{h1}\left[\frac{1 + (-1)^j}{2} - \frac{\alpha_1}{\alpha} \cdot \sigma(\mathbf{X}_{j-1})\right] + \\ + \beta_{h2}\left[\frac{1 + (-1)^{j+1}}{2} - \frac{\alpha_2}{\beta} \cdot \sigma(\mathbf{X}_{j-1})\right] \quad (10)$$

$$X_{N+1,i} = \frac{1 + (-1)^{i+1}}{2}\alpha$$

$$X_{N+2,i} = \frac{1 + (-1)^i}{2}\beta$$

From the given definitions, it is easy to demonstrate that the following relation holds:

$$X = \tilde{B}\,\tilde{\sigma}$$

where X denotes the matrix

$$X \equiv \begin{pmatrix} X_{12} & X_{13} & X_{14} & X_{15} & \ldots \\ & & & & \\ X_{N2} & X_{N3} & X_{N4} & X_{N5} & \ldots \\ \alpha & 0 & \alpha & 0 & \ldots \\ 0 & \beta & 0 & \beta & \ldots \end{pmatrix}$$

Let us now require that the coupling coefficients be such that the rank of $\tilde{B}$ is 2. It follows then that, if this requirement is satisfied, the matrix X has also rank 2, from which it follows:

$$X_{h,m+1} = X_{h,m-1}, \quad m > 2 \quad (h = 1, 2, \ldots, N)$$

We can therefore conclude that the machine can only perform reverberations of period 2.

Of course, this method can be generalized to the case of reverberations of arbitrary period K. For this, it suffices to consider $2K$ N-component vectors:

$$\alpha_i, \beta_j \quad (i, j = 1, 2, \ldots, k)$$

at K arbitrary non-zero numbers $\alpha, \beta, \gamma, \delta \ldots$, and to require that the matrix $\tilde{B}$ be of rank K. The number of arbitrary parameters becomes then $2KN + K$. The method presented here requires, obviously, the intro-

duction of K controlling elements, which cause the machines to perform only reverberations of period K.

Let us now go back to the case of a separable matrix, i.e. $b_{hk} = a_h b_k$, for which we solved already the reverberation problem.

We have, by means of the method expounded here, for reverberations of period 1,

$$\tilde{B} = \begin{pmatrix} a_1 b_1 & \alpha_1 b_2 \ldots a_1 b_N & b_1 \\ \ldots\ldots\ldots\ldots\ldots & & \\ a_N b_1 & a_N b_2 \ldots a_N \beta_N & \beta_N \\ \alpha_1 & \alpha_2 \ldots \alpha_N & a \end{pmatrix}$$

By imposing that the rank of $\tilde{B}$ be 1, we deduce:

$$\alpha a_h b_k = \alpha_k \beta_h$$

therefore:

$$X_{h,m+1} = \sum_k \left(a_h b_k - \left(\frac{\alpha_k}{\alpha}\beta_h\right)\sigma(X_{k,m}) + \beta_h = \beta_h\right.$$

i.e. we have a reverberation of period 1. For reverberations of period 2, we find likewise:

$$X_{h,2m} = \beta_{h1} + \sum_k \left(b_{hk} - \beta_{h1}\frac{\alpha_{1k}}{\alpha} - \frac{\beta_{h2}}{\beta}\alpha_{2k}\right)\sigma(X_{k,2m-1})$$

$$X_{h,2m+1} = \beta_{h2} + \sum_k \left(b_{hk} - \beta_{h1}\frac{\alpha_{1k}}{\alpha} - \frac{\beta_{h2}}{\beta}\alpha_{2k}\right)\sigma(X_{k,2m}),$$

Since $\tilde{B}$ has rank 2, it follows:

$$b_{hk} - \frac{\beta_{h1}}{\alpha}\alpha_{1k} - \frac{\beta_{h2}}{\beta}\alpha_{2k} = 0$$

so that we may conclude:

$$X_{h,2m} = \beta_{h1}, \; X_{h,2m+1} = \beta_{h2}$$

in conformity with our previous statement that all reverberations are of period 2.

Thanks are due to Drs. A. de Luca and L. M. Ricciardi for many interesting discussions.

REFERENCES

1. Caianiello, E. R. (1961). "Outline of a theory of thought processes and thinking machines." *J. Theoret. Biol.* **1**, 204–235.
2. Caianiello, E. R. (1966). "Decision equations and reverberations" Kybernetik, in print. See also a report by the same author in the Proc. of the 1965 Ravello School on "Functional Analysis and Optimization," to be published by Academic Press.

W. L. KILMER,

Department of Electrical Engineering, Michigan State University, East Lansing, Michigan.

W. S. MCCULLOCH,

Research Laboratory of Electronics, Massachusetts Institute of Technology Cambridge, Massachusetts

J. BLUM,

Massachusetts Institute of Technology, Instrumentation Laboratory Cambridge, Massachusetts

E. CRAIGHILL

Department of Electrical Engineering, Michigan State University East Lansing, Michigan

and D. PETERSON

Massachusetts Institute of Technology, Instrumentation Laboratory Cambridge, Massachusetts

*On a Cybernetic Theory of the Reticular Formation**

ABSTRACT

Throughout the life of the vertebrates, the core of the central nervous system, sometimes called the reticular formation, has retained the power to commit the whole animal to one mode of behavior rather than another. Its anatomy, or wiring diagram, is fairly well known, but to date no theory of its circuit action has been proposed that could possibly account for its known performance. Its basic structure is that of a string of similar modules, wide but shallow in computation everywhere, and connected not merely from module to adjacent module, but by long, jumpers between distant modules. Analysis of its circuit actions hereto-

* The major portion of this work was supported by Air Force Cambridge Research Laboratories Grant AF 19(628)-5076, administered through Michigan State University. Support was also given by the following: Air Force Office of Scientific Research Grant AF-AFOSR-1023-66 through Michigan State University; DSR Project 55-257, sponsored by the Bioscience Division of the National Aeronautics and Space Administration, Contract No. NSR 22-009-138 through the Massachusetts Institute of Technology Instrumentation Laboratory; the National Institutes of Health Grant NB-04985-03 through the Massachusetts Institute of Technology; U. S. Air Force (Research and Technology Division) Contract AF 33(615)-1747 through the Massachusetts Institute of Technology; and the Teagle Foundation, Inc., through the Massachusetts Institute of Technology.

fore proposed in terms of finite automata or coupled nonlinear oscillators has failed.

We now propose nonlinear, probabilistic hybrid computers as proper modules, and describe a behavioral simulation of an anastomatically-coupled linear array of 12 such computers. The model contains about 2200 wires, yet still behaves as an integral unit, rolling over from stable mode to stable mode according to abductive logical principles, and as directed by its succession of input 60-tuples. We discuss the model's design in terms of modular input focusing, module decoupling, redundancy of potential command, and gains around reverberatory loops. Future work on memory mechanisms and biological experiments is indicated.

INTRODUCTION

Throughout the vertebrate phylum, the reticular formation (RF) is the nervous center which integrates the complex of sensory-motor and autonomic-nervous relations so as to permit an organism to function as a unit instead of a mere collection of organs. The RF consists generally of the nervous core of the spinal cord, but bulges somewhat in higher animals in the lower spinal (lumbar) region, and in regions corresponding to the neck and brain stem areas of man (see Fig. 1). In the highest vertebrates it comprises about 1/1000th of the central nervous system. The RF receives relatively unrefined information from all of the sensory-motor systems which link the organism to its environment (visual, auditory, vestibular, etc.) as well as from all of the internal housekeeping systems which insure the organism's internal wellbeing (visceral, cardiovascular, respiratory, etc.). Its primary job is to commit the organism to either one or another of about 16 gross modes of behavior—i.e., run, fight, sleep, speak, etc.*—as a function of the nerve impulses that have played

* To illustrate, the RF "modes" for cat might be:

1, 2 lie	9 vocalize
symmetrically or not	10 surrender
3 sit	11 eat
4 stand	12 drink
5, 6 locomote	13 vomit
symmetrically or not	14 fornicate
7, 8 fight	15, 16 sleep
anger, full flight	EEG high voltage slow wave
	EEG low voltage fast activity

This classification is rather arbitrary, but the point is that nobody would list more than 30 modes.

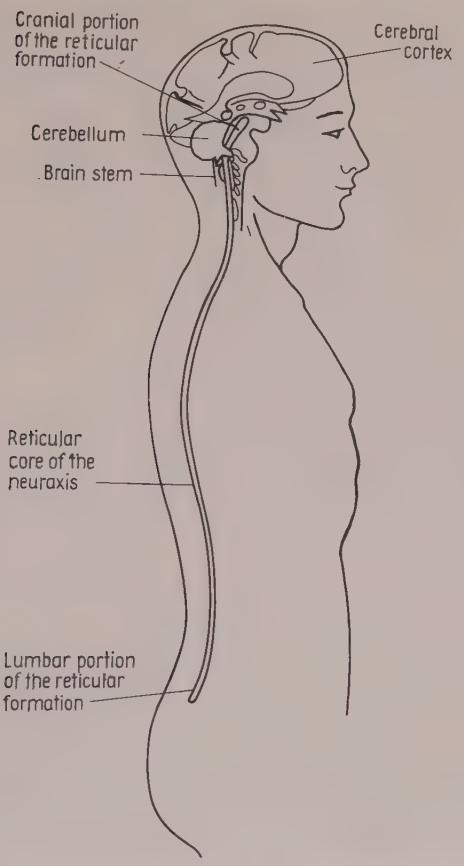


Fig. 1. Location of the reticular formation in the cranium, brain stem, and spinal cord of man.

in upon it during the last fraction of a second, and also to send out control directives to the other more specialized nerve centers so that they in turn can behave in an integrated fashion.*

* The RF plays a role in animal central nervous systems analogous to that of the admiral in a battle fleet. In the highest vertebrates, where a vast number of variations on modal behavioral themes are mediated through other brain regions, the RF can also be viewed as a generalized information filter and function setter for the rest of the central nervous system. Some prefer this viewpoint, but it does not seem to differ basically from ours above.

At the millipede stage of evolution, the RF is the entire nervous system. By the pigeon stage, it has grown, or separated out, several comparatively specialized computers for making finer discriminations between sensory stimuli, and for computing more precise motor control signals, than it could possibly produce by itself and still maintain its overall command and control function. Chief among these specialized computers are the visual, vestibular, bodily-sensory, and auditory systems, and the cerebellum to compute precise autocorrelations for actions of the pigeon on the pigeon, and the pigeon on its world (as required for pecking, control in flight, etc.). The pigeon RF has also evolved specialized mechanisms for programming its associated bodily movements (required for running, fighting, feeding, mating, etc); a set of well localized feedback paths, called simple reflexes; and a set of regenerative nerve loops for controlling various types of internal rhythms (cardiovascular, respiratory, digestive, etc.). It is still clear, though, that the pigeon, sea gull, and other organisms of that evolutionary rank behave in a distinctly modal fashion.

By the human stage of evolution, the RF has grown a cortical mantle over the rest of its phylogenetically older structures. These older structures, when left by themselves, are only concerned with the rather more immediate preservation of the individual and its species.[1] But in humans, we find new and different types of functions, like language; we also find that the behavioral influences of many of the older functions, like anger, are greatly modified, and that an enlarged frontal lobe has mushroomed the development of long range judgment and deliberative purposes in the organism. We find, too, that the visual, auditory, bodily-sensory, and motor-outflow computers are larger and more specific than ever. Yet for all the RF's reliance on the discriminatory, associational, memory, computing, and programming powers of the cortex, it has never relinquished its central command function to the cortex.[2] It couldn't have, and have survived. For only its computations are wide enough (have sufficient scope) to be able to encompass the crucial information in every eventuality, and also shallow enough (do not have too much logical depth) to always arrive at a modal decision within a fraction of a second given sufficient information.

In the present report we are concerned only with that aspect of the RF which makes the decision to commit the organism, henceforth denoted RF*. RF* is RF minus everything on the RF's input side (the dorsal-

lateral RF) and output side (the ventral-lateral RF, basal ganglia, etc.), all of the reflexes along the neuraxis that are handled locally, and all of the respiratory and other rhythmic operational aspects which are functionally separable from the modal decisionary task.

We will first sketch the known neurophysiology and neuroanatomy of the whole RF, and then propose a theoretical framework for RF* that we think stands a chance of being right enough to eventually be of some use in understanding real RF's. What we are after is a way to think about how RF*'s always arrive at integrated modal decisions within their characteristic anatomy instead of disagreeing among their several parts in the face of competitive or even contradictory input signals. But we proceed humbly, for to date no one has been able to invent a satisfactory theory for reasoning about RF's.

NEUROPHYSIOLOGY AND NEUROANATOMY OF THE RF

The Scheibels[3] have so far done the most definitive neuroanatomy that we have on the RF. In their milestone report of 1962 they caricatured its anatomical structure in the brain stem by comparing it to a stack of poker chips. In each chip region the dendritic processes of RF neurons ramify in the plane of the chip face, often covering nearly half of the face area. This causes a very large degree of overlap and intermingling among dendrites of nearby neurons, as shown schematically for just the brain stem region in Figure 2. (This is very similar to Scheibel's Figure 1 in reference 3.) The dendritic organization of the nerve nuclei that furnish inputs to the RF is predominantly longitudinal, as seen in Figure 2b. The axons out of these nuclei, and the axonal collateralizations out of all of the longitudinal fibre tracts that feed into the RF, turn off sharply to reach into it in the planes of the RF's greatest dendritic ramification. Since in this process as many as a half dozen or more input systems may synapse on a single RF neuron, and each RF input nucleus and fibre tract in general feeds very many RF cross-sectional levels, the Scheibels suggest that the RF might tolerate considerable puddling of information at each of its cross-sectional levels, but demand somewhat greater informational rigor between levels.

The order of magnitude of the number of afferents to the RF, the number of RF neurons, and the number of RF efferents is accepted as

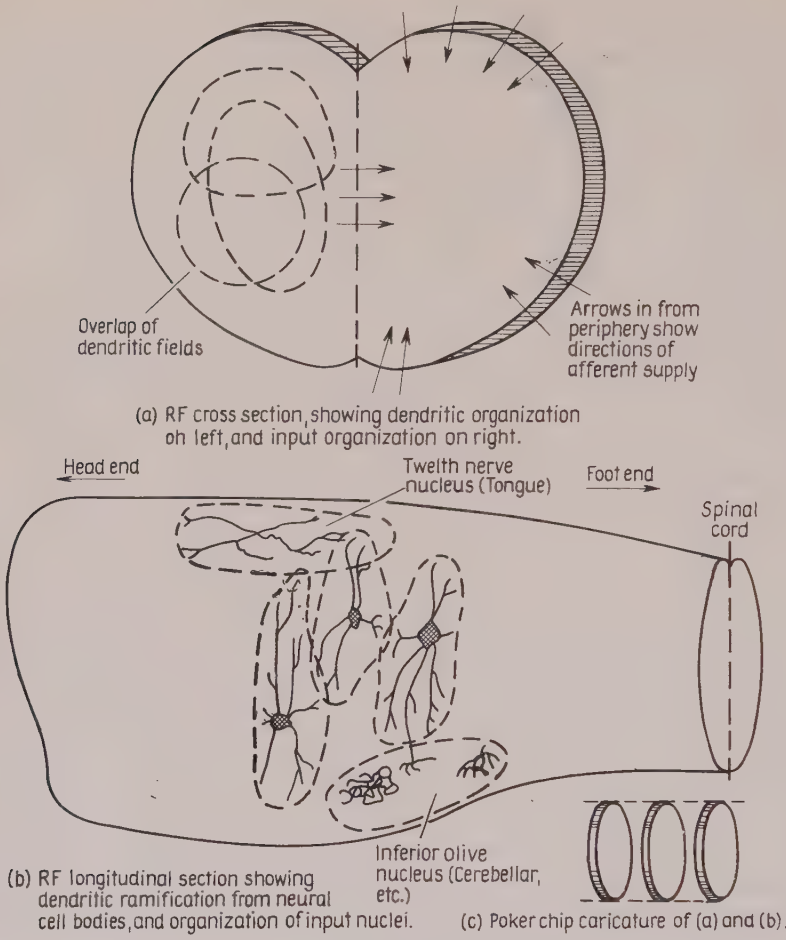


Fig. 2. Brain stem RF dendritic anatomy.

being nearly the same. RF dendrites generally appear to fan out about 60° ventrally from more ventral RF cell bodies, and about 180° dorsolaterally from more dorsal RF cell bodies. All processes of more laterally situated RF neurons are in general smaller than their more medial counterparts. Smaller RF neurons are, of course, concerned more with local operations, and larger ones more with global functions. In general, more ventral RF neurons participate more extensively in effector functions, more dorsal RF neurons in sensory functions, and more lateral RF neurons in vegetative functions.

The RF axonal anatomy corresponding to Figure 2b is shown in Figure 3. (This drawing is essentially Scheibel's Figure 4 in reference 3.) There a characteristic RF axonal process is seen coursing its way longitudinally over a major portion of the brain stem. Collaterals branch off into other RF levels and various RF input nuclei, as well as into both corticifugal (i.e., descending) and corticipetal (i.e., ascending) neural fibre tracts. A good many RF axons also project nonspecifically* into cerebral regions, as well as directly out to the level of the first synapse in each of the sensory systems (e.g., just behind the retina of the eye, and in the inner ear). In short, the RF sits athwart all incoming and outgoing nervous transactions carried out over the entire neuraxis, and it both samples and modulates their spatio-temporal information sequences so as to command the gross modal operation of the organism.

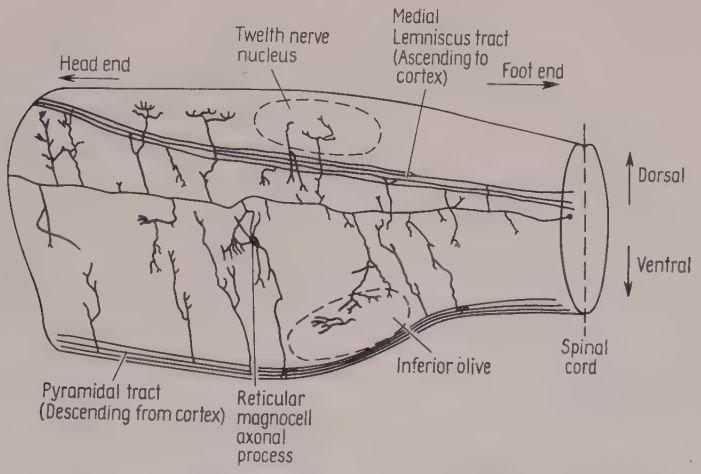


Figure 3. RF axonal anatomy corresponding to Figure 2 dendritic anatomy.

Essentially all that is known about the neural architecture of the RF is that there is a full range of neural cell body sizes.[3] But neither near circumscription of neural groups nor laminar or other striking distributive organization is evident. Thus no hypotheses have yet been develo-

* "Nonspecific" projection tells cortex what functions to compute, but does not furnish the information for computation. The projection alluded to is mostly indirect, i.e., it passes through at least one thalamic synapse.

ped relating the geometric forms of RF neurons to the functions they compute.

Axonally, we know that near any given RF neural cell body there may be tens of thousands of both fast conducting (100 meters/sec) insulated fibres and slow conducting (a few meters/sec) uninsulated fibres, but the functional significance of this has only been guessed at. We know that RF neurons characteristically respond to exceptionally wide ranges of stimuli involving perhaps several sensory and vegetative modalities.[2, 4, 5, 6, 11] For example, in cat there are RF neurons which increase their firing rates during asphyxia, tickling of hair cells in the nose, and postural unbalance. Other RF neurons respond to visceral disorders, crude body interface phenomena (touch, pressure, and cold), and certain phases of anti-gravity bodily kinetics; and still others to cerebral control signals, raw information from head end distance receptors (eye and ear), and signals from neuroendocrine receptors in the hypothalamus. We know there is massive reticular involvement in motor outflow and attention-focusing affairs.[7] For example, rats do not ordinarily distinguish yellow; but if they are hungry and smell cheese, RF directed outflow sets up visual computations which enable them to. A modal decision, then, can amount essentially to a very broad command to attend, e.g., to running, fighting, etc. We know that RF neurons are the first to adapt out their responses to meaningless stimuli (e.g., gunfire at a shooting range),[2, 3] and the first to condition to sensory indication of imminent painful stimuli.[2, 3, 5] So we understand something of how vertebrate organisms take habits and conduct their neuronal affairs at the reticular level.

There is strong evidence to suggest that the RF can change modal commitments at a steady rate of not more than about 3 times per second (spinal reflexes, some of whose paths can be traversed in about 20 milliseconds, notwithstanding), but must be driven with pulse repetition rates of the order of a few hundred per second for this to occur. Too low and too high pulse rates inside reticular tissue have very little overall effect.[8] If the RF is engaged in a significant overall decisionary activity, probably the focusing down affects following the crest of this activity persist for a minute or more.[5, 9] We conjecture that cortical perceptions (is there a lion behind that bush or not?) are produced at the rate of about ten per second,[46] and that this is the main limiting temporal factor in those cortico-reticular exchanges that primarily concern modal decisions. We

note effects that humoral and hormonal rhythms, with periods of hours, days, months, and years, can have on the overall sensitivity and set of an RF.

We know essentially nothing about the kinds of spatiotemporal information codes that the RF employs to cope with its horrendous intrareticular communication and computation problem (though some ambiguous signal channeling as a function of pulse repetition rate has often been noted experimentally — see, e.g., reference 2), nor do we know anything fundamental about the patterns of RF neural cooperation and dissociation (either physiological or psychiatric) that arise during modal decisionary activities.

In short, a good deal is known about the RF's input and output systems, but somewhat less about the RF's neuroanatomy, especially the detailed connection patterns among the various neural types, and their respective counts with changes of position in RF. Practically nothing is known about how RF inputs start appropriate computations racing up and down the net so as to always yield effectively unanimous decisions for modal command and control signals.

We are convinced that if the RF is ever to be really understood, we must have a theoretical model that will enable us to intuit from it logically sophisticated experiments of sufficient cybernetic dimensionality and complexity to take into account those differences which make a difference. N one-dimensional experiments can never take the place of one N-dimensional experiment in a thoroughly N-dimensional system.

LOGICAL REQUIREMENS OF THE THEORY

First of all, we must understand what kind of logic the RF* described in our introduction performs. C. S. Pierce called it the logic of relations, but for its clearest statement we go straight to the father of modern biology, Aristotle. He described three kinds of logic: deductive, inductive, and abductive. RF* does the latter. Its scheme is to go from facts and rules to cases: i.e., *facts* of sensory and internal perception as represented over the RF* input channels, and *rules* for deciding on the organism's mode of behavior as a function of the RF*'s ability to classify its immediate environmental stimuli, to *cases*. (For example, does the present input derive from a "case" of the stomach's inability to play its digestive role?

If so, the "rule" is vomit.) An organism's case structure is always the result of its evolutionary, and to a much lesser extent, its developmental endowment. RF*'s abductive task is rendered at once easier and more difficult by the correlated natures of its inputs (e.g., recognition of a case under the mating rule may well involve visual, olfactory, and bodily touch information that generally blends correlated reports on a world that is itself covarying).[6, 11]

After each new modal decision, RF* keys the proper output program, and from then on as far as it is concerned everything follows in a completely deductive manner. For example, the programmed output of the basal ganglia throws its keyed signal sequences for walking down over the interlocked entrainments of nerve centers in the arms and legs, and they in turn embellish the details of the orders given them as the contingencies of rough terrain, etc., demand; and so on out to the periphery, where the effector signals are transformed into smooth and complicated actions.)[6, 4] Most of this deductive activity apparently begins in the lateral RF right off the RF core. Actually, RF* no doubt never computes single modal decisions directly, but rather their half-center representations. To illustrate, Figure 4 shows the half-center dimensions for the lumbar enlargement of a dog. The advantage of such a representation is that for

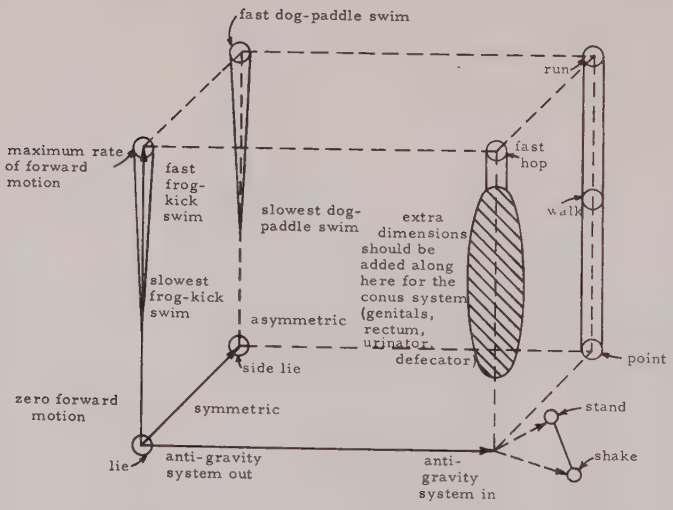


Fig. 4. Half-center dimensions for lumbar RF of dog. (Note: Heavy lines show dimensions.)

n dimensions, a single 2^n-valued function can be replaced by n two- (or sometimes more) valued functions.

Conditioning, habituation, and long-term learning in the RF require only inductive logic, and quite probably at this stage are best studied in cortex.

CONVERGENCE ON A MODEL

To a certain extent, any chunk of nerve tissue that has to perform both an analyzer and an integrator function, as RF* does, can be viewed as an assembly of coupled nonlinear oscillators. (In a very rich sense, all neural tissue amounts to such an assembly, and certainly behaves nonlinearly overall.) In fact, the variety of abductive logic our RF* employs on its highly correlated input sequences strongly suggests a model of rather tightly-coupled multistable oscillator units cooperating probabilistically so as to admit at any given time only one of a small number of possible stable overall operating modes. Thus we turned to Wiener's work on correlation-coupled nonlinear oscillators, which shows that there are forbidden zones about each stable point, and suggests that generally such systems behave as required.[17] To illustrate, Figure 5 (Wiener's Figure 8.4 in reference 17) depicts for such a system the shape of the probability distribution of oscillator frequencies about a normalized stable point, as calculated by Wiener. The crucial thing about Figure 5 as far as our RF* model is concerned is that there is enough variability

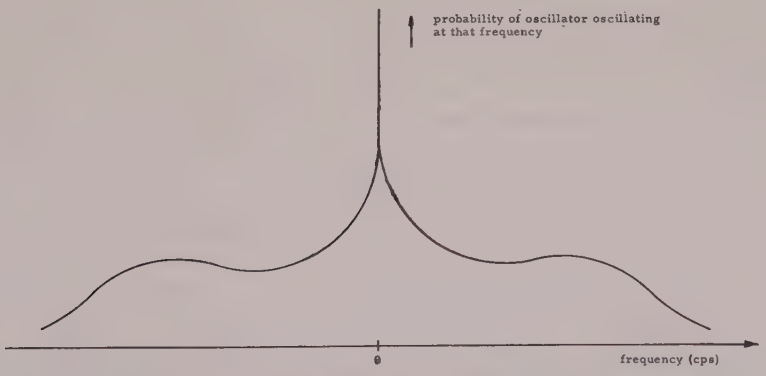


Fig. 5. Probability distribution of frequencies in a system of coupled nonlinear oscillators. After Wiener (ref. 17).

about the stable point to permit flexible system operation. Unfortunately, we found that a central defect of all such systems is that there is no reasonable analytical or experimental way of determining anything basic about the transient behavior between stable mode points following input changes. The same holds true for every sufficiently complex nonlinear artificial neural net theory we know of to date. (See, for example, references 18, 19, 20, and 21.) But to be able to follow such transients is central to our task, so we had to regard these systems as useless. Recalling that single nonlinear oscillators a la Minorsky[22] behave too rigidly, and that linear systems are unable to exhibit the necessary memory, cell assembly, and modal features,† we abandoned all coupled-oscillator and neural-net approaches to RF∗ theory construction as utterly hopeless.

Evidently we required considerably more initial structure than they would afford, i.e., we needed a well-developed set of strategies for designing a skeletal model of RF∗ behavior. Then we could revert to a computer-simulation and mathematics to investigate the complex behavioral consequences of varying these strategies. In order to pursue this plan without doing too much violence to the biology, we returned to the Scheibels' stack of poker chips analogy. The result is described below.

THE PRESENT MODEL

This section describes our present RF∗ model, which we denote S-RETIC. It is a caricature of the poker chip analogy for the brain stem RF, and was constructed to be as simple as we could make it without doing too much violence to our intuitive notion of reticular operation.

In the model we replace adjacent groups of the Scheibels' approximately 100-micra-thick chip regions by single modules which contain nonlinear, probabilistic hybrid computers. We require that all modules in the resulting columnar array be similar and operate on the same synchronous time scale. The modules are interconnected to a degree and in a way suggested by the known RF axonal anatomy. The anatomy

† In other words, the long-term responses of linear systems are determined in a 1 : 1 manner from their input drives, whereas this is not true for nonlinear systems. So in this sense, linear systems are irredundant, whereas nonlinear ones have a chance of being redundant in the right way for modeling neural behavior.

also serves as the guide in specifying extra-S-RETIC inputs to each of the modules.

The procedure is somewhat different on the S-RETIC output side. Since it is mainly the computational structure of single RF modal decisions that we are interested in, each of the S-RETIC modules is endowed with only mode-indicator outputs. It is assumed that the effects of each S-RETIC modal change show up appropriately at some slightly later time at the S-RETIC's input. This departs from the RF biology in that the S-RETIC outputs have no direct way of influencing the organism's input and output systems which feed it. Such an overall-output approach seems justified by results like Doty's, which show that even brain stem swallowing motoneurons "seem to have an unpredictable and random pattern from one swallow to the next, though the overall schedule of excitation and inhibition among the participating muscles is highly constant."[10]

Figure 6 shows a reduced schematic of our S-RETIC. All σ_i and γ_i lines are binary (an arbitrary but convenient basis of information coding); the M_i are S-RETIC's logic modules; the S_i correspond to the various exteroceptive and interoceptive sensory and internuncial systems which feed inputs directly into the RF at DI; the Ω_i (only Ω_7 shown) are the modular mode-indicating output lines; and the upper and lower boundaries, T and B, correspond roughly to the diencephalic and high cervical regions of the higher vertebrates, respectively. For clarity, each type of connection appears in Figure 6 only once, whereas actually the connection types proximate to M_7 recur at all corresponding similar locations over the entire figure. Thus, if a connection type diverges from or converges to one or a group of S_j or M_i in Figure 6, it does likewise at every corresponding similar location in S-RETIC. Thus each M_i receives inputs from several but not all S_j, and each S_j feeds several but not all M_i. Also, each M_i sends information directly to several but not all other M_i, and receives information directly from several but not all other M_i.

The M_i-to-M_j connections are arranged so that, in general, modules close together are information-coupled more closely than modules far apart. This is in line with the neuroanatomy. Similar restrictions are also all that govern the terminal distributions of the A and C bundles, though the anatomy of Nauta and Brodal suggests some further specificity in this regard. The S_j output and M_i ascending and descending bundle sizes are delimited to 5, 4, and 4, respectively; and the degrees

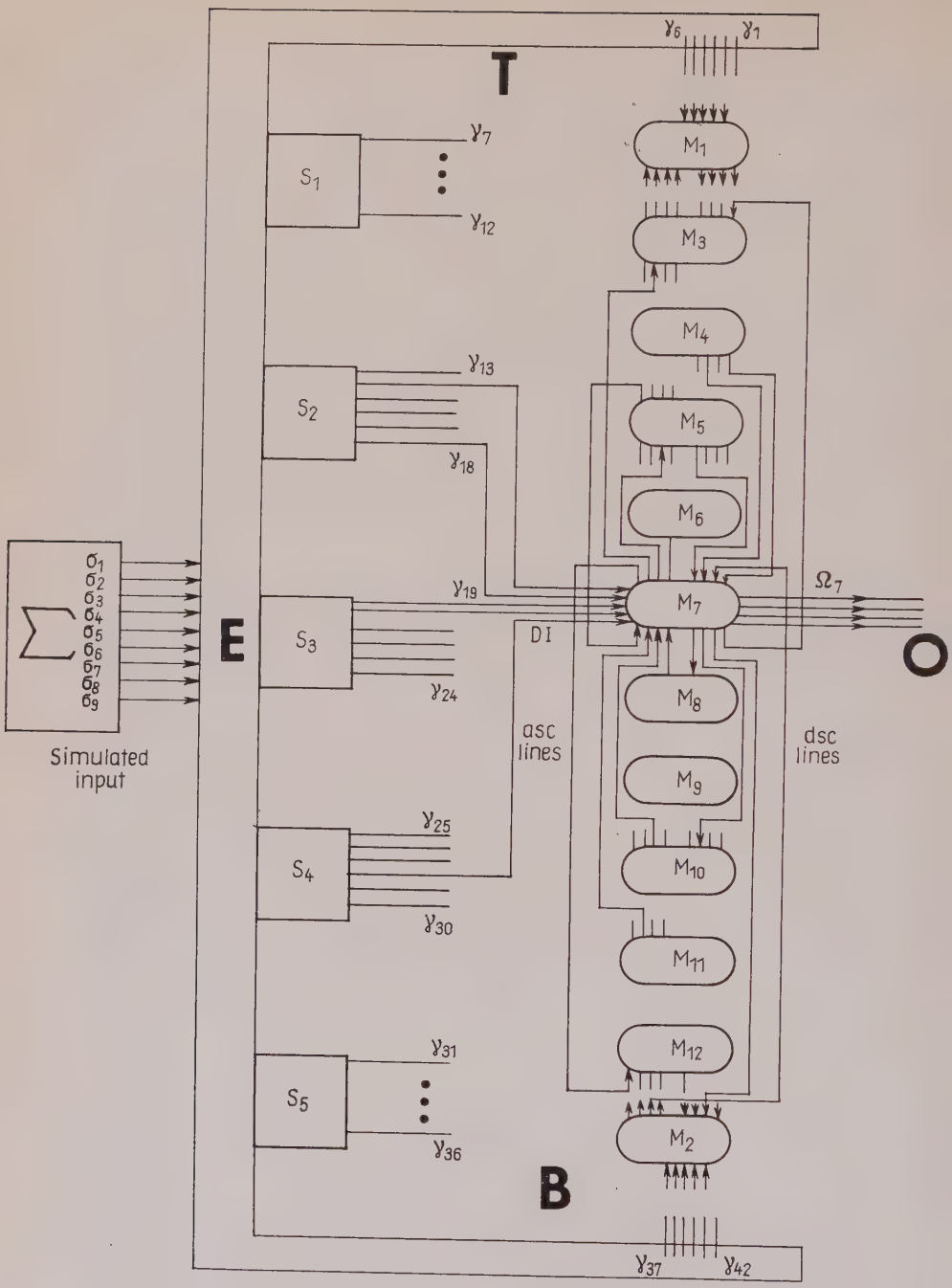


Fig. 6. S-RETIC simulation model.

of S_j-to-M_i fan-in and fan-out are delimited as suggested by the RF input anatomy (involving nuclear regions, fibre tracts, and the somewhat more localized lateral reticular structures). The precise nature of these delimitations is suggested in Figure 6, and is specified in detail in Appendix 3. The specification was made with the intention of imposing fairly even "use" distributions on the γ_i and σ_i. The corresponding γ_i connection relations are important only because the information on these lines is highly (but nontrivially) correlated, with the relative degree of correlation between each pair of γ_i determining their relative proximity in the $\gamma_{11}, \ldots, \gamma_{42}$ ordering. The γ_i are realizations of all 42 symmetric switching functions of the form $(\sigma_i \wedge \sigma_j) \vee (\sigma_j \wedge \sigma_k) \vee (\sigma_k \wedge \sigma_i)$, with i, j, and k pairwise different, $i, j \leqq 7$, and $k = 8$ or 9. (Cf. Appendix 1 for tabulation of functions.) This keeps the percentage of 1's on the γ_i's, about the same as that on the σ_i's (cf. Appendix 2 for details), and preserves some useful distance properties in the passage from σ_i to γ_i signal sets. The M_i-to-M_j connections are made randomly so that the probability of a direct M_i-to-M_j connection is inversely proportional to the absolute magnitude of the square root of $(i - j)$. (Cf. Appendices 6 and 7 for connection table and details.) Σ and E in Figure 6 are thus included only to simulate an RF environment that engenders input signals from a covarying world. (For example, if a runaway car stops abruptly at a wall at the bottom of a hill, a witness is likely to hear a crash. His visual and auditory pathways then transmit correspondingly covariant signals into his reticular formation.)

Before proceeding to a formal specification of each M_i in our Figure 6 RF$*$ model, we might ask if general automata structures of that ilk have been studied mathematically. The answer is yes, and for more on this we turn to the theory of iterative net automata; for there is much there that illuminates the basic nature of our S-RETIC modal computation analysis problem. Mainly, iterative net theory, and the unpublished lucubrations that have gone into shaping it, tell us where the central difficulties lie, and what we must do and not do if we are to slip our way around and through them to satisfactory S-RETIC designs. The connection between our Figure 6 S-RETIC schema and iterative nets is just that the latter are very special cases of the former. To see this, let us review the usual definition of an iterative net. We define the M_i in Figure 7 for all i between l and n inclusive as follows: σ_i is M_i's external input; α_i is its ASC input; δ_i is its DSC input; μ_i is its external output; and λ_i is the de-

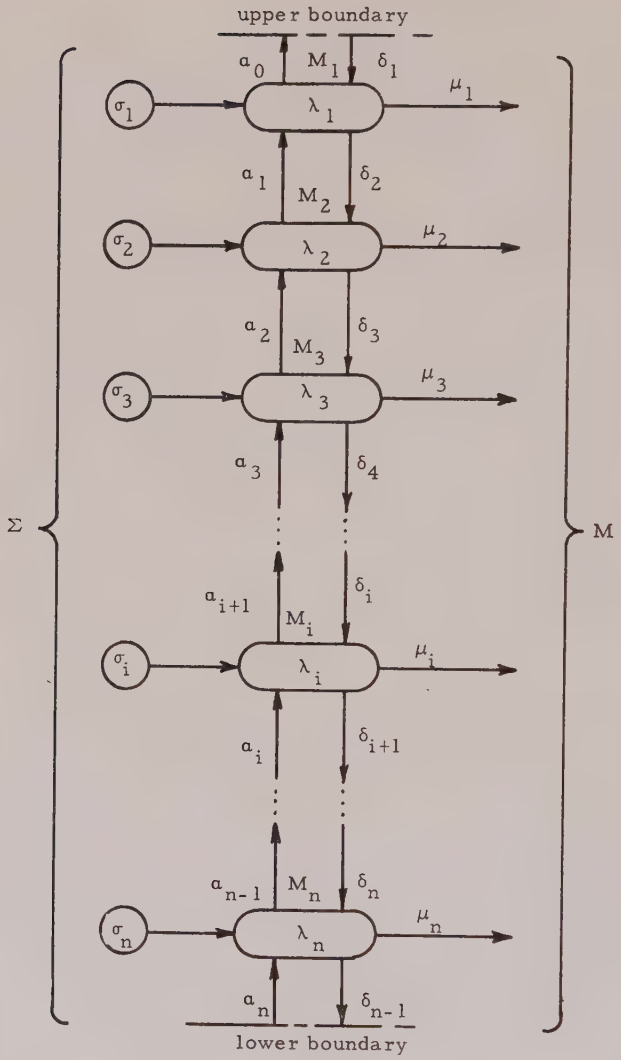


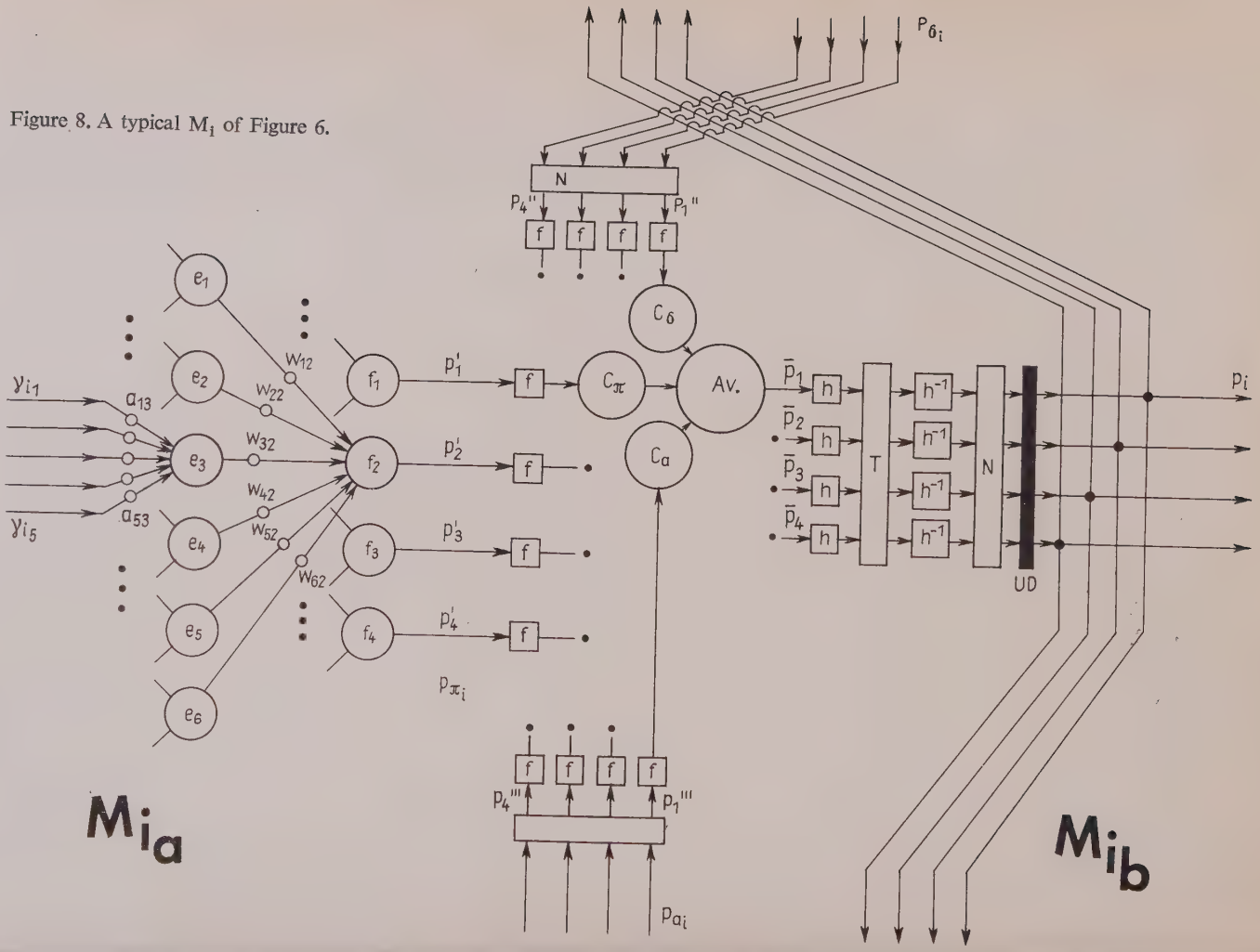
Fig. 7. A highly specialized instance of the *S*-RETIC schema.

scription of its time-invariant logic. This logic defines for each of the (finitely many possible) σ_i, α_i, δ_i input arguments of M_i a corresponding set of α_{i-1}, δ_{i+1}, η_i values which appear out of M_i one time later. (It has been shown that the different α, δ output timing is essentially inconsequential.) We denote the result a BITN (for Bilateral ITerative Net), and call the M_i of BITN's "memoryless modules" if and only if each M_i performs only combinational logic: i.e., does not allow its λ_i to vary as a function of M_i's past experience. Thus memoryless module BITN's are evidently just cut down versions of S-RETIC's. We give in Appendix 8 a precise description of all of our BITN results to date that are even remotely relevant to the S-RETIC problem. The tenor of these results is that there do not, and cannot, for deep underlying logical reasons, exist complete mathematical theories for coping with any *completely general* aspect of our S-RETIC design and analysis problems. Realizing this, we decided we had to specify a set of particulars for our Figure 6 model if we were ever to understand anything really nontrivial about its organizational implications.

Figure 8 indicates the scheme we chose for these particulars. It arose out of our desire to embody the logic, if not the mechanisms, of coupled nonlinear oscillator manifolds which operate on information according to strategies a) and b) developed later in this section. In Figure 8 the e_i in M_{i_a} are all linear threshold logic elements. The small circles on each part of their input lines are gain settings, or weights, and they can be either positive or negative. Each threshold logic element works as follows: Assume such an element has input lines $i_1, i_2, \ldots, i_n$ whose weights are $w_1, w_2, \ldots, w_n$, respectively. If at time t the element's jth input line is excited, set i_j equal to 1; otherwise, set i_j equal to 0. Then if the element's threshold is 0, its output is 1 at if and only if $\sum_{j=1} i_j w_j$ exceeds 0 at t; otherwise it is 0. The input connections to e_j, $1 \leqq j \leqq 6$, in Figure 8 are all the same, except each of the e_j in general has a different set of input weights, a_{ij}.

The f_j in Figure 8 amount to only half a threshold element: f_j's output is $\sum_{i=1}^{m} (w_{ij})$ (e_i's output), and so is analog. The w_{ij} are all required to be non-negative, and the p_i give the probabilities, as computed by M_{i_a} from only $\gamma_{i1}, \ldots, \gamma_{i_5}$, that the present overall γ_i signal configuration properly corresponds to an ith mode output indication. Actually, in

Figure 8. A typical M_i of Figure 6.



the S-RETIC simulation reported here, each M_{i_a} was replaced by an equivalent table giving M_{i_a} input-output correspondents. (Cf. Appendix 4 for an example.) These correspondents were derived from a priori specification of each overall γ_i signal configuration to one of the four modal regions in a straightforward statistical fashion. (Cf. Appendix 5 for details.)

In Figure 8, the p_i'' and p_i''' signals into the center of M_{i_a} give the modal

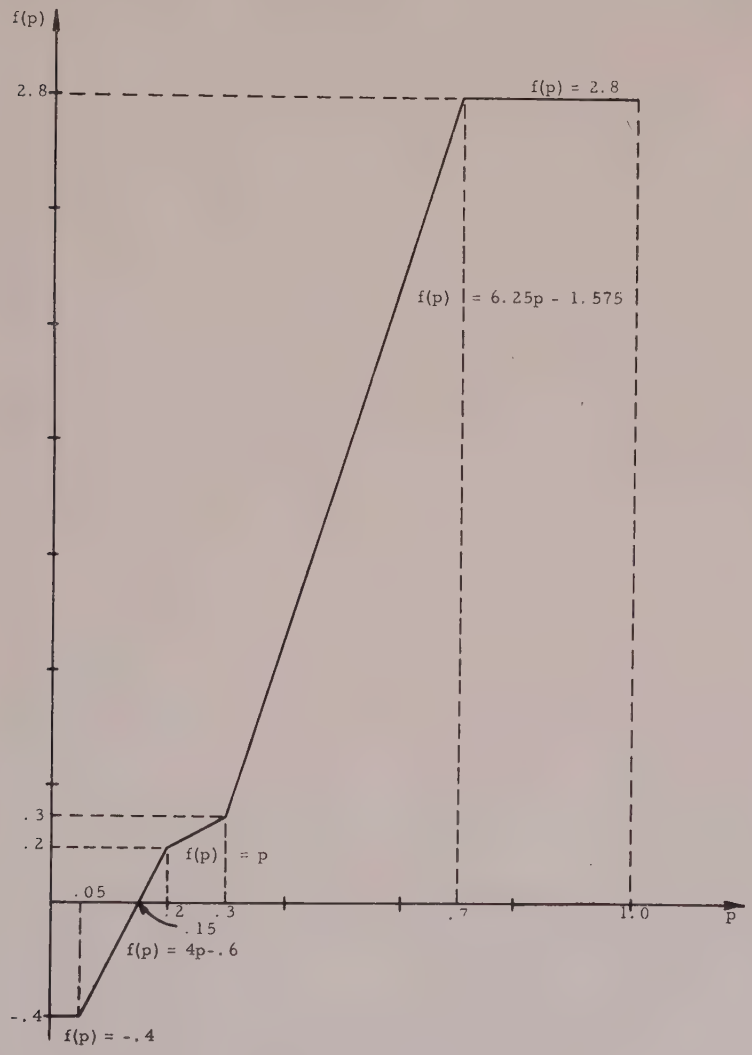


Fig. 9. The f(p) function

probabilities as computed presently by selected M_j above and below M_i, respectively. Each N box is a normalizer such that the sum of its four analog outputs equals 1. The twelve normalized p_i', p_i'', and p_i''' values are linewise operated on by a nonlinear function, f, as shown in Figure 9. Then the $\bar{p}_i$ are arrived at through C_α, C_π C_δ multiplier units and an Av averaging unit according to the formula

$$\bar{p}_i = \frac{C_\pi f(p_i') + C_\delta f(p_i'') + C_a f(p_i''')}{C_\pi + C_a + C_\delta},$$

where the factors of $C_\pi = C_{\pi 1} C_{\pi 2} Q$ and C_α and C_δ are determined as indicated below.

Since $(\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4)$ does not, in general, have all its components $\geqq 0$ and sum to 1, we always restore this vector in the h, T, h^{-1} blocks to a probability vector (p_1, p_2, p_3, p_4) for which this is true. The result is then delayed one time unit UD, and used as M_i's output to the ascending α stream, the descending δ stream, and the overall S-RETIC output bundle.

Since all of the M_i may not agree with probability 1 on which mode the overall γ_i signal configuration properly corresponds to, we specify a general output modal decisionary scheme as shown in Figure 10. There if $\sum_{i=1}^{10} \sum_{j=1}^{4} w_{ij}$ (p_j component for M_i) $> T_0$, E_0 is 1; and otherwise 0. For two half-center T_0, E_0 output systems then, the four combinations 00, 01, 10, 11 of the two output half-centers' on-off activities define four output modes for S-RETIC. Such an output criterion allows the effector system into which S-RETIC feeds to do a small amount of decoding, but does not at all push the heart of the modal computation into the effector units. (Muscle systems into which motor neurons discharge can easily perform at least this sophisticated a decoding operation.) (Actually, in the simulation reported here, we use an even simpler output criterion for the sake of convenience: if $\geqq 6$ modules indicate the jth mode with probability $\geqq 5$, S-RETIC is said to converge to the jth mode. This output convergence criterion is most reasonable if one assumes that S-RETIC always predicates its modal computations on the present system mode, K. Then the probabilities p_i out of M_j actually become P_r (transition from mode k to mode i) $\triangleq p(k \to i)$[18]

We now return to formula (1), which reflects all our design strategies. We put them under two headings: a) decoupling; and b) redundancy of

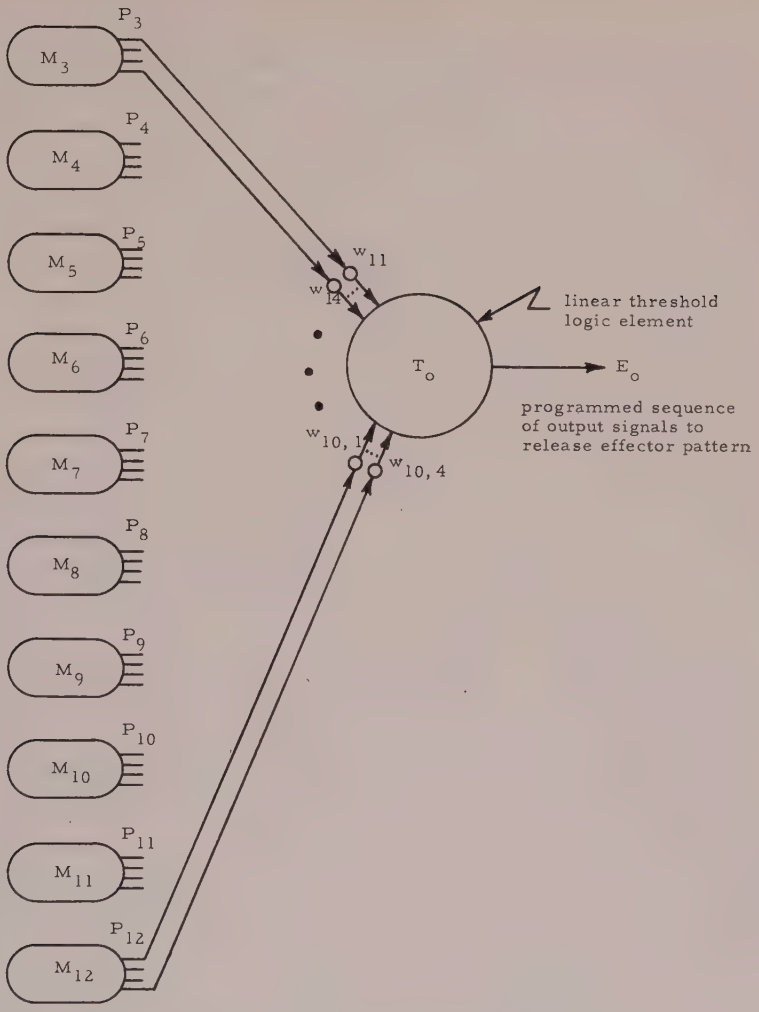


Fig. 10. S-RETIC output scheme.

potential command. The idea is to have an interconnected system of M_i such that if at any discrete time unit, t, the overall γ_i signal configuration changes and then remains fixed, the M_i will compute cooperatively from t to $t + T$, T a dozen or so, and then converge to a nearly uniform agreement on the correct output modal indication.

The b) strategy is to enable those M_i in S-RETIC which currently have the most crucial information to also have the greatest authority in recruiting other M_i of the heterarchy over to their persuasion. In S-RETIC we identify the crucial M_{i_α} as M_i whose output p_i vectors have components with values furthest from 0.25. We call such vectors "peaked." The b) strategy is realized by the f function and the C_{π_2}, C_α, and C_δ factors as follows: The f function serves to exaggerate the probabilistic modal indications of vector components which pass through them to a degree determined by the extent to which these indications depart from the neutral .25 point. Thus the f functions promote rapid overall computational convergence by amplifying the differences between the 1st, 2nd, 3rd, and 4th modal component effective gains around interconnected $M_{i_1}, M_{i_2}, \ldots, M_{i_N}, M_{i_1}$ loops in accordance with the differences between the corresponding p_{α_i}, $p_{\delta_i^j}$ modal probability values. One consequence of this, as we shall see, is that some M_i tend to pick up, or recruit, other more equivocally indicating M_j over to their modal persuasions by a logic strikingly parallel to the frequency domain logic mediated by Manifolds of coupled nonlinear oscillations.[17, 18] This is redundancy of potential command.

The C_{π_2}, C_α, and C_δ factors of each M_i are always 1, 1, and 1 if the corresponding p_{π_i}, p_{α_i}, and p_{δ_i} are not peaked (do not have component values greatly different from .25). But if at any time instant (i.e., computation clock time) any of the $f(p_{\pi_i})$, $f(p_{\alpha_i})$ and/or $f(p_{s_i})$ components are $\geqq 1$ or $\leqq 0$, C_{π_2}, C_α and/or C_δ are set to 1.5, 2 and/or 2 respectively for that time instant. In this way if any M_i has p_{π_i} and p_{α_i} vectors, nearly $=$ (.25, .25, .25, .25), say, and a $p_{\delta_i} =$ (.7, .1, .1, .1), say, p_{π_i} and p_{α_i} will not tend so much to overwhelm p_{δ_i}'s proper effect on M_i's output p_i vector. (The assymetry between C_{π_2}, C_α, and C_δ was found necessary because of the two other factors in C_π.) This is also an aspect of potential command redundancy.

The a) strategy involves both a local M_i and global S-RETIC module decoupling following certain overall γ_i signal configuration changes. The purpose of this strategy is to prevent S-RETIC from being either too trigger-happy to start completely new modal computations after slight and unimportant γ_i changes, or too prone to lock forever on output modal indications previously well intrenched at S-RETIC's output. (Monkeys and pigs have the most trigger-happy and sluggish RF's, respectively, that we know of among the higher vertebrates.) The idea is

to quench α and δ system signals after significant γ_i changes for sufficient degrees and durations to allow for proper injections of new mostly-γ_i-derived M_i output signals into the α and δ streams.

The a) strategy's global decoupling is expressed by the Q factor of C_π. If S-RETIC is converged to mode j at $t-1$ and there are any γ_i changes from $t-1$ to t, Q is increased by an amount and for a duration that is roughly proportional to the degree of entrenchment of S-RETIC in mode j at $t-1$. The values of Q are determined from the following table:

number of M_i for which the jth component of $p_i \geqq 0.65$ at $t-1$	Value of Q at t	Value of Q at $t+1$	value of Q at $t+2$	Value of Q at $t+3$
0 to 3	1.5	1.0		
4 to 7	2.0	1.5	1.0	
8 to 10	2.5	2.0	1.5	1.0

The right of the table indicates that Q is never reduced to less than 1.0. If at any time, on the basis of a new γ_i change, a new Q is computed which exceeds the value Q has decayed to from the last Q computation, then and only then is Q set to its newly computed value. The same Q is used in the (1) formula of M_3, M_4, ..., and M_{12}.

The a) strategy's local decoupling is expressed by C_{π_1}. It is determined separately for each M_i according to the following table (except that at $t = 0$, when the S-RETIC simulation begins, all modules are completely decoupled, i.e., $C_\pi = \infty$).

Number of γ_{i_j} changes into M_i from $t-1$ to t	Value of C_{π_1} at t	Value of C_{π_1} at $t+1$	Value of C_{π_1} at $t+2$	Value of C_{π_1} at $t+3$
0	2			
1	4	2		
2	6	4	2	
3 to 5	8	6	4	2

C_{π_1} is kept $\geqq 2$ rather than $\geqq 1$ in order to keep M_i's output p_i normally about equally dependent on p_{π_i} and $(p_{\alpha_i}, p_{\delta_i})$.

We now indicate the details of the h, T, and h^{-1} blocks in Figure 8. Their aggregate purpose is to restore $\bar{p}$ to a probability vector such that

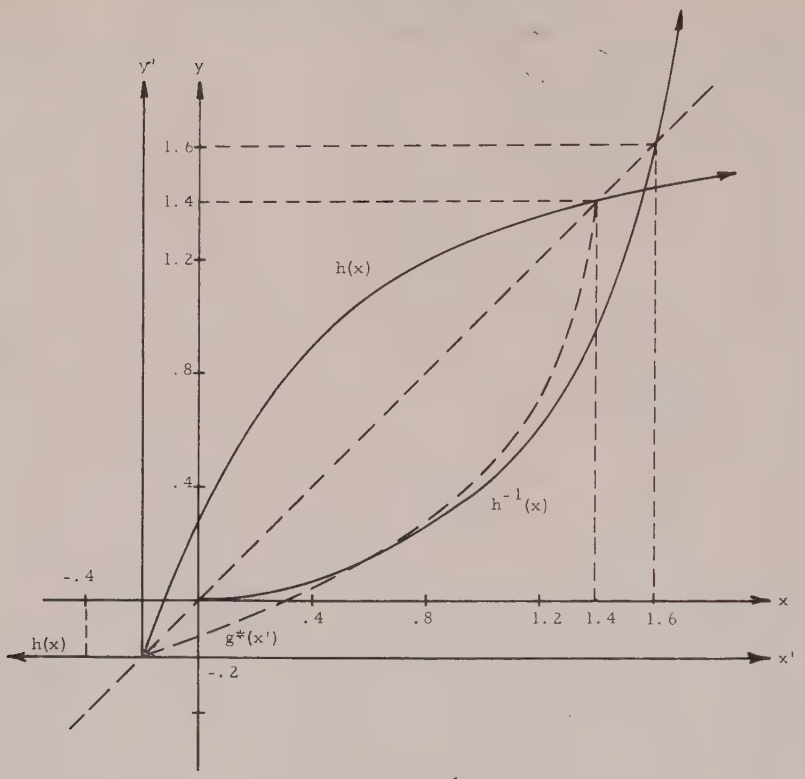


Fig. 11. $h(x)$ and $h^{-1}(x)$ curves.

the relative significance of its components is not greatly distorted in the process. Since differences between small $\bar{p}$ probability components (e.g., between .25 and .05) are generally more significant than equal differences between large $\bar{p}$ probability components (e.g., between .65 and .85), we first pass $\bar{p}$ componentwise through an exponential $h(x)$ of the form shown in Figure 11. We then add the absolute magnitude of the most negative resulting component to each $\bar{p}_i$ to get all $\bar{p}_i \geqq 0$. We finally pass the result of this componentwise through a "translated inverse" of $h(x)$, denoted $h^{-1}(x)$, as shown in Figure 11. The equations for $h(x)$ and $h^{-1}(x)$ are derived as follows:

$$y' = y + .2$$

$$x' = x + .2$$

$$y'_{y*}(x') = e^{ax'} - 1 = 1.6 = e^{a1.6} - 1$$

therefore $$a = \frac{\log e^{2.6}}{1.6}$$

therefore $$y_{g*}(x) = e^{ax}e^{.2a} - 1.2$$

therefore $$y_g(x) = \boxed{e^{ax}e^{.2a} - 1 = h^{-1}(x)}$$

$$y'_h(x) = 1.6 - y'_{g*}(1.6 - x')$$

$$y'_h(x') = 1.6 - [e^{a(1.6-x')} - 1]$$

therefore $$y_h(x) = \boxed{2.4 - e^{1.4a}e^{-ax} = h(x)}$$

The delimiting of $h(\bar{p}_i)$ by the $h(x)$ function has proved significantly useful in several S-RETIC simulations. It prevented single M_i that were very wrong in their p_i outputs from dominating the rest of the model's behavior. This and the 2.8 and $-.4$ cutoffs on the $f(p)$ curve gave S-RETIC some built-in protection against "pathological" M_i. Note that due to the h, T, h^{-1} blocks, each component of M_i's output p_i is functionally dependent on all components of p_{π_i}, p_{α_i}, and $p\delta_i$.

As a more general point, we note that RF biology recommends to our use the following M_i design strategy, which not so identically is also aligned with Leibnitz's notion of the diversified monad.[47] (Especially for $\{\sigma_i\}$, $\{\gamma_i\}$, $\{\gamma_{i_j}\}$, and $\{M_i\}$ set sizes of over 100, 1000, 100, 100, respectively, which is what we are really thinking of. We chose the small numbers 9, 42, 5, 12, because they were the smallest we thought we could get away with without completely violating the RF concept of operation.) Each M_i of S-RETIC should receive a selection of γ_i inputs which just enables it to get a good picture, or relatively high resolution view, of the signal state in a certain small portion of the $\{\sigma_1, \sigma_2, \ldots, \sigma_9\}$ bundle, but which only permits it a progressively poorer picture of the state in portions of the $\{\sigma_1, \sigma_2, \cdots, \sigma_9\}$ bundle more outlying from its "area centralis." Figure 12 depicts the essentials of this notion. Over a cross-section of the $\{\sigma_1, \ldots, \sigma_9\}$ bundle, the M_i for which the figure was drawn derives from its γ_i inputs exactly k units of information on the signal state of all the σ_i lines within each marked off area of the cross-section (A or B, for example). So A lies within the M_i's area centralis, and B its peripheral low resolution area.

The idea is to have each M_i's area centralis displaced from each other M_i's area centralis, but such that each area of the bundle's cross-section

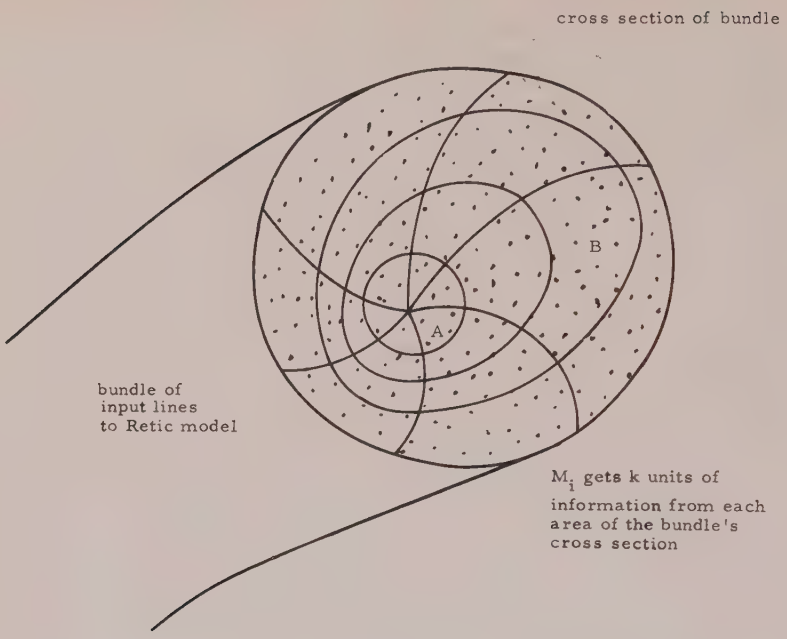


Fig. 12. Basis of pseudo-hologrammatic module design philosophy.

is at least near the area centralis of some M_i. Then each M_i knows something about all of S-RETIC's input affairs, so is diversified, but is a specialist on only a subset of them, which admits the necessary degree of overall S-RETIC acuity. It's as if each M_i had a hologrammatic view of S-RETIC's world as reflected from the ith one of a row of crazy-house mirrors, each of which greatly magnified some aspets of its reflected object, but graded off to a large demagnification of others. Then just as two eyes make one in sight, nM_i compose to one S-RETIC. An important consequence of this strategy is that one can shoot holes through S-RETIC, and what's left performs with an overall decisionary acuity that is roughly proportional to the number of good M_i it has left. Careful checking will reveal that our S-RETIC employs this strategy.

Before beginning a description of our S-RETIC simulation results, we will try to convey some more intuition on what to look for by giving the following précis of our RF* modeling problem. It is fundamentally that of appropriately matching: 1) the set of all possible correlated overall RF* model inputs; 2) the manner in which the RF* model's regional (i.e., M_i type) logic allows initial ascending and descending (i.e., α and δ type)

signal sequences to evolve through it during a modal computation; and 3) the nature of the possible sequences of changes out of the correspondents to A, C, and the S_j in Figure 6, and also the nature of their associated sequences of modal specifications.

Beyond this, it is important to emphasize a few basic organizational and operational aspects any satisfactory S-RETIC-like model, denoted Retic below, must have. First, it must have sufficient input scope with respect to the overall central nervous system (CNS) model in which it resides so that it can receive the crucial S_i information in every eventuality; and it must have sufficient computational capacity so that it can arrive at the right modal decision, regardless of whether or not conflicting and/or competitive demands appear over different A, C, and S_j systems. For it is established that real reticular formations must be able to cope with virtually every possible sufficiently correlated barrage of input signals. Second, a Retic must keep its flow of computation close enough to its input receiving areas so that all input changes can quickly exert their influence over its output and modal calculations; and the more important the input changes, the more quickly and profoundly must these influences be exerted. This is only reasonable in command and control systems for which momentary delays and wrong outputs can mean failure or annihilation. Thus, unlike some cortical systems, Retics must be pre-eminently interruptible, and not given to long periods of indecision because of excessive logical depth. Yet Retics must not be allowed to compute new modal commitments too quickly, for this would make them too vulnerable to noise and meaningless distractions (like "dreams," for example). Third, the logical design of a Retic must be extremely economical. Otherwise the heavy decisionary demands placed upon it would make it too large and to slow. Aside from a Retic's conditioning, habituation, long-term learning plasticity, and spatio-temporal coding of information, it is essentially a combinational logic circuit with very many highly correlated inputs and a small number of possible stable outputs. The main economy of any Retic organization of the general type suggested in Figure 6 stems from its repeated use of a fixed amount of modular logic throughout each modal computation. That is, logic signals are recirculated from combinations of M_i units to combinations of M_i units at successive time instants during each modal computation until an actual or approximate decisionary equilibrium is reached. Then the computation is said to be complete and the modal outputs are produced.

In general, such a scheme enables the logic of each M_i to be used at nearly full capacity throughout each modal computation, and also enables each Retic input channel to be monitored continuously. This is vastly different from the way conventional one-way-flow combinational logic nets work in engineering systems.

To recapitulate, a Retic must be a wide, shallow, anastomotic logic net, consisting of a logical heterarchy of rather tightly coupled and similar computing modules, each the equivalent of about one neuron deep.

SIMULATION RESULTS

Appendix 9 consists of a few selections from our S-RETIC simulation data. The "ith run, *j*th Σ_i, *k*th cycle" byline which appears at the top of each page refers respectively to which of our three simulation runs the associated data were obtained from, which Σ_i of the run is currently under test, and which computational pass through the M_e since the introduction of the present Σ_i the data pertain to. Recall that $C_\pi = \infty$ only on the first cycle of each run, and that our convergence criterion requires $\geqq 6$ M_i with the same p_i component $\geqq 0.5$ for a converged output indication.

Our S-RETIC simulation experience to date has revealed the following important S-RETIC behavioral features:

1) S-RETIC always converges, and usually in fewer than 15 cycles.

2) Once S-RETIC converges, it does not deconverge until a new Σ_i is introduced. This apparently would also be true for the more general Figure 10 convergence criterion.

3) S-RETIC rolls over from one output modal indication to another quite easily under strong p_{π_i} vector provocation. As this provocation becomes weaker and weaker, the rollover requires more and more computation time. This is as it should be. In fact, simulation experience has shown that an S-RETIC modal computation time is very roughly proportional to the (abductive) logical complexity (as judged by us) of the corresponding modal decision. I mean by "complexity" here the degree of p_{π_i} vector competition and conflict, and the intricacy of the pattern of modal unbalances among the p_{π_i} and p_{δ_i} vectors at the start of a new modal computation. Simulation experience has also shown that as S-RETIC's p_{π_i} vector provocation becomes less and less forceful, more and more modal decisionary hysteresis appears from one Σ_i to the next. For example, in the listing of our "Simulation Results" in AFCRL Report

No. AFCRL-66-356, dated February 1966, Run No. 1's 7th Σ_i is the same as its 9th Σ_i. But the different S-RETIC states the cycle before its first 7th Σ_i cycle and the cycle before its first 9th Σ_i cause significant differences in the 14-cycle results for these two cases. See also Run No. 2, 8th Σ_i, ibid, for a good example of such "initial condition" effects, and the consequence of perhaps excessive h, T, h^{-1}, N "clipping." (We say "perhaps" because S-RETIC needs some built-in protection against pathological M_i that might try to bully the rest of the net. We provide such protection by clipping the f function at top and bottom, and by flattening our h curve at the high end. But maybe we overdid it, or did it in too coarse a fashion.) See also Run No. 2, 3rd Σ_i, ibid, for initial condition effects.

4) S-RETIC resolves p_{π_i} conflict in a desirable fashion. (Cf. ibid, Run No. 1, 2nd Σ_i, 4th Σ_i, 5th Σ_i; Run No. 2, 2nd Σ_i, 5th Σ_i, 6th Σ_i; and Run No. 3, 3rd Σ_i.)

5) In ibid, Run No. 2, 1st Σ_i, the average of the mode 1 components of the p_{σ_i} vectors is 0.245, and that of the corresponding mode 4 components is 0.279, So S-RETIC's computation in this case went against the average. It should have, for the M_i with $p_{\pi_i} = (.7, .1, .1, .1)$ was the only one that reflected very much confidence in its modal preference, and therefore we say it is the only one that had very much "information". Run No. 3, 3rd Σ_i, of ibid, shows another example of this (it would obviously have converged after a few more cycles.)

Simulation experience has shown that regardless of how the rows of the p_{π_i} vectors table in Run No. 2, 1st Σ_i of ibid (for example) are permuted among the M_i, the result of the modal computation is essentially the same. Yet clearly there does not exist a set of 12 linear or nonlinear 4-component vector functions $F_i(p_i)$ such that $\sum_{i=3}^{12}$ always has its highest modal component the first one, regardless of how the p_{π_i} vector table rows are permuted. Thus S-RETIC is a better computer for making potential command decisions than any Figure 6—Figure 10 computer that does not have significant inter-M_i connections—even if the Figure 10 scheme is allowed to have nonlinear w_{ji}'s, or if a "$\geq n_1$ modulas $\geqq P_c$ on the same mode" convergence criterion is used.

Note that a best computer program for telling what mode any given arbitrary set of 12 p_π vectors should correspond to is simply an S-RETIC simulation.

6) Run No. 3, 1st Σ_i, of ibid shows the desirable S-RETIC feature of arbitrarily deciding on some mode not ruled out by p_{π_i} component "inhibition" when the set of p_{π_i} vectors default in producing a sufficiently positive reason for choosing one of the other modes. In several of our very first simulation trials, this desirable feature did not come through. What we needed but did not have in the undesirable cases was a mechanism to cause noise (only roundoff so far in our simulation) and gratuitous circuit particularities to force modal convergences in cases of insufficient positive reason among the p_{π_i} vectors. Our recourse was to add in at the 15th nonconvergent cycle of each modal computation that got that far without converging the multiplication of each jth component f curve by

$$G_j = \frac{4 \sum_{i=3}^{12} (j\text{th component of } p_i)}{\sum_{i=3}^{12} (\Sigma \text{ of all components of } p_i)} \triangleq \frac{4 \sum_{i=3}^{12} P_{ij}}{10}.$$

In a restricted sense, this amounted to the insertion of another strategy. This new strategy solved all our convergence problems.

7) Doubtless our simulation data could be improved upon by further adjusting S-RETIC's various design parameters. But it is not our purpose, nor should it be S-RETIC's sine qua non, to have the best possible parameter adjustments. After all, real RF's continue to function all the way from coma to convulsion.

Our S-RETIC was never upset by low level random noise injection into the simulation runs.

8) We will not dwell on it here, but anyone who carefully surveys Appendix 9 will note many other interesting, yet hardly crucial, facets of S-RETIC "decisionary motion" (in the sense of "equation of motion" for the system): e.g., the *rate* of aggregate swelling of modal components among the M_i determines loosely the degree of regenerative gain for these components between time steps; the degree of "dissociation" (or the prevalence of apparently uncooperative phenomena, amenable to, say, a simple statistical mechanical description) among the p_i is *not* particularly related to the variances among the jth components of those p_i;

the "recruitment" of M_i is *not* just the reverse of p_i inhibition of components in other M_i; etc.

Appendix 10 gives a macro level flow chart for our simulation program, which was run on the MIT Instrumentation Laboratory Honeywell 1800 Computer, and written in MAC language.

CONCLUSIONS

We can safely infer from our simulation results and the expressed cooperation of our design strategies, that the unmitigated effect of proportionately increasing the numbers of everything in our S-RETIC model would be to improve its performance in every important respect. To see this, consider distributing a wide range of M_i variants over S-RETIC's length so as to give it a near continuum of M_i area centralae displacements over its input bundle. Do this such that M_i's with most similar S-RETIC environmental acuities are furthest removed from each other physically. The result is an RF* model vastly more reliable and competent per unit length than S-RETIC, and probably only as much slower than S-RETIC as its ratio (No. of α, δ output line splits/M_i) : (No. of M_i) is lower (assuming 4 modes). All we have to insure is a sufficiently high percentage of adequately peaked p_{π_i} vectors for each S-RETIC input possibility. But this is easy—either by M_i design, by individual and aggregate M_i training, or by growth from a primitive substrate. Mutatis mutandis above, we can see that proportionately longer RF* models than S-RETIC should perform well also, but with considerably less overall head-to-foot behavioral integrity than S-RETIC has.

It appears that we have for the first time, then, supplied a paradigm for getting a family of more than two informationally-coupled automata to work together in a slightly biological fashion. To the extent that our result was inspired by the biology and is a good command and control computer in its own right, we make a claim for bionics.

Our S-RETIC is not just a glorified pattern recognition net, because it satisfies the additional temporal constraints of a real-time reticular formation model."

We think that appropriate extensions of our work might some day help us to better understand how vertebrate neural affairs are conducted in general, especially since the RF is that part of the vertebrate CNS out of which the rest has evolved. To argue the point, let us correspond S-RETIC transactions among M_{i_b}'s to those among RF neural cell bodies

and proximal dendrites, and S-RETIC transactions among M_{i_a}'s to those among RF distal dendrites. Now we hope to substantially enrich our M_{i_a}'s, and subsequently insert enough inter-M_{i_a} connections with sufficiently flexible delay-chains to enable us to realize various RF-like conditioning, habituation, extinction, and self-organizing phenomena in the augmented S-RETIC. We then hope to gain a better understanding of RF time-binding mechanisms, and possibly neural memory mechanisms in general. Our aim will be to intuit from our model, the best RF biology available at the time, and considerable empirical fumbling with RF experiments ourselves, a few detailed examples of what a really n-dimensional experiment on an RF might be like. Our prejudice is that probably nothing but a cybernetic logical approach can possibly yield an answer to this question.

The foregoing is our response so far to the challenge of the RF.

APPENDIX I

γ_k Function Table

$$\gamma_k = (\sigma_i \wedge \sigma_j) \vee (\sigma_j \wedge \sigma_l) \vee (\sigma_l \wedge \sigma_i)$$

k	i	j	l	k	i	j	l
1	1	2	8	22	2	7	9
2	1	2	9	23	3	4	8
3	1	3	8	24	3	4	9
4	1	3	9	25	3	5	8
5	1	4	8	26	3	5	9
6	1	4	9	27	3	6	8
7	1	5	8	28	3	6	9
8	1	5	9	29	3	7	8
9	1	6	8	30	3	7	9
10	1	6	9	31	4	5	8
11	1	7	8	32	4	5	9
12	1	7	9	33	4	6	8
13	2	3	8	34	4	6	9
14	2	3	9	35	4	7	8
15	2	4	8	36	4	7	9
16	2	4	9	37	5	6	8
17	2	5	8	38	5	6	9
18	2	5	9	39	5	7	8
19	2	6	8	40	5	7	9
20	2	6	9	41	6	7	8
21	2	7	8	42	6	7	9

APPENDIX 2

TABLE 1: Some σ_i; F, C. γ_i Relationships for E, with only Seven σ_i and the γ_i Comprising the Set of all 35 3-variable Symmetric Switching Functions of the Form

$$\gamma_i = (\sigma_j \wedge \sigma_k) \vee (\sigma_k \wedge \sigma_l) \wedge (\sigma_l \vee \sigma_j)$$

Number of the seven σ_i which equal 1	Number of $\sigma_1, \cdots, \sigma_7$ combinations for which this can happen	Number of the 35 γ which equal 1 in each of these combinations
1	7	0
2	21	5
3	35	13
4	35	22
5	21	30
6	7	35

APPENDIX 3

γ_j-to-M_i Connection Table

i	j	i	j	i	j	i	j
	1		6		12		3
	14		17		18		15
1	23	4	28	7	23	10	22
	32		29		37		32
(M_u)	34		33		42		39
	13		5		9		1
	24		8		13		12
2	31	5	21	8	18	11	26
	38		25		28		34
(M_L)	41		34		36		41
	4		10		7		3
	11		16		10		6
3	15	6	26	9	19	12	7
	20		27		30		19
	40		39		35		30

APPENDIX 4

Exemplary $\{\gamma_{i_i}\}$, $p_{\pi i}$ Table

$(i = 6)$

Ordered 5-tuple of γ_{i_j} Inputs to M_6	p_π Vector Corresponding to Overall Set of γ_{i_j} Input Signals			
$\gamma_{6_1}\gamma_{6_2}\gamma_{6_3}\gamma_{6_4}\gamma_{6_5}$	p_1'	p_2'	p_3'	p_4'
00000	0.273	0.152	0.333	0.242
10000	0.625	0.125	0.000	0.250
01000	0.563	0.375	0.000	0.063
11000	0.250	0.250	0.250	0.250
00100	0.250	0.350	0.150	0.250
10100	0.357	0.214	0.071	0.357
01100	0.833	0.167	0.000	0.000
11100	0.500	0.167	0.167	0.167
00010	0.000	0.308	0.385	0.308
10010	0.500	0.000	0.333	0.167
01010	0.692	0.154	0.077	0.077
11010	0.583	0.167	0.083	0.167
00110	0.125	0.333	0.292	0.250
10110	0.333	0.000	0.167	0.500
01110	0.444	0.222	0.111	0.222
11110	0.400	0.200	0.200	0.200
00001	0.000	0.040	0.440	0.520
10001	0.000	0.333	0.111	0.556
01001	0.333	0.222	0.333	0.111
11001	0.250	0.375	0.167	0.208
00101	0.083	0.083	0.417	0.417
10101	0.200	0.250	0.250	0.300
01101	0.500	0.167	0.333	0.000
11101	0.640	0.120	0.120	0.120
00011	0.000	0.000	0.750	0.250
10011	0.000	0.000	0.500	0.500
01011	0.214	0.143	0.357	0.286
11011	0.200	0.400	0.200	0.200
00111	0.333	0.000	0.333	0.333
10111	0.000	0.000	0.438	0.563
01111	0.000	0.750	0.125	0.125
11111	0.212	0.182	0.364	0.242

APPENDIX 5

The Preparation Scheme for Appendix 4.

1. First of all we constructed a chart of the form:

Σ_i Δ $\sigma_1, \sigma_2, \cdots, \sigma_9$	Mode assignment of Σ_i point	Values of γ_1 γ_2 $\cdots$ γ_{42}
0 0 $\cdots$ 0	1	0 0 $\cdots$ 0
all 9-tuples	.	.
.	.	.
.	.	.
.		

using the γ-function chart of Appendix 1, and (actually, several different) assignments of Σ_i points that we found* yielded enough "interesting and reasonable"† sets of p_π vectors for enough Σ_i to enable us to perform a meaningful simulation.

2. Then we constructed a chart of the form:

Σ_i $\sigma_1, \sigma_2, \cdots, \sigma_9$	γ5-tuple into M_1 M_2 $\cdots$ M_{12}
0 0 $\cdots$ 0	00000 00000 $\cdots$ 00000
all 9 tuples	.
.	.
.	.

3. Then for each M_i we constructed a chart of the form:

γ5-tuple into M_i	$p_1' =$ proportion of all those Σ_i corresponding to the γ5-tuple in question which are mode 1 Σ_i	$p_2' =$ proportion of all those Σ_i corresponding to the γ5-tuple in question which are mode 2 Σ_i	p_3' similarly	p_4' similarly
00000	.25	.3	.2	.25
all different 5-tuples that appear under M_i in the previous chart	.	.	.	.

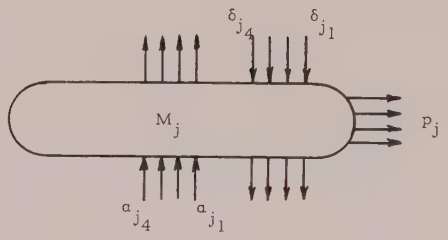
This is the Chart in Appendix 4.

* After much labor. This aspect of our simulation design is currently one of the most difficult and crucial.

† Cf. the text Section on Simulation Results.

APPENDIX 6

M_i-to-M_j Connection Table



Connections into the jth module, for $j =$	Let α_{j_k} or δ_{j_k} in the figure above come from the ith module and go into the jth module. It carries the kth component of p_i, which we denote p_{i_k}. Below we list for each αj_k and δj_k the i of the corresponding p_{i_k}. This gives the module of origin of the connection.							
	α_{j_1}	α_{j_2}	α_{j_3}	α_{j_4}	δ_{j_1}	δ_{j_2}	δ_{j_3}	δ_{j_4}
3	9	2	4	11	11	4	10	1
4	8	10	10	6	3	12	8	5
5	6	7	7	10	7	1	6	9
6	2	9	3	12	5	8	9	4
7	11	6	5	9	8	5	12	10
8	3	5	2	7	10	3	5	12
9	5	11	6	4	6	10	11	3
10	12	3	9	3	4	7	1	8
11	10	4	12	2	1	6	3	7
12	7	8	8	8	9	11	7	11

APPENDIX 7

Distribution of the $|i-j|$ in the M_i-to-M_j Connection Table.

$\lvert i-j\rvert$	Number of M_i-to-M_j connections with this $\lvert i-j\rvert$	Ideal distribution to satisfy p_π (an M_i-to-M_j connection with $\lvert i-j\rvert = k$) $\Delta p_k = \frac{C_0}{\sqrt{\lvert i-j\rvert}}$ where C_0 is a connection constant such that $\sum_{k=1}^{10} 88\, p_k = 88$, with roundoff to the nearest integers
1	18	18
2	16	15
3	12	12
4	11	10
5	9	8
6	7	7
7	6	6
8	5	5
9	3	4
10	2	3

88 = total = (8 × 10 from M_3 through M_{12}) + (4 from M_1) + (4 from M_2)

APPENDIX 8

RESULTS ON ITERATIVE NETS

1. Steady-State Memory

We have proved that for every pair of integers k and j, there is a memoryless module design such that all corresponding k or fewer module BITN's have a single-valued $\Sigma \rightarrow M$ relation at equilibrium (i.e., no variable values tending to change); but that all corresponding $(k+1)$ or more module BITN's have at least one value which has jM values that satisfy the $\Sigma \rightarrow M$ relation at equilibrium. Thus our BITN's can jump from no overall memory to any arbitrary amount of it (at least j overall memory states in the above) upon the addition of a single module at the right net size. We also showed that such changes in overall memory capacity can be effected by changing the α_n and δ_1 boundary values under the right conditions. Further, we gave an analysis algorithm* to

* A finite, completely defined computation procedure which is guaranteed to produce the correct result in a finite number of steps.

determine the critical k for any arbitrary memoryless module at which overall memory is introduced by adding a single module or by changing the BITN's boundary settings. Finally, we showed that in BITN's of the post-critical $(k + 1)$ module size, all α_j and δ_j values (there are no μ_i values here) of any one equilibrium configuration of internal BITN variable values must differ from their respective correspondents in any other such configuration, Σ constant.

2. Transients

1. In case of memoryless modules which admit only single-valued $\Sigma \rightarrow M$ relations at equilibrium in every corresponding BITN, Hennie showed that the modules can be decomposed into independent ascending and descending information processing components as shown in Figure A 8-1.

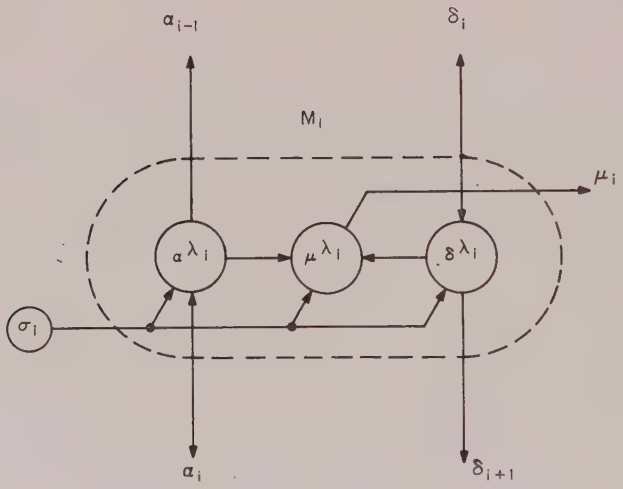


Fig. A 8-1. Hennie's decomposed modules.

i. For BITN's in which this decomposition has been performed, we derived algorithms to determine the transient character of the BITN responses to single σ_i changes from equilibrium. That is, we showed how to determine whether all transients caused by such perturbances from equilibrium must propagate all the way to a boundary of the BITN, or whether they must all be confined within

a bounded number of modules about the σ_i change in question, or what.

ii. For cases where the "i" decomposition can be performed but has not been, we showed that both the bounded and boundary transient questions of "i" are recursively unsolvable.* We also showed that the question of whether or not a single σ_i change from equilibrium can possibly cause a BITN to enter a cycle (i.e., oscillation) at some later time, given a sufficient number of BITN modules, is likewise recursively unsolvable. Thus there is no completely general algorithm to tell whether or not an arbitrary module design of the type in question will have all of its corresponding networks compute down to final variable values following a single input perturbation from equilibrium.

2. For cases where cells of the type in "1" are not present, we proved that for every integer k there is a module design such that no corresponding k or fewer module BITN could ever enter a cycle (oscillation) after being disturbed from equilibrium by a single σ_i change, but every corresponding $(k + 1)$ or more module BITN could easily do so. We also gave an especially simple proof that the results in "1, ii" above, which apply a fortiori here, can be derived very easily under the less stringent conditions of "2". Finally, we proved that there can be no essential distinction between the class of memoryless module BIT's and the class of BITN's whose modules can remember any finite but bounded amount.

* Recursive unsolvability is a difficult concept at best, so we will only attempt here to explain as much as seems important to our RF problem. Suppose we have a completely precise yes-no question, Q, about a perfectly well-defined type of BITN behavior; and suppose we wish to construct an algorithm that will determine for any arbitrary module design whether or not there is a corresponding BITN of any finite size (i.e., number of modules) for which the answer to Q is yes. Since we must allow for any finite module, we might as well construct our algorithm in stages to cover all possible modules having up to k possible (σ, α, δ) values (i.e., λ argument values), as k increases step by step over the integers. Now, to say that Q is recursively unsolvable means that we can never reach a sufficiently large k to make our algorithm complete. In other words, our algorithm must be augmented to cover essentially new kinds of cases an infinite number of times in order to be able to answer Q for every finite module design. Thus Q is a question with which nothing can be done in a totally comprehensive, finite, algorithmic way.

3. Cycling

We also showed that generally it is impossible to define a finite set of generators that implicitly describe all of the "essentially" distinct cycling (i.e., oscillatory) modes that can be supported in BITN's from the set of all BITN's that correspond to a given module design.

4. Separability

We answer negatively the question of whether a module not of the type in "B, 1" can always be decomposed into partially independent α and δ components somewhat after the fashion of Figure A 8-2. The special difficulty of this question on how precisely to frame the meaning of "partially independent." Loosely stated, our result says that in almost all cases, no one or more of the solid arrowheads in the completely general module representation shown in Figure A 8-2 can be left out without changing the function computed by many of the corresponding n-module BITN's, $n = 1, 2, 3, \ldots$

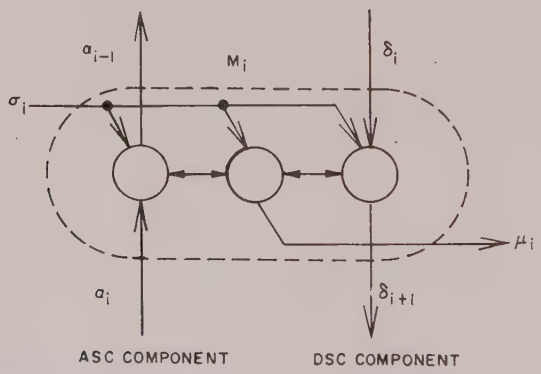


Fig. A8-2. Module decomposition for inseparability result.

The foregoing BITN results have been immensely valuable to us as theoretical guidelines for our S-RETIC problem. The unsolvability results show us that no operational definition of Figure 7 whose strict algebraic aspects reduce to recursively unsolvable BITN problems could possibly be developed with enough generality to permit the *level-by-level* refinement any really successful reticular formation theory would require

(cf. footnote on recursive unsolvability). The inseparability and cycling results tell us how the (no doubt) enormous possible efficiency and requisite variety of the S-RETIC might be concerted. They also suggest several senses in which these aspects might be removed from our Retic decision problem without appreciable effect. Our critical size results show how new dimensions of iterative net behavior (roughly in the additional half-center output sense) can arise from incremented net lengths and altered boundary conditions. And our algorithmic procedures provide conceptual tools with which to think further about more general iterative net behavior.

APPENDIX 9. SAMPLE SIMULATION RESULTS

RUN Number 2

First cycle of simulation with complete decoupling of modules (i.e., c_α, $c_\delta = 0$)

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	10101	0.200	0.250	0.250	0.300				
2	101	0.200	0.250	0.250	0.300				
3	1100	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
4	10011	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
5	10101	0.700	0.100	0.100	0.100	0.890	0.036	0.036	0.036
6	10011	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
7	10111	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
8	10001	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
9	11101	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
10	11001	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
11	11011	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278
12	11110	0.200	0.250	0.250	0.300	0.221	0.249	0.249	0.278

7th cycle, 1st Σ_i, 1st convergent cycle, mode = 1

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	10101	0.200	0.250	0.250	0.300				
2	101	0.200	0.250	0.250	0.300				
3	1100	0.250	0.250	0.250	0.250	0.465	0.179	0.090	0.264
4	10011	0.200	0.250	0.250	0.300	0.626	0.110	0.112	0.150
5	10101	0.700	0.100	0.100	0.100	0.785	0.084	0.037	0.093
6	10011	0.200	0.250	0.250	0.300	0.534	0.078	0.087	0.299
7	10111	0.200	0.250	0.250	0.300	0.584	0.121	0.048	0.245
8	10001	0.200	0.250	0.250	0.300	0.465	0.105	0.109	0.318
9	11101	0.200	0.250	0.250	0.300	0.716	0.083	0.051	0.148
10	11001	0.200	0.250	0.250	0.300	0.521	0.122	0.124	0.231
11	11011	0.200	0.250	0.250	0.300	0.386	0.141	0.093	0.378
12	11110	0.200	0.250	0.250	0.300	0.570	0.092	0.053	0.283

1st cycle, 2nd Σ_i, no convergence

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	1011	0.250	0.250	0.250	0.250				
2	1110	0.050	0.050	0.050	0.850				
3	111	0.250	0.250	0.250	0.250	0.430	0.169	0.181	0.219
4	11101	0.250	0.250	0.250	0.250	0.452	0.170	0.189	0.187
5	1001	0.250	0.250	0.250	0.250	0.441	0.197	0.171	0 189
6	11100	0.250	0.250	0.250	0.250	0.343	0.193	0.198	0.263
7	1011	0.400	0.200	0.200	0.200	0.610	0.124	0.124	0.141
8	111	0.250	0.250	0.250	0.250	0.453	0.178	0.171	0.196
9	1100	0.250	0.250	0.250	0.250	0.504	0.153	0.155	0.185
10	1110	0.250	0.250	0.250	0.250	0.438	0.174	0.105	0.101
11	110	0.250	0.250	0.250	0.250	0.344	0.184	0.183	0.287
12	1010	0.250	0.250	0.250	0.250	0.497	0.151	0.161	0.189

2nd cycle, 2nd Σ_i, no convergence

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	1011	0.250	0.250	0.250	0.250				
2	1110	0.050	0.050	0.050	0.850				
3	111	0.250	0.250	0.250	0.250	0.345	0.185	0.228	0.240
4	11101	0.250	0.250	0 250	0.250	0.353	0.203	0.214	0.227
5	1001	0.250	0.250	0.250	0.250	0.374	0.212	0.201	0.212
6	11100	0.250	0.250	0.250	0.250	0.277	0.234	0.236	0.251
7	1011	0.400	0.200	0.200	0.200	0.506	0.168	0.158	0.166
8	111	0.250	0.250	0.250	0.250	0.391	0.217	0.178	0.212
9	1100	0.250	0.250	0.250	0.250	0.343	0.214	0.221	0.219
10	1110	0.250	0.250	0.250	0.250	0.371	0.192	0.213	0.222
11	110	0.250	0.250	0.250	0.250	0.247	0.206	0.204	0.341
12	1010	0.250	0.250	0.250	0.250	0.438	0.186	0.166	0.208

8th cycle, 2nd Σ_i, 1st convergent cycle, mode = 4

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	1011	0.250	0.250	0.250	0.250				
2	1110	0.050	0.050	0.050	0.850				
3	111	0.250	0.250	0.250	0.250	0.159	0.067	0.164	0.609
4	11101	0.250	0.250	0.250	0.250	0.176	0.092	0.113	0.618
5	1001	0.250	0.250	0.250	0.250	0.133	0.169	0.100	0.595
6	11100	0.250	0.250	0.250	0.250	0.077	0.126	0.145	0.651
7	1011	0.400	0.200	0.200	0.200	0.260	0.137	0.051	0,550
8	111	0.250	0.250	0.250	0.250	0.199	0.115	0.057	0.627
9	1100	0.250	0.250	0.250	0.250	0.106	0.080	0.124	0.689
10	1110	0.250	0.250	0.250	0 250	0.201	0.079	0.148	0.570
11	110	0.250	0.250	0.250	0.250	0.182	0.095	0.089	0.632
12	1010	0.250	0.250	0.250	0.250	0.164	0.108	0.078	0.648

1st cycle, 3rd Σ_i, no convergence

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	10	0.250	0.250	0.250	0.250				
2	1011	0.200	0.250	0.250	0.300				
3	11011	**0.700**	0.100	0.100	0.100	0.815	0.037	0.036	0.109
4	10101	0.200	0.250	0.250	0.300	0.136	0.124	0.141	0.597
5	11001	0.307	0.230	0.307	0.153	0.185	0.167	0.167	0.479
6	11000	0.250	0.250	0.250	0.250	0.161	0.127	0.144	0.565
7	10001	0.150	0.150	0.150	0.550	0.066	0.062	0.056	0.814
8	10011	0.200	0.250	0.250	0.300	0.151	0.164	0.187	0.495
9	1011	0.200	0.250	0.250	0.300	0.156	0.173	0.185	0.483
10	110	0.200	0.250	0.250	0.300	0.127	0.124	0.173	0.574
11	1011	0.200	0.250	0.250	0.300	0.194	0.176	0.187	0.440
12	1001	0.200	0.250	0.250	0.300	0.159	0.157	0.154	0.528

2nd cycle, 3rd Σ_i, no convergence

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	10	0.250	0.250	0.250	0.250				
2	1011	0.200	0.250	0.250	0.300				
3	11011	0.700	0.100	0.100	0.100	0.834	0.044	0.039	0.081
4	10101	0.200	0.250	0.250	0.300	0.321	0.110	0.141	0.427
5	11001	0.307	0.230	0.307	0.153	0.147	0.134	0.129	0.587
6	11000	0.250	0.250	0.250	0.250	0.173	0.156	0.103	0.566
7	10001	0.150	0.150	0.150	0.550	0.080	0.073	0.076	0.769
8	10011	0.200	0.250	0.250	0.300	0.235	0.120	0.181	0.461
9	1011	0.200	0.250	0.250	0.300	0.200	0.217	0.217	0.365
10	110	0.200	0.250	0.250	0.300	0.201	0.080	0.294	0.423
11	1011	0.200	0.250	0.250	0.300	0.191	0.184	0.178	0.445
12	1001	0.200	0.250	0.250	0.300	0.145	0.201	0.164	0.488

3rd cycle, 3rd Σ_i, 1st cycle of convergence, mode = 4

		p_π vectors				p_i vectors			
Module	Template	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	10	0.250	0.250	0.250	0.250				
2	1011	0.200	0.250	0.250	0.300				
3	11011	0.700	0.100	0.100	0.100	0.779	0.044	0.064	0.111
4	10101	0.200	0.250	0.250	0.300	0.347	0.078	0.150	0.423
5	11001	0.307	0.230	0.307	0.153	0.166	0.161	0.125	0.547
6	11000	0.250	0.250	0.250	0.250	0.173	0.159	0.117	0.549
7	10001	0.150	0.150	0.150	0.550	0.105	0.082	0.082	0.729
8	10011	0.200	0.250	0.250	0.300	0.298	0.054	0.131	0.515
9	1011	0.200	0.250	0.250	0.300	0.206	0.209	0.203	0.380
10	110	0.200	0.250	0.250	0.300	0.229	0.070	0.368	0.330
11	1011	0.200	0.250	0.250	0.300	0.196	0.153	0.146	0.502
12	1001	0.200	0.250	0.250	0.300	0.125	0.157	0.129	0.587

REFERENCES

Neurophysiology

1. MacLean, P. (1958) "The Limbic System with Respect to Self-Preservation and the Preservation of the Species," *J. Nervous and Mental Disease* **127**, 1–11.
2. Jasper, H. H., and Procter, L. D., *et al.* (1958). *Reticular Formation of the Brain*, Little, Brown and Co., Boston.
3. Scheibel, M. and A. (1962). "On Neural Mechanisms for Self-Knowledge and Command," Mitre Report SS-3, First Congress on the Information Systems Sciences, Mitre Corp., Boston, Mass.
4. Nauta, W. J. (1963). "Central Nervous Organization and the Endocrine Motor System," *Advances in Neuroendocrinology*, University of Illinois Press, Urbana, Illinois.
5. Amassian, V., *et al.* (1961). "Patterns of Activity of Simultaneously Recorded Neurons in Midbrain Reticular Formation," *Annals New York Acad. Sci.* **89**, 883–895.
6. Wolstenholme, G., and O'Connor, C., eds. (1958). *Neurological Basis of Behavior*, Little, Brown and Co., Boston.
   (4) "The Behavior of Chronically Decerebrate Cats," Bard, P., and Macht, M., 55–70.
   (11) "Some Basic Mechanisms of the Translation of Bodily Needs into Behavior," Dell, P., 187–203.
   (15) "Some Aspects of the Neurophysiological Basis of Conditioned Reflexes and Behavior," Gastaut, H., 255–276.

7. Rosenblith, W., ed. (1961), *Sensory Communication*, MIT Press, Cambridge, Mass.
(26) "Reticular Mechanisms of Sensory Control," Hernandez-Peon, R., 497–520.
8. Jasper, H. H., and Phanopeat, P. Personal communication on spike and dome waves in cortex, our interpretation.
9. Wall, P. D. Personal communication.
10. Doty, R., and Bosma, J. (1956). "An Electromyographic Analysis of Reflex Deglutition," *J. Neurophysiol.* **19**, 44–60.
11. Field, J. (ed.-in-chief) (1959). *Handbook of Physiology*, Section 1: "Neurophysiology, Vol. 1," American Physiological Soc., Washington, D. C.
12. Doty, R. (1961). "The Role of Subcortical Structures in Conditioned Reflexes," *Annals New York Acad. Sci.* **92**, 939–945.
13. Riss, W. (1962). "The Evolution of the CNS in Relation to the Evolution of Behavior, I. The Theoretical Evolution of Prolonged Responses to Somatic Stimulation," *Trans. New York Acad. Sci., Ser. II* **24**, 6, 630–641.
14. Scalia, F. (1962). "The Evolution of the CNS in Relation to the Evolution of Behavior, II. A Hypothesis on Spinal Cord Activity Waves in Relation to a Theory of the Evolution of Locomotion," *Trans. New York Acad. Sci., Ser. II*
15. Krieg, W. J. S. (1963). *A Polychrome Atlas of the Brain Stem*, Brain Books, Evanston, Illinois.
16. Krieg, W. J. S. (1963). *Brain Mechanisms in Diachrome*, Brain Books, Evanston, Illinois.

Nonlinear Oscillations and Neural Nets

17. Wiener, N. (1958). *Nonlinear Problems in Random Theory*, MIT Press, Cambridge, Massachusetts, and John Wiley and Sons, New York.
18. Pringle, J. W. S. (1951). "On the Parallel Between Learning and Evolution," *Behavior* **3**, 174, 90–110.
19. Beurle, R. L. (1956). "Properties of a Mass of Cells Capable of Regenerating Pulses," *Phil. Trans. Royal Soc., Ser. B* **240**, 55–94.
20. Smith, D. R., and Davidson, C. H. (1962). "Maintained Activity in Neural Nets, *J. Assoc. Comp. Mach.* **9**, 268–279.
21. Block, S. H. (1963). "A Neural Net for Adaptive Behavior," Rand Report RM-3868-PR, Rand Corp., Santa Monica, California.
22. Minorsky, N. (1962). *Nonlinear Oscillations*, Van Nostrand, Princeton, New Jersey.
23. Cesari, L., LaSalle, J., and Lefschetz, S. (eds.), (1950–1960). "Contributions to the Theory of Nonlinear Oscillations," Vol. 1–5, Annals of Math. Studies, Princeton University Press, Princeton, New Jersey.
24. LaSalle, J., and Lefschetz, S. (1961). *Stability by Liapunov's Direct Method*, Academic Press, New York.
25. Pontryagin, L. S., *et al.* (1962). *The Mathematical Theory of Optimal Processes*, Interscience, New York.
26. Farley, B. (1963). Films on artificial neural net operation, shown at various cybernetics conferences. Films at MIT Lincoln Laboratories, Bedford, Massachusetts.
27. Hale, J. (1963). *Oscillations in Nonlinear Systems*, McGraw-Hill, New York.

Automata Theory and Information Theory

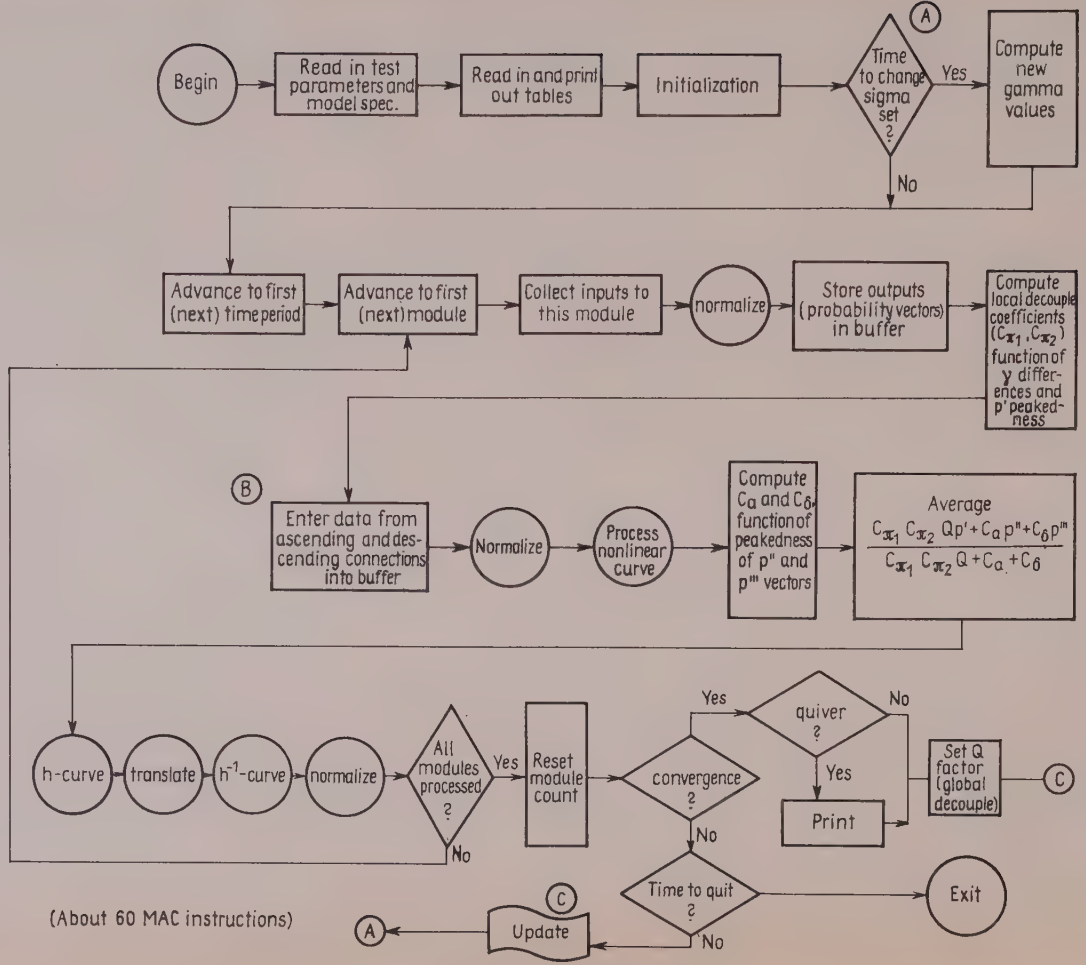
28. Davis, M. (1958). *Computability and Unsolvability*, McGraw-Hill, New York.
29. Kemeny, J. G., and Snell, J. L. (1960). *Finite Markov Chains*, Van Nostrand, New York.
30. Hennie, F. (1961), *Iterative Arrays of Logical Circuits*, MIT Press, Cambridge, Massachusetts, and Wiley, New York.
31. Kilmer, W. (1961). "Transient Behavior in Iterative Combinational Switching Networks," Proc. AIEE Symp. Switching Theory and Logical Design, 114–128.
32. Schutzenberger, M. P. (1961). "A Remark on Finite Transducers," *Information and Control* **4**, 185–196.
33. Shannon, C. E. (1961). "Two-Way Communication Channels," Fourth Berkeley Symp. Probability and Stat., J. Neyman, ed., Vol. 1, Univ. of California Press, Berkeley, 611–644.
34. Wolfowitz, J. (1961). *Coding Theorems of Information Theory*, Prentice-Hall, Englewood Cliffs, New Jersey.
35. Kilmer, W. (1962). "Iterative Switching Networks Composed of Combinational Cells," *IRE Trans. Electronic Computers* **EC-11**, 123-131.
36. Kilmer, W. (1962). "On Cycling Behavior in 1-Dimensional Bilateral Iterative Networks," Montana State College Electronics Research Laboratory Report.
37. Winograd, S. (1962). "Bounded-Transient Automata," Proc. AIEE Symp. Switching Theory and Logical Design.
38. Carlyle, J. (1963). "Reduced Forms for Stochastic Sequential Machines," *J. of Math. Anal. and Applica.* **7**, 167–175.
39. Jelinek, F. (1963). "Coding for Discrete Memoryless Two-Way Channels," Ph. D. Dissertation, Dept. of Electrical Engineering, MIT, Cambridge, Massachusetts. Also subsequent derived publications in several technical journals.
40. Kilmer, W. (1963). "On Dynamic Switching in One-Dimensional Iterative Logic Networks," *Information and Control* **6**, 399–415.
41. Rabin, M. (1963). "Probabilistik Automata," *Information and Control* **6**, 230–245.
42. Winograd, S., and Cowan, J. (1963). *Reliable Computation in the Presence of Noise*, MIT Press, Cambridge, Massachusetts.
43. Winograd, S. (1963). "Redundancy and Complexity of Logical Elements," *Information and Control* **6**, 177–194.
44. Carlyle, J. (1964). "On the External Probability Structure of Finite-State Channels," Department of Engineering Report, UCLA, Los Angeles, California.
45. Kilmer, W. (1964). "Topics in the Theory of One-Dimensional Iterative Networks," *Information and Control* **7**, March.

General

46. Craik, K. J. W. (1952). *The Nature of Explanation*, Cambridge University Press, New York.
47. Leibnitz, *Selections*, ed. by Phillip P. Wiener, Scribner's, New York (1951).
48. Turing, A. M. (1936–1937). "On Computable Numbers with an Application to the Entscheidungsproblem," *Proc. London Math. Soc., Ser.* 2, **42**, 230–265. See also correction in **43**, 544–546.

49. Shannon, C. E. (1949). *Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois.
50. Scholl, D. A. (1956). *The Organization of the Cerebral Cortex*, Methuen and Co., Ltd., London.
51. Rabin, M., and Scott, D. (1959). "Finite Automata and Their Decision Problems," *IBM J. Res. Develop.* **3**, 114–125.
52. Ashby, W. R. (1960). *Design for a Brain*, Wiley, New York.
53. Minsky, M., and Selfridge, O. (1961). "Learning in Random Nets," Fourth London Symp. on Information Theory, Butterworths, London, 335–347.
54. Pask, G. (1961). *An Approach to Cybernetics*, Harper Bros., New York.
55. Wiener, N. (1961). *Cybernetics*, 2nd ed., MIT Press, Cambridge, Massachusetts.
56. Shannon, C. E., and McCarthy, J., eds. (1956). *Automata Studies*, Princeton University Press, Princeton, New Jersey.
57. *Recent Developments in Information and Decision Processes*, ed. Machol and Gray, Macmillan, New York (1962).
    (1) Wiener, N. "The Mathematics of Self-Organizing Systems, Recent Developments."
    (2) Robbins, H., and Samuel, E. "Testing Statistical Hypotheses—The 'Compound Approach'."
58. Goodwin, B. C. (1963). *Temporal Organization in Cells*, Academic Press, New York.
59. Moruzzi, G. (1963). "The Physiology of Sleep," Endeavour, London.
60. Arbib, M. (1964). *Brains, Machines, and Mathematics*, McGraw-Hill, New York.
61. Young, J. Z., *A Model of the Brain, Oxford*, 1964.
62. McCulloch, W. S. *Embodiments of Mind*, MIT Press, 1965.
63. Nilsson, Nils J., *Learning Machines, Foundations of Trainable Pattern Classifying Systems*, McGraw-Hill, New York, 1965.
64. Ebert, James D., *Interacting Systems in Development*, Holt, Rinehart, and Winston, New York, 1965
65. Gose, E., "An Adaptive Network for Producing Real Functions of Binary Inputs," *Information and Control*, **8**, 111-123 (1965).
66. Minsky, M., *Matter, Mind and Models*, Proc. IFIPS Congress, May 1965, Spartan Books, Washington, D. C.
67. Gabor, D., "Hologram Pattern Recognition," *Nature*, Oct. 30, 1965.
68. *The Nature of Psychology*, selections from K. J. W. Craik, ed. by S. L. Sherwood, Cambridge University Press, 1966.
69. Heisenberg, W. *Physics and Philosophy*, Harper and Bros., New York, 1958.

APPENDIX 10



Appendix 10. Flow chart of program

ROBERTO MORENO-DIAZ

Instrumentation Laboratory

Massachusetts Institute of Technology

Cambridge, Massachusetts

*An Analytical Model of the "Bug Detector" Ganglion Cell in the Frog's Retina**

INTRODUCTION

Since Lettvin *et al.*, (Ref. 1) and Maturana *et al.* (Ref. 2) published their measurements of signals in the optic fibers of the frog, considerable effort (Refs. 3 and 4) has been given to developing models that could account for the properties they found. Such models are of importance to engineers and neurophysiologists for two reasons. First, models provide clues on which to base advanced and versatile engineering systems. Secondly, models provide a basis of thought consistent with reported neurophysiological findings. Such a basis could be useful to neurophysiologists in interrelating experimental results.

Lettvin, Maturana, and co-workers have distinguished four major groups of retinal ganglion cell which report to the tectum. These have been designated as follows: Group 1—edge detector, Group 2—bug detector, Group 3—dimming detector, and Group 4—event detector ganglion cells. Of these, relatively simple explanations can be given (Refs. 5 and 6) to the operations of the Group 1, 3, and 4 ganglion cells. The Group 2 or bug detector ganglion cell, however, is a more intricate and the most exciting cell to model because it is sensitive to small dark convex objects which move centripetally with respect to the responsive retinal field (RRF) of this cell. In essence, it is the most specialized pattern recognition cell of the frog's retina. Gaze and Jacobson (Ref. 7) suggest that the Group 2 operation may be due to the existence of an excitatory area surrounded by an inhibitory ring, such that large objects will cause in-

* This paper was sponsored by the Biosciences Division of the National Aeronautics and Space Administration, NSR 22-009-138.

hibition, whereas small objects will be detected by the cell. However, Grusser *et al.* (Ref. 8) reported that in their experiments no special construction of the receptive field, with respect to inhibitory or excitatory areas, was found. Grusser *et al.* (Ref. 9) pointed out that patterns moved outside the RRF can have an inhibitory effect on the response elicited by a small moving object inside the RRF, i.e., the inhibitory effect of the supposed ring appears only if the object moves. Evidence (Ref. 8) has been given suggesting that these cells are directionally sensitive, although the argument is not definitive.

We present an analytical model consistent with the findings of the aforementioned authors. In structuring our model, we follow the anatomy of the Group 2 ganglion cell as understood by Lettvin *et al.* They identified (Ref. 10) the Group 2 ganglion cell as a multilevel E-shaped neuron from Ramon y Cajal's drawings. Accordingly, we distribute cell computations in three layers. Those computations are, in general, compatible with commonly accepted neural processes. It has not been necessary to postulate an exclusively inhibitory ring, although the model cell receives information from an area wider than the responsive retinal field. Some cellular properties appear as consequences of the model structure. As a consequence, it is not necessary to make ad hoc hypotheses to explain each of them.

The operation of the model can be summarized as follows. First, a convex function Φ, depending upon the penetration of an object into the responsive retinal field (RRF) is defined. It is only significant when the object moves centripetally. Secondly, a similar function, Ψ, is defined, which is dependent on the size of the object, being a maximum for one particular size. The coincidence of both is computed by the product $\Phi\Psi$. Thirdly, an inhibitory effect, X, is defined which acts upon the function $\Phi\Psi$. The inhibition is large for bright objects and small for dark objects. As a result, an activity function, Ω, is obtained. The pulse repetition frequency of the cell is assumed to be proportional to Ω.

THE MODEL

We assume that, for the purpose of the Group 2 ganglion cell operation, the photoreceptors are connected to two different types of bipolar cells, the outputs of which are pulses of width δt and amplitude r. Each bipolar

cell performs a different operation on the retinal image. Let us call $n_I(t)$ and $n_{II}(t)$ the number of bipolar cells (belonging to types I or II) that fire at time t as a response to a changing image on the retina.

We postulate:

a. $n_I(t)$ is proportional to the total length of the edges, coincident with a local dimming, in the retinal image.

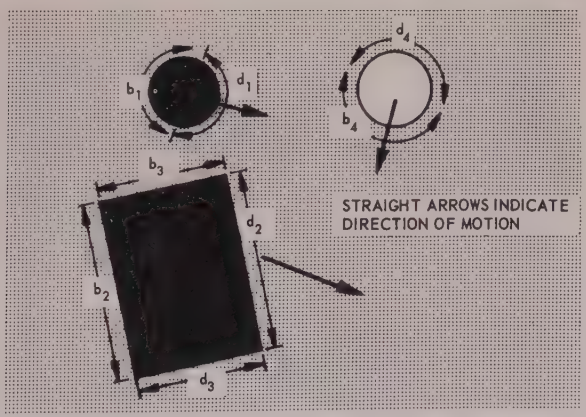
b. $n_{II}(t)$ is proportional to the total length of the edges, coincident with local brightening, in the retinal image.*

Type I bipolar cells we term contrast-dimming detectors, whereas type II bipolar cells we term contrast-brightening detectors. In both cases, spatio temporal changes of the illumination on the photoreceptors feeding each bipolar cell are necessary to fire types I and II bipolar cells.

Figure 1 illustrates $n_I(t)$ and $n_{II}(t)$ for several bright and dark moving objects.

Consider one group 2 ganglion cell. It receives signals from type I and II bipolar cells, and processes these signals in three computation layers (Fig. 2).

In Layer 1, pulses emanating from type I bipolar cells are collected over a circular area equal to the RRF of the ganglion cell. Each pulse



$$n_I(t) \propto d_1 + d_2 + d_3 + d_4$$

$$n_{II}(t) \propto b_1 + b_2 + b_3 + b_4$$

Fig. 1. Equivalences of $n_I(t)$ and $n_{II}(t)$ for several moving objects.

* In Ref. 11, a model is described in which photoreceptors and bipolar cells perform in a manner similar to that postulated here.

originates a signal level that is maintained for a time Δt.* We may regard this as a short term memory. Let $n_1(t)$ be the number of pulses impinging on Layer 1 at time t. $N_1(t)$, number of existing signal levels at time t, will be equal to the total number of pulses that have reached the Layer 1 in the previous time interval $(t - \Delta t, t)$. Note that $N_1(t)$ is proportional to the area that has been scanned by dimming, within the

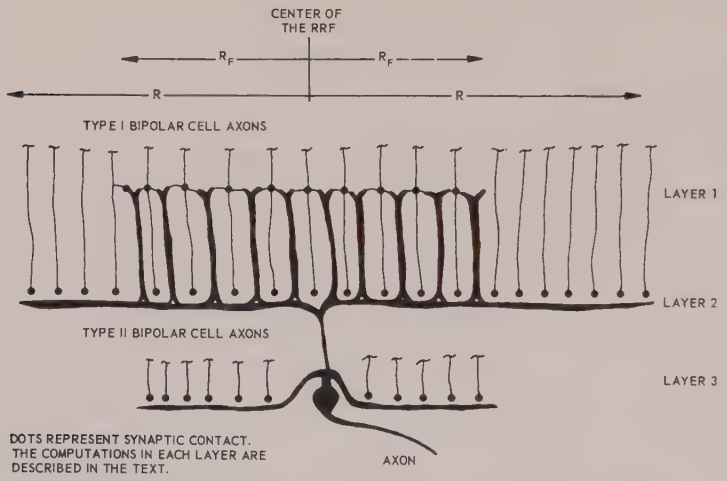


Fig. 2. Diagram of a section of the Group 2 ganglion cell model.

RRF, due to a moving object. Further, each existing signal level at time t is affected by divisional inhibition,† by the ensemble of incoming pulses at time t. Thus, $N_1(t)$ new signal levels are originated, each of them having an identical value, α, defined by:

$$\alpha = a/[1 + bn_1(t)] \tag{1}$$

where a and b are constants. For $bn_1(t) \gg 1$, Eq. (1) becomes

$$\alpha = K/[n_1(t)] \tag{2}$$

where $K = a/b$.

* Δt is made equal to the transit time across the RRF of the slowest object to be detected by the cell.

† Divisional Inhibition, proposed by Lettvin (Ref. 12) can be formulated as follows: Let E and I be the excitatory and inhibitory signals, respectively, expressed as dimensionless numbers. After inhibition, the resulting signal is $E/(1 + I)$; $I \gg 1$, $E/(1 + I) \simeq E/I$.

In Layer 2, three operations are distinguished. First, the $N_1(t)$ signal levels from Layer 1, each of them having a value $K/n_1(t)$, interact in a manner such that a signal, $\Phi[N_1(t)/n_1(t)]$, is obtained, which has the convex shape shown in Fig. 3. $\Phi[N_1(t)/n_1(t)]$ is maximum for a particular

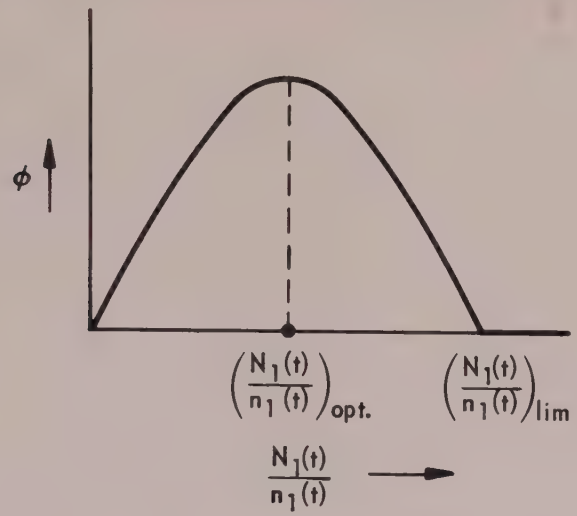


Fig. 3. Shape of the curve $\Phi\left(\frac{N_1(t)}{n_1(t)}\right)$.

value $[N_1(t)/n_1(t)]_{opt}$ and is zero for $N_1(t) = 0$ and for $[N_1(t)/n_1(t)] \geqq [N_1(t)/n_1(t)]_{limit}$. The ratio

$$[N_1(t)/n_1(t)] = \frac{\text{area scanned by contrast-dimming in the RRF}}{\text{length of contrast-dimming in the RRF}}$$

provides a measure of the penetration of a round-shaped dark object moving into RRF. By appropriately choosing the value of constants in the function Φ, we can make Φ significant only when the image moves centripetally across the RRF.

A particular function, Φ, containing the previously mentioned characteristics can be obtained by assuming that the active level signals interact by processes commonly accepted in neurophysiology; namely, lateral divisional inhibition, adaptation, and spatial summation. To demonstrate, let us assume that each afferent active line is laterally inhibited by all

the others. If the signal level of each line is $K/n_1(t)$, the total inhibition upon each line is:

$$\xi = k[N_1(t) - 1] \cdot K/n_1(t) \tag{3}$$

where $N_1(t)$ is the total number of active lines and k is a constant.

If $N_1(t) \gg 1$, Eq. (3) becomes

$$\xi = K_1[N_1(t)/n_1(t)] \tag{4}$$

where $K_1 = kK$.

As a result of divisional inhibition*, the signal A_C in each active line becomes:

$$A_C = [K/n_1(t)]/[1 + K_1[N_1(t)/n_1(t)]] \tag{5}$$

If we assume that each active line is adaptive, i.e., its threshold, θ, increases proportionally to the incoming signal,

$$\theta = A[K/n_1(t)] \tag{6}$$

where A is a constant. This is a form of linear adaptation.

From Eqs. (5) and (6), the net signal in each line is

$$A_C - \theta = \frac{K/n_1(t)}{1 + K_1[N_1(t)/n_1(t)]} - A[K/n_1(t)] \tag{7}$$

By spatial summation over all the $N_1(t)$ active lines at time t, we have

$$\Phi\left[\frac{N_1(t)}{n_1(t)}\right] = \sum_{\substack{\text{all active} \\ \text{lines}}} A_C - \theta = K\left[\frac{N_1(t)/n_1(t)}{1 + K_1[N_1(t)/n_1(t)}\right] - A[N_1(t)/n_1(t)] \tag{8}$$

Equation (8) is plotted in Figure 4 for $K_1 = 1/0.22\ R_F$ and $A = 0.25$. R_F is the number of contrast-dimming bipolars contained in one radius of the *RRF*. K_1 and A have been chosen in a manner such that Φ is significant only for a round-shaped dark object when it moves centripetally across the RRF. Again, note that the essential feature of $\Phi[N_1(t)/n_1(t)]$ is its convex shape as shown in Fig. 3. There exists an infinite number of functions with these characteristics. Among them, Eq. (8) is an example, which is compatible with neurophysiological facts.†

* See footnote on page 484.

† A similar shaped curve is obtained in Ref. 11 by assuming non-linear divisional inhibition of the type E/e^I, where E is the excitation and I the inhibition.

The second operation in Layer 2 is performed on afferent pulses from type I bipolar cells over a circular area of radius R, which is wider than the *RRF*. Let $n_2(t)$ be the number of incoming pulses collected at time t over this area. $n_2(t)$ is proportional to the total length of contrast-dimming within the circular area of radius R.

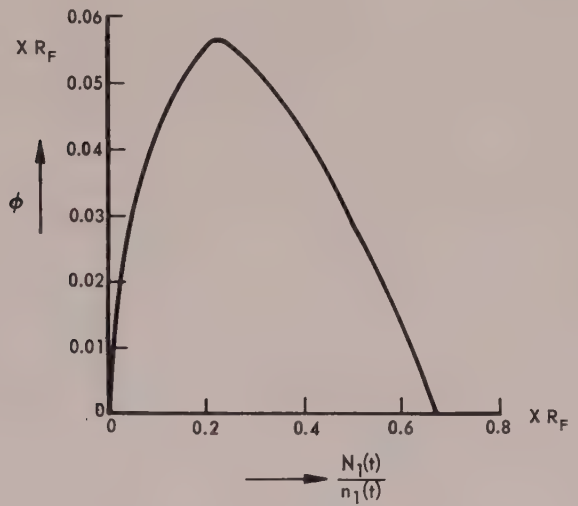


Fig. 4. Example of function $\Phi\left(\frac{N_1(t)}{n_1(t)}\right)$.

The $n_2(t)$ pulses interact in a manner such that a function, $\Psi[n_2(t)]$, is obtained which is similar to Φ. Thus, $\Psi[n_2(t)]$ is maximum for $n_2(t)_{\text{optimum}}$ and it is zero for $n_2(t) = 0$ and for $n_2(t) \geqq n_2(t)_{\text{limit}}$. Again, constants in $\Psi[n_2(t)]$ can be computed to adjust $n_2(t)_{\text{opt}}$ and $n_2(t)_{\text{limit}}$ to the experimental results.

As an example, we assume again that the $n_2(t)$ pulses interact by lateral divisional inhibition and that there exists adaptation and spatial summation. Ψ can be expressed as

$$\Psi[n_2(t)] = K'\left[\frac{n_2(t)}{1 + K_1' n_2(t)} - Bn_2(t)\right] \tag{9}$$

where K', K_1', and B are constants.

Results (Refs. 1, 2, and 9) showed that when experimenting with dark discs, the ganglion cell output is maximum for a disc of radius $\approx R_F/2$

and is zero for discs of radii larger than R_F. Using these findings, we then compute $K_1' = 3/\pi R_F$ and $B = 0.25$. The radius, R, of wider circular area is estimated as being $R = 3.2\ R_F$ by using the criterion (Refs. 1 and 2) that a straight band wider than R_F does not produce a response. The results of Gaze and Jacobson (Ref. 6) appear then as a consequence of this restriction.

The third Layer 2 operation is a multiplication of the functions Φ and Ψ. We do not have enough neurophysiological evidence to support this assumption, although we will tacitly assume its validity.* Thus the activity function

$$\Phi[N_1(t)/n_1(t)]\,\Psi[n_2(t)]$$

is generated and we consider this as the output of Layer 2.

In Layer 3, the outputs from the type II bipolar cells are of concern. These afferent pulses are collected over the *RRF*, and they generate signal levels that are maintained for a time Δt. Let $N_3(t)$ be the number of these levels at time t. $N_3(t)$ is then proportional to the area that has been scanned by contrast-brightening within the RRF in the time interval $(t - \Delta t, t)$.

The $N_3(t)$ signal levels are spatially summed and generate a signal

$$X[N_3(t)] = CN_3(t) \tag{10}$$

where C is a constant. This signal affects, by divisional inhibition, the output from Layer 2. Therefore, we have

$$\Omega = \frac{\Phi\Psi}{1 + X} \tag{11}$$

We assume that the pulse frequency, f, of the ganglion cell is proportional to Ω.

Thus,
$$f = f_0\Omega = f_0(\Phi\Psi/1 + X) \tag{12}$$

where f_0 is a constant.

The value of constant C in Eq. (10) can be chosen to achieve a cell output for bright objects suitably less than that for dark objects.

* An explanation of this hypothesis and the shapes of the curves Φ and Ψ may be given in terms of probability. We will discuss this point of view in a later paper, since it could be applied to the description of any nerve cell.

DISCUSSION

The performance of the model can be derived from Eq. (12). For purposes of illustration, we will assume that Φ and Ψ are given by Eqs. (8) and (9), respectively. Constant C is fixed by the arbitrary condition, that the maximum response for a bright disc RRF/8 wide is 1/10 of that which would have resulted without the inhibition produced by X. This condition gives $C = 60\ R_F^{-2}$. The maximum pulse frequency is approximately 40 pulses $\cdot$ sec^{-1} (Ref. 8), which fixes $K = K' = 1$ and $f_0 = 300/R_F^2$ sec. The pulse frequency of the ganglion cell model is plotted versus the penetration of the leading edge of the object crossing the RRF in Figure 5. Curves (a), (b), and (c) are for dark discs of radii 0.125 R_F, 0.5 R_F, and 0.75 R_F, respectively. Curves (d), (e), and (f) are for bright discs of radii 0.125 R_F, 0.5 R_F, and 0.75 R_F, respectively.

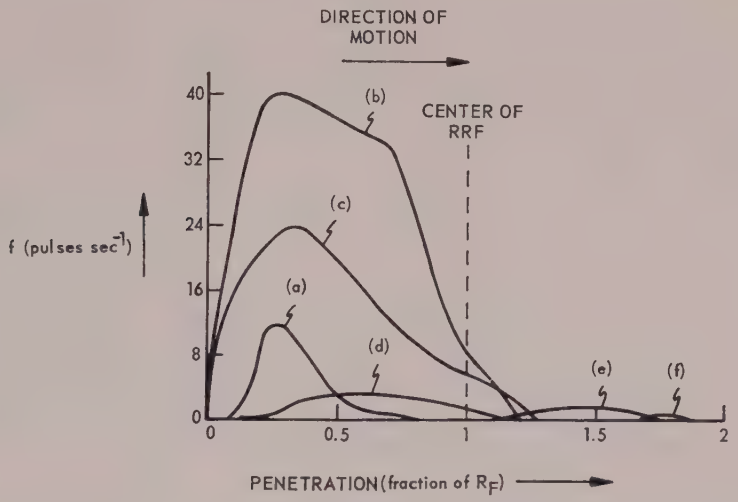


Fig. 5. Output of the model versus penetration of different objects into the RRF.

The following are some of the consequences that may be derived from the characteristics of the model.

a) No response occurs to a general change in illumination. (In agreement with Refs. 1, 2, and 10).

b) A corner may produce a response. (In agreement with Ref. 2).

c) Several small object images moving simultaneously in the RRF may produce either very small or null responses. This is in accord with Lettvin's observation quoted in Refs. 1 and 2.
d) No evidence of the annulus surrounding the RRF can be detected by fixed dark or light spots with a simultaneous moving testing spot. This is in agreement with Ref. 8.
e) If the spots in the surrounding ring move, the response to the testing spot may be either increased or decreased, depending on the size of the spots. This is in agreement with Ref. 9.

The Group 2 ganglion cells respond for approximately one second after an object has entered and stops in the RRF. This response is erased by a corresponding step to darkness (Refs. 1, 2, and 10). However, in the model, the response disappears when the object is stopped within the RRF. The persistence of the response might be explained by feedback from the tectum, as suggested by Lettvin *et al.* (Ref. 10). If we assume that tectal feedback acts on the type I of the bipolar cells, and that the feedback has the same effect as that of dimming, the ganglion cell will provide an output as long as feedback exists. This can be formulated in the following manner. Boolean magnitudes $C(t)$, $D(t)$, $F(t)$, and $B_I(t)$ are defined as follows:

$C(t)$ is 1 if contrast exists, at time t, in the field of a type I bipolar cell
0 if there is no contrast
$D(t)$ is 1 if dimming occurs, at time t, in the field of a type I bipolar cell
0 if there is no dimming
$F(t)$ is 1 if there is feedback from tectum, at time t, on a type I bipolar cell
0 if there is no feedback
$B_I(t)$ is 1 if the type I bipolar cell fires at time t
0 if it does not fire

The condition for type I bipolar cell firing is then the boolean expression:

$$B_I(t) = C(t) \cdot [D(t) + F(t)] \tag{14}$$

Feedback from the tectum must be maintained for approximately 1 sec after local dimming has disappeared. This feedback might be provided by the newness cells of the tectum (Ref. 10).

REFERENCES

1. Lettvin, J. Y., *et al.* "What the Frog's Eye Tells the Frog's Brain," *Proc. I.R.E.*, **47**, 1940–1959, November 1959.
2. Maturana, H. R., *et al.* "Anatomy and Physiology of Vision in the Frog (Rana Pipiens)," *J. of General Physiology*, **43**, 129–175, July 1960.
3. Herscher, M. B., and Kelley, T. P. "Functional Electronic Model of the Frog's Retina," *I.E.E.E.* Transactions, MIL–7, 98–103, 1963.
4. Sutro, L. L., *et al.* "1964 to September 1965 Advanced Sensor and Control System Studies," R-519, Instrumentation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 1965.
5. Sutro, L. L., *et al.* "1963 Advanced Sensor Investigations," R-470, Instrumentation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 1964.
6. Shypperheyn, J. J. "Contrast Detection in Frog's Retina," *Acta Physiol. Pharmacol. Neerl.* **13**, 231–277, 1965.
7. Gaze, R. M., and Jacobson, M. "Convexity Detectors in the Frog's Visual System," *Proc. Physiol. Soc.*, Edinburgh Meeting, July 1963.
8. Grusser-Cornehls, U., Grusser, O. J., and Bullock, T. H. "Unit Responses in the Frog's Tectum to Moving and Nonmoving Visual Stimuli," *Science*, **141**, 820–822, August 1963.
9. Grusser, O. J., Grusser-Cornehls, U., and Bullock, T. H. "Functional Organization of Receptive Fields of Movement Detecting Neurons in the Frog's Retina," *Pflügers Arch. ges. Physiol.*, **279**, 88–93, 1964.
10. Lettvin, J. Y., *et al.* "Two Remarks on the Visual System of the Frog," *Sensory Communications*, W. Rosenblith, ed., M.I.T. Press, Massachusetts Institute of Technology, 1961.
11. Moreno-Diaz, R., "An Analytical Model of the Bug Detector Ganglion Cell in the Frog's Retina," E-1858, Instrumentation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, November 1965.
12. Lettvin, J. Y. "Form-Function Relations in Neurons," *Research Laboratory of Electronics Quarterly Progress Report*, No. 66, M.I.T., Cambridge, Massachusetts, July 1962, pp. 333–335.

R. J. SWALLOW

Electrical Engineeting Department

Purdue University

Lafayette, Indiana

A Neuron-Glial Cell Model Deduced from Neurological and Psychological Considerations

ABSTRACT

Considerations of brains at both the neurological and psychological levels have suggested 1) properties for a neurglion—a neuron plus its surrounding glial cell medium, 2) ways these neurglions may be dendritically and axonally connected, and 3) sets of initial values for the dendritic and axonal connection strengths between neurglions.

This paper is divided into two parts. The first part presents briefly the suggested properties for a neurglion and its nets, outlining many of the outstanding structural and behavioral properties of the nets. The second part mathematically describes and studies the nets, illustrating or explaining some of the unique behavioral (including learning) properties which were outlined in the first part.

In particular, explanations are offered for 1) why the lower brain stem is normally inflexible, 2) why man worsens in his ability to form permanent memories as he ages, and 3) why man is curious. Strongly responsible for the above explanations and for the model's successful adherence to a diversified amount of data, is the rule by which connection strengths alter. As a consequence of this rule a neurglion will not exhibit conditioned reflex behavior unless its dendritic and axonal input pulse state is sufficiently "new".

BIOLOGICAL MODEL

Neurglions are assumed to be functionally identical except that each contains a neuron that can either inhibit or excite other neurons, but cannot do both. The neuron of each neurglion is assumed to pulse whenever its psp*, consisting of the sum of synaptically modified axonal pulses, exceeds a threshold. This threshold is assumed to be infinite for a short

* Postsynaptic potential

period just after the neuron pulses and then is assumed to decay exponentially from a finite value toward a resting value. When a neuron excites, a pulse is assumed to propagate out along its dendrites and axons. Axons and dendrites are assumed to grow and branch only if pulsing, and to grow toward pulsing neurons, if any, or else along surfaces. The glial cell medium is assumed to allow pulsing axons and dendrites to grow and branch if the average area of the excited neuronal membrane that it contains (or possibly just the non-axonal membrane) is below a critical value. A neurglion and its variables are diagrammed in Figure 1.

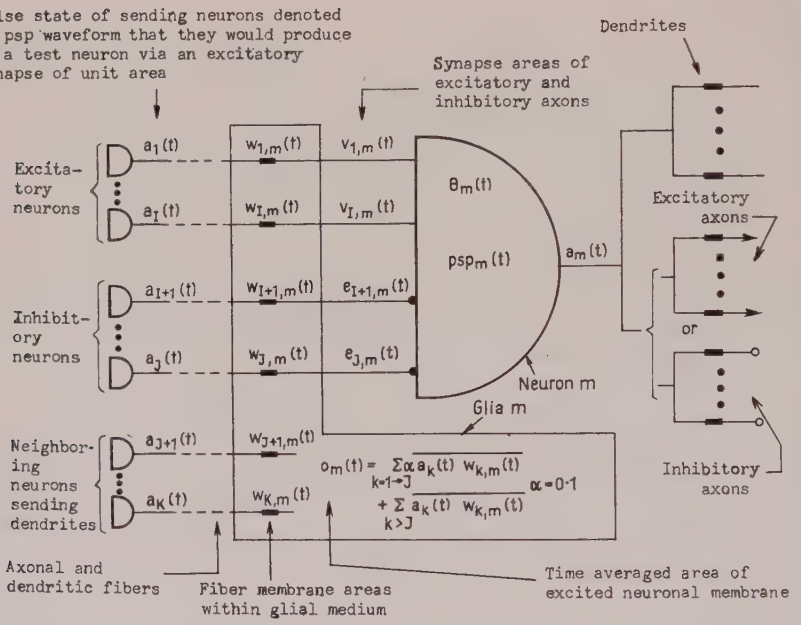


Fig. 1. Diagram of a neurglion (a neuron plus its immediately surrounding glial cell medium).

As a consequence of the above, the number of fibers in various areas of brains should increase in time. No learning should take place wherever the average area of excited membrane exceeds the critical value—thus offering a means of explaining why the lower brain stem appears inflexible. Learning should occur in the lower brain stem but only if this membrane excitation level is reduced somehow—say by sensory deprivation, nerve

severance, etc. As the area of membrane increases in regions which have exhibited learning, learning should occur less and less often—thus explaining man's apparent poorer ability to form permanent memories as he ages. The conditions for fiber growth can be shown to require that learning will occur only if the excitation pattern of the entire fiber input system to a neurglion possesses a critical newness as compared with the previous patterns which were present during fiber growth in previous learning periods. This condition on whether or not learning will occur offers a means of explaining man's curiosity drive and explains his other basic drives of hunger, sex, and pleasure as special cases. Since fiber growth is considered involved in learning, the arbitrary border between growth of a brain and learning is eliminated – learning involving the latter stages of growth under environmental influences.

Neighboring excitatory neurglions are assumed to be arranged into groups where all those of a group have strong axonal connections with a single inhibitory neurglion which in turn sends back strong axonal connections to them. There are no other axonal inputs to the inhibitory neurglion. There are both dendritic and other excitatory axonal inputs to the excitatory neurons of a group. See Figure 2.

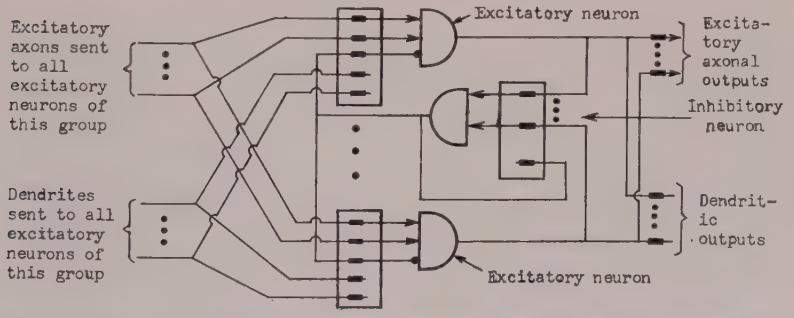


Fig. 2. Diagram of a group of excitatory neurglions made mutually inhibiting by an included inhibitory neurglion.

As a consequence of the above, the neurglion that pulses in a group is the one which is measuring the greatest similarity between its incoming excitatory axonal pulse pattern and the average of those patterns which were new when they occurred while it was pulsing previously. Thus the pulse responses from a group depend upon relative similarities rather than upon absolute similarities. The responses from a group will depend

upon the absolute similarities only if these patterns are defined over part of the axonal input system to the group, the other part of the input system biasing the response.

Each group is assumed to contain neurons of different ages. By means of an assumed genetic program, all neurons are assumed to be internally excited for a short period of time as they become of a critical age. The initial connection strengths are assumed negligibly small to begin with and to increase under the growth conditions listed above. The result is that all neurons can be excited by some sensory pattern, that certain sensory patterns can be guaranted to excite a given muscle, and that learning will occur as soon as sensory stimuli produce pulse patterns which differ from those which occurred during the setting in of the initial connection strengths.

Regions which are a few relays away from sensory neurons are found to saturate in their conditionability much earlier than those regions more relays away—thus giving an explanation why the cortex is the seat of learning in later life.

MATHEMATICAL MODEL

1. Mathematical Representation

Due to limited space we shall state the equations which relate the variables of a neurglion in time, and then without proof we shall present the mathematical results of a study of these equations. Details of the analysis can be obtained from "Toward Brain Models which Explain Human Mental Capabilities", TR-EE66-1, Report 1 of the Electrical Engineering School of Purdue University, West Lafayette, Indiana.

Referring to Figure 1, we assume

$$psp_m(t) = \sum_{k=1}^{I} a_k(t - \tau_{k,m})\, v_{k,m}(t) - \sum_{k=I+1}^{J} a_k(t - \tau_{k,m})\, e_{k,m}(t) \qquad (1)$$

where the $\tau_{k,m}$ are pulse propagation time delays.

The condition for pulsing by neuron m is taken to be

$$psp_m(t) - \theta_m(t) = 0 \quad \text{when} \quad \frac{d}{dt}(psp_m(t) - \theta_m(t)) > 0 \qquad (2)$$

defining pulsing times t_n. In terms of these times t_n the output pulse state can be written

$$a_m(t) = \sum_n p_\tau(t - (t_n + \tau_d)) \quad \tau_d \approx 1/2 \text{ msec}; \quad \tau \simeq 5 \text{ msec} \tag{3}$$

where

$$p_\tau(t - x) = U(t - x)\, e^{\frac{-t+x}{\tau}} \tag{4}$$

The threshold $\theta_m(t)$ is taken to vary as follows.

$$\theta_m(t) = \sum_n \{\infty[U(t - t_n - \delta)^* - U(t - t_n - \tau_a)] + k_\theta p_{\tau_r}(t - t_n - \tau_a)\} + \theta_0 \tag{5}$$

(δ infinitesimal; $\tau_a \approx 1/2$ msec; $\tau_r \simeq 1$ msec; $0.\infty = 0$)

The axonal connection strengths are taken to vary as follows.

$$\frac{dv_{k,m}(t)}{dt} = a_k(t - \tau_{k,m})\, a_m(t)\, U(\theta_L - o_m(t)) \quad k = 1 \to I \tag{6}$$

$$\frac{de_{k,m}(t)}{dt} = \beta \cdot \text{same as above} \quad k = I + 1 \to J \tag{7}$$

where

$$o_m(t) = \sum_{k=1}^{J} \alpha \overline{a_k(t - \tau_{k,m})\, w_{k,m}(t)} + \sum_{k=J+1}^{K} \overline{a_k(t - \tau_{k,m})\, w_{k,m}(t)} \quad \alpha = 0{,}1 \tag{8}$$

The rates of change of the membrane areas are taken to obey the following expressions.

$$\frac{dw_{k,m}(t)}{dt} = a_k(t - \tau_{k,m})\, U(\theta_L - o_m(t)) \quad k = 1 \to K \tag{9}$$

Analytic Results

Assume a connection scheme as in Figure 2, where the excitatory connections from the N excitatory to the inhibitory neuron is large enough to guarantee that the inhibitory neuron fires whenever any one of the N neurons fires and where the inhibitory neuron is not pulsing at a rate above 100 cps (2 pulse decay time constants).

* $U(x) = 1$ if $x \gg 0$; $= 0$ otherwise.

By means of manipulating the above equations, we arrive upon the following expressions where the ith possible excitatory axonal pulse state is denoted by $A_i = \{{}^{i}a_1\, {}^{i}a_2 \ldots {}^{i}a_I\}$.

$$P_m(t) \simeq \frac{\sum_i q_i(t)\, {}^{i}\varepsilon_m(t) - \theta_m(t)}{\sum_i {}^{i}\varepsilon_m(t)} \tag{10}$$

N equations, one for each neuron in the group
where

$${}^{i}\varepsilon_m(t) = \int_{\substack{\Delta t \\ i}} a_m(t)\, U(\theta_L - o_m(t))\, dt \tag{11}$$

$\underset{i}{\Delta t}$ = time that A_i was input pulse state

which crudely is proportional to the time A_i was the input pulse state when neuron m pulsed while learning was allowed ($o_m(t) < \theta_L$), and where

$$q_i(t) = \sum_{k=1}^{I} a_k(t)\, {}^{i}a_k \tag{12}$$

which crudely represents the similarity of the pulse state at time t with pulse state A_i.

The neuron that will pulse at time t will be the one with the maximum $P_m(t)$ within the group of N neurons. This expression for $P_m(t)$ can be reduced further to the following expression.

$$P_m(t) \simeq \sum_{k=1}^{I} a_k(t)\, \bar{a}_k = \bar{q}(t) \tag{13}$$

where the $\bar{a}_k$ is the average value of $a_k(t)$ during times that neuron m pulsed when $o_m(t) < \theta_L$. Thus the excitatory neuron that pulses in the group is the one "seeing" the incoming pulse state as most "similar" to the average it "saw" on occassions when learning "onto" neuron m occurred.

Further analysis yields the following expression for $o_m(t)$.

$$o_m(t) = \alpha \sum_i \overline{q_i(t)\, {}^{i}\nu_m(t)} + \alpha \sum_i \overline{q_i'(t)\, {}^{i}\eta_m(t)} + \sum_i \overline{q_i''(t)\, {}^{i}\mu_m(t)} \tag{14}$$

where we have also labeled the input pulse states of the inhibitory input system to each neurglion by B_i's ($B_i = \{{}^{i}a_J\}$) and the input pulse states

of the dendritic input system by C_i's ($C_i = \{{}^i a_{J+1}\, {}^i a_{J+2} \ldots {}^i a_k\}$, where

$$q_i'(t) = a_J(t)\, {}^i a_J \tag{15}$$

which can be interpreted as the similarity of incoming inhibitory pulse state with one of its possible states B_i, and where

$$q_i''(t) = \sum_{k=J+1}^{K} a_k(t)\, {}^i a_k \tag{16}$$

which can be interpreted as the similarity of the incoming dendritic pulse state with state C_i, and finally where

$${}^i\nu_m(t) = \int_{\Delta t_i} U(\theta_L - o_m(t))\, dt \tag{17}$$

$${}^i\eta_m(t) = \int_{\Delta' t_i} U(\theta_L - o_m(t))\, dt \tag{18}$$

$\Delta' t_i$ those times the inhibitory pulse state was B_i

and

$${}^i\mu_m(t) = \int_{\Delta'' t_i} U(\theta_L - o_m(t))\, dt \tag{19}$$

$\Delta'' t_i$ those times the dentritic pulse state was C_i

The ${}^i\nu_m(t)$, ${}^i\eta_m(t)$, and ${}^i\mu_m(t)$ can be interpreted respectively as the total time A_i, B_i, and C_i were present when learning was allowed ($o_m(t) < \theta_L$).

If the incoming pulse patterns (A, B, or C) are similar to others that were present often in the past during learning periods, there will be large terms in the expression for $o_m(t)$ making $o_m(t)$ large. Otherwise, $o_m(t)$ will be small. Thus $o_m(t)$ is a measure of the oldness of the incoming pulse state as compared to others that have come before during learning periods. A period of time will be a learning period if the incoming pattern is not too old, i.e. is sufficiently new.

3. Discussion

If we imagine that the lower brain stem has, so to speak, received a diverse sample of input pulse states such that there are no more "new" ones (possibly different ones, but not new as we have defined new) which can occur under normal conditions, learning will be impossible. If, how-

ever, an abnormal situation such as severing a nerve (or temporarily inhibiting a nerve's conduction of a usually large number of pulses per unit time) occurs, the q's in $o_m(t)$ will become smaller, possibly reducing $o_m(t)$ below θ_L.

Since it is a rarer and rarer occasion that new patterns can be found (each incoming pattern being forced to have some similarity to a previous pattern due to the ever increasing sample of previous patterns), there will be fewer and fewer occasions when learning will occur. This offers an explanation why man's ability to form permanent memories decreases as he ages.

Because newness is rewarding in this model where the rule for changing connection strengths between neurons is basically a conditioned reflex rule, we find ourselves in a unique position of having a neuron design which may make for nets which have the goal for newness, i.e. a curiosity drive. Why a child learns to crawl, then walk may be no more than the result of the fact that such behavior yields new input pulse states which are rewarding.

Is this the universal goal we are looking for in self-organizing systems?

W. H. BROCKMAN†
Department of Electrical Engineering, Iowa State University
Ames, Iowa.
K. S. FU
School of Electrical Engineering, Purdue University,
Lafayette, Indiana.

A Stimulus Conditioning Learning Model and its Application to Pattern Recognition

INTRODUCTION

This paper describes the development, analysis and simulation of a learning model based on a psychological theory of learning. The model takes its inspiration from the "two-factor" learning theory of O. H. Mowrer.[1, 2] The basic tenets of this theory are: (1) All learning is basically stimulus conditioning. (2) Organisms increase the stimulation which is conditioned to an internal emotional response denoted as "hope" and decrease the stimulation conditioned to "fear". Hope, for example, arises in the following fashion. An organism is deprived of a basic necessity of life, such as food, thus creating a hunger "drive." If a stimulus, such as a bell ringing, is applied concurrently with food, the drive is reduced. Hope of drive reduction is conditioned to the sound of the bell. In like fashion, fear is generated by an increase in drive.

Habit formation is assumed to occur when proprioceptive stimuli associated with some response are conditioned to hope through a drive reduction. Inhibition of a response occurs when the proprioceptive stimuli are conditioned to fear. In a discrimination problem, organisms use a vicarious trial and error method to tentatively sample the expected outcome of each of the possible responses. This provides an immediate comparison of the favorability of each response.

DEVELOPMENT OF THE LEARNING MODEL

The mathematical model of the process outlined above assumes the external stimuli applied to the system to be n in number and to have

† Formerly of School of Electrical Engineering, Purdue University.

binary properties. The whole stimulus pattern can be represented as a stimulus vector S.

$$S = [s_1, s_2, s_3, \ldots s_n] \tag{1}$$

whose elements are defined as in (2)

$$s_i = \begin{cases} 1 \text{ if stimulus } i \text{ is present in the pattern} \\ 0 \text{ if stimulus } i \text{ is not present in the pattern} \end{cases} \quad i = 1, \ldots, n \tag{2}$$

The responses are assumed to be binary, mutually exclusive, and exhaustive. Some m responses are assumed. A response vector is defined as

$$R = [r_1, r_2, \ldots, r_m] \tag{3}$$

where the elements r_j are binary

$$r_j = \begin{cases} 1 \text{ if the response } j \text{ is chosen} \\ 0 \text{ if the response } j \text{ is not chosen} \end{cases} \quad j = 1, \ldots, m \tag{4}$$

A measure of the hope and/or fear conditioned to the conjunctive occurrence of external stimuli and internal proprioceptive stimuli associated with a given response is defined in (5) and (6).

H_{ij} = Hope conditioned to the simultaneous presence of s_i and the proprioceptive stimuli correlated to r_j. (5)

F_{ij} = Fear conditioned to the simultaneous presence of s_i and the proprioceptive stimuli correlated to r_j. (6)

The favorability of making a given response r_j when considering only one stimulus element s_i is

$$f_{ij} = H_{ij} - F_{ij} \tag{7}$$

and the total favorability of response r_j is

$$f_j = \sum_{i=1}^{n} s_i [H_{ij} - F_{ij}] \tag{8}$$

Decisions as to the appropriate response to a given stimulus pattern are made by choosing the response with the greatest favorability as defined in (8).

The training rules making use of reward (drive reduction) or punish-

ment (drive induction) applied by an external trainer are as follows. Let H_{ij_N} be the value of hope on trial N.

$$H_{ij_{N+1}} = \alpha H_{ij_N} + (1 - \alpha)\,\Phi_k \quad 0 < \alpha < 1$$
$$F_{ij_{N+1}} = \alpha F_{ij_N} + (1 - \alpha)\,\Psi_k \tag{9}$$

When reward is applied Φ_k = maximum possible value of hope and $\Psi_k = 0$. When punishment is applied $\Phi_k = 0$ and Ψ_k = maximum possible value of fear. The learning parameter is α.

When training is terminated, all responses made are assumed to be correct. The index for termination of training is specified as the point at which the total number of correct trials is a given fraction of the total number of trials.

ANALYSIS OF THE MODEL

It must first be established that the system learns. That is, the system parameters must be nonstationary as a function of the history of the training sequence. Consider a general hope value H_N. The changes in H_N are defined as:

$$H_{N+1} = \alpha H_N + (1 - \alpha)\,\Phi_k$$
$$\Phi_1 = \Phi \text{ for reward} \quad \Phi_2 = 0 \text{ for punishment.} \tag{10}$$

Let the probability that Φ_k be used be π_k.

$$P(\Phi_k) = \pi_k \sum_{k=1}^{2} \pi_k = 1 \tag{11}$$

On the Nth trial in the sequence there exist 2^N possible sequences of application of Φ_k leading to that trial. Let $P_j(N)$ denote the probability that the jth sequence will take place. Let $H_j(N)$ be the value of hope on the Nth trial given that sequence j has taken place. The expected value of hope on trial N is denoted as $\bar{H}_N$. It is assumed that either Φ_1 or Φ_2 was applied on each trial.

$$\bar{H}_N = \sum_{j=1}^{2^N} P_j(N)\, H_j(N) \sum_{j=1}^{2^N} P_j(N) = 1 \tag{12}$$

On the Nth trial either Φ_1 or Φ_2 will be applied

$$\bar{H}_{N+1} = \sum_{j=1}^{2^N} P_j(N) \sum_{k=1}^{2} [\pi_k \alpha H_j(N) + \pi_k(1 - \alpha)\Phi_k] \tag{13}$$

$$\bar{H}_{N+1} = \alpha \bar{H}_N + (1 - \alpha)\bar{\Phi} \quad \text{where} \quad \bar{\Phi} = \sum_{k=1}^{2} \pi_k \Phi_k \tag{14}$$

The above expression for $\bar{H}_{N+1}$ defines a geometric progression.[3] If $H_0 = 0$,

$$\bar{H}_N = \sum_{i=0}^{N-1} \alpha^i (1 - \alpha)\bar{\Phi} = (1 - \alpha^N)\bar{\Phi} \quad \text{for} \quad |\alpha| < 1 \tag{15}$$

The variance of the hope level, $\sigma^2_{H_N}$, can be developed in a similar fashion

$$\sigma^2_{H_N} = \overline{H_N^2} - \bar{H}_N^2 \tag{16}$$

where $\overline{H_N^2}$ is the second order moment of H_N and $\bar{H}_N^2$ is the square of the mean.

$$\overline{H_{N+1}^2} = \alpha^2 \overline{H_N^2} + 2\alpha(1 - \alpha)\bar{H}_N\bar{\Phi} + (1 - \alpha)^2 \overline{\Phi^2} \quad \text{where}$$

$$\overline{\Phi^2} = \sum_{k=1}^{2} \pi_k \Phi_k^2 \tag{17}$$

If the variance of the asymptotic values Φ_k is defined as

$$\sigma^2_{\Phi} = \overline{\Phi^2} - \bar{\Phi}^2 \tag{18}$$

Then it is found that

$$\sigma^2_{H_{N+n}} = \alpha^2 \sigma^2_{H_n} + (1 - \alpha)^2 \sigma^2_{\Phi} \tag{19}$$

A closed form expression can be derived from the geometric progression given in (19) as

$$\sigma^2_{H_N} = (1 - \alpha^{2N}) \frac{(1 - \alpha)}{1 + \alpha} \sigma^2_{\Phi} \tag{20}$$

The mathematical expressions above may be interpreted as learning phenomena if it is recalled that $H_0 = \bar{H}_0 = 0$. As the training sequence progresses, the expected values of hope and fear, $\bar{H}_N$ and $\bar{F}_N$, increase to non-zero asymptotic values. The variances likewise approach limit conditions. The system thus starts in a naive state and progresses to the condition where the values of hope and fear reflect the probabilistic history

of the rewards and punishments. Note that as α approaches unity, the variance approaches zero, thus the actual values of H_N and F_N do not vary appreciably from the means.

APPLICATION TO PATTERN RECOGNITION

Although there exist many pattern recognition problems to which the model might be applied, only one class will be analyzed here. Consider the situations where two patterns are to be applied. When S_1 is applied r_1 is the appropriate response. Similarly $S_2 - r_2$ constitute a favorable pair. The stimulus vectors are assumed to have x elements in each which do not occur in the other and some y elements which are common. For example if $x = 2$ and $y = 4$, the stimuli might be

$$S_1 = [1, 1, 0, 0, 1, 1, 1, 1]$$
$$S_2 = [0, 0, 1, 1, 1, 1, 1, 1] \tag{21}$$

If on each trial either reward or punishment is applied then

$$\pi_{i1} = 1\,\pi_{i2} = 0$$

For $i = 1, \ldots, x$

$$\varrho_{i1} = 0\,\varrho_{i2} = 1 \tag{22}$$
$$\pi_{i1} = 0\,\pi_{i2} = 1$$

For $i = x + 1, \ldots, 2x$

$$\varrho_{i1} = 1\,\varrho_{i2} = 0 \tag{23}$$

$$\pi_{ij} = \pi_j$$

For $i = 2x + 1, \ldots, 2x + y$

$$\varrho_{ij} = 1 - \pi_{ij} \tag{24}$$

where π_j is the probability that r_j will be rewarded and π_{ij} is the probability that r_j will be rewarded when s_i is present. ϱ_{ij} is the probability that r_j will be punished when s_i is present. The asymptotic values of hope and fear are made equal to maintain symmetry.

$$\Phi_{ij} = \Psi_{ij} = \Phi > 0 \tag{25}$$

For all $i = 1, \ldots n$ $j = 1, \ldots, 2$

Before the model is able to solve the discrimination problem, the correct response corresponding to the most frequent stimulus will predominate. The dominant stimulus occurs with probability π_k where

$$\pi_k = 0.5 + \varepsilon \quad \varepsilon > 0 \tag{26}$$

The stimulus which is dominated will virtually always be incorrectly discriminated. The expected values of favorability index for the two stimuli in the asymptotic case are given below for $\pi_1 = 0.5 + \varepsilon$.

where $f = f_{r_1} - f_{r_2}$

$$\bar{f}_1 = \Phi x + 2\varepsilon\Phi y$$
$$\bar{f}_2 = -\Phi x + 2\varepsilon\Phi y \tag{27}$$

Since Φ and ε are positive, $\bar{f}_1$ is always positive and S_1 will always be correctly discriminated. $\bar{f}_2$ is negative only when

$$\frac{x}{y} > 2\varepsilon \tag{28}$$

Equation (28) imposes a sufficient limitation on x and y for solution of the discrimination problem with asymptotic zero error rate.

The actual values of f_1 and f_2 depart from the mean when $\alpha < 1$, thus the effects of such changes must be taken into account. A recursive formula for the change in hope and fear values can be developed for $m = 2$. Let $\bar{\Delta}H$ be the expected asymptotic change in H.

$$\bar{\Delta}H = 2(1 - \alpha)(1/4 - \varepsilon^2)\Phi$$
$$\bar{\Delta}F = 2(1 - \alpha)(1/4 - \varepsilon^2)\Phi \tag{29}$$

The deviation from the mean of the favorability index for the stimuli s_i ($i = 2x + 1, \ldots, 2x + y$) is thus

$$y\left[\overline{\Delta H} + \overline{\Delta F}\right] = (1 - \alpha)(1 - 4\varepsilon^2)\Phi y \tag{30}$$

Since the deviation is given in (30), the discrimination problem approaches solution when

$$\Phi x \geqq (1 - \alpha)(1 - 4\varepsilon^2)\Phi y \tag{31}$$

The constraint on the learning rate parameter is

$$\alpha \geqq 1 - \frac{x}{(1 - 4\varepsilon^2)\, y} \tag{32}$$

The larger the value of α, the slower the learning proceeds. Thus the optimal value of α is that given by equality in (32).

The conditions necessary for the dominated stimulus to be correctly discriminated half of the time, leading to an error rate of

$$E = \left(\frac{1}{4} - \frac{\varepsilon}{2}\right) 100\,\%$$

can be evaluated by noting that the use of the optimal α yields

$$y[\overline{\Delta H} + \overline{\Delta F}] = x\Phi \tag{33}$$

In order to maintain a fixed mean value of favorability, the index must fall above or below the mean by a distance of Φx half of the time. This gives rise to the situation where

$$\bar{f}_2 = -\Phi x + 2\varepsilon\Phi y - \Phi x \quad \text{or} \quad \frac{x}{y} \geqq \varepsilon \tag{34}$$

This is the limiting case for the specified error rate.

COMPUTER SIMULATION AND RESULTS

Experimental investigation of the properties of the learning model was done using an IBM 7094 computer. Both the stimuli and the model were

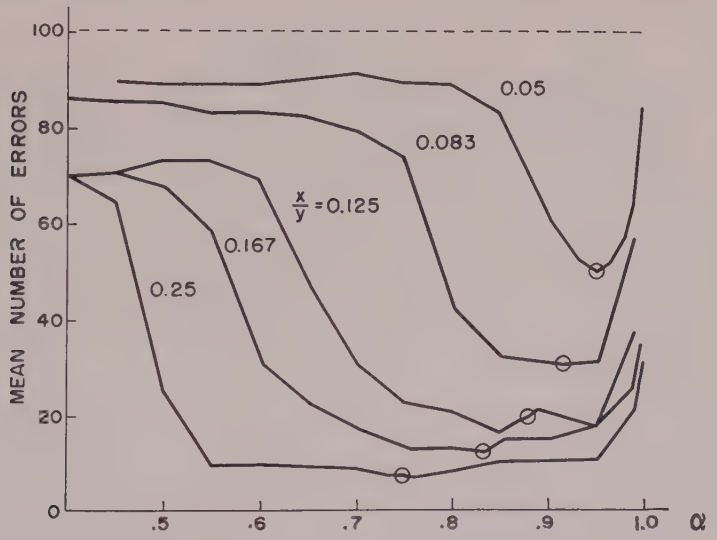


Fig. 1. Mean Error as a Function of α Where $\varepsilon = 0.0$.

simulated. Simple patterns of the type discussed in part IV were applied to the model and the performance for several values of α is given in Figures 1 and 2. The circles indicate the optimal values for α computed for each ratio x/y. The dotted lines indicate the probability matching mode where S_2 is not ever correctly discriminated. Note that α becomes critical for difficult problems. The conditions noted in (34) were found to be conservative.

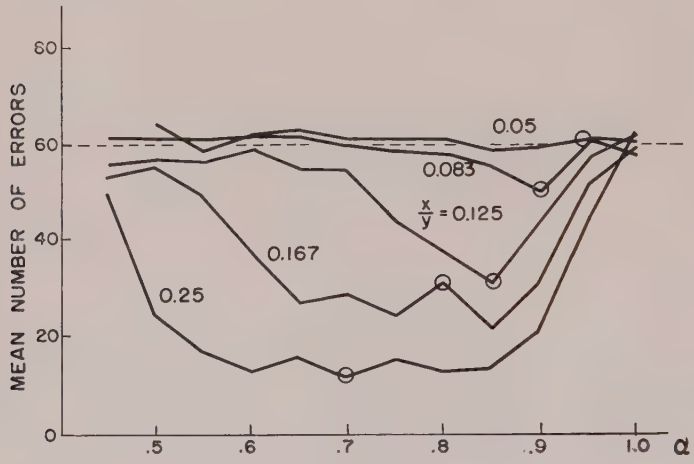


Fig. 2. Mean Error as a Function of α Where $\varepsilon = 0.20$.

The learning model was also applied to the recognition of handwritten block letters represented in quantized form by a 12 × 12 matrix. The 144 cells in the matrix were logically reduced to form 48 stimulus elements for the system. Fifty each of A's and B's were applied. In each case the training sequence consisted of 10 each A's and B's chosen arbitrarily as prototypes. Training data were provided on these 20 characters until a specified performance criterion was reached. Training was then terminated. Figure 3 illustrates the performance for various values of the minimum performance criterion. The problem was extended to three classes (A, B, C) with results of 97% accuracy at the termination of training.

The meteorological prediction of precipitation was simulated by using as stimuli quantized representations of the 500 millibar altitude wind direction and geographical areas of precipitation. The forecasts were made for a 24-hour period at Indianapolis, Indiana. A polar coordinate representation of the United States as shown in Figure 4 was used to

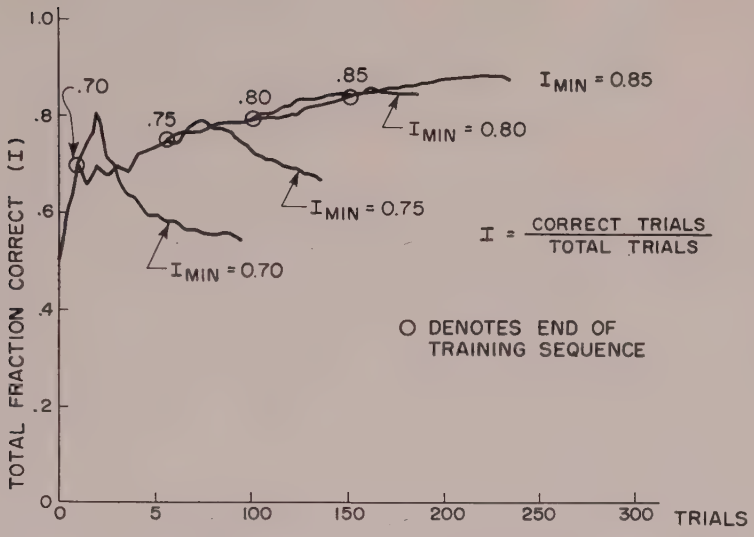


Fig. 3. Performance for Two Classes of Block Letters.

Fig. 4. Precipitation Occurrence Grid.

generate a binary vector indicating the location of precipitation. Training data covered 30 days in October, 1962 and recognition without training was tested on the subsequent 30 days in November, 1962. The accuracy during recognition was found to be 73.5 per cent. This represents a "skill score" of 38.4 per cent.[4] This is interpreted as recognition of precipitation in 38.4 per cent of the cases where climatological averages would be incorrect.

The diagnosis of bundle branch block cardiac abnormalities from characteristics of the electrocardiogram was simulated on 72 EKG records, 36 of which were normal. Quantized representations of the following data were used: (1) Auricular rate; (2) Ventricular rate; (3) $P - R$ interval; (4) QRS interval, and (5) $Q - T$ interval. A typical electrocardiogram and the significant data are shown in Figure 5. A total of 31 stimulus elements were generated. Recognition, following training, was found to be 89.6%.

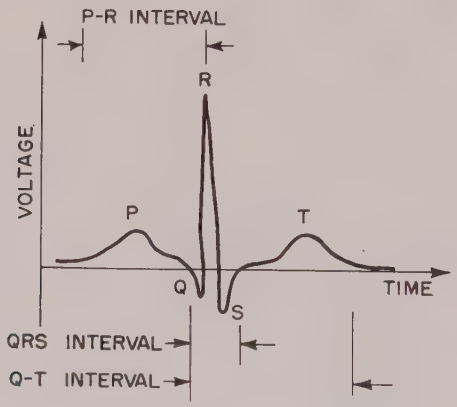


Fig. 5. The Normal Electrocardiogram

A comparison between human performance and the model was carried out. Two classes of patterns to be discriminated were applied to humans and simulated on the computer. The classes were of the form shown in Figure 6. The performance of the human subjects is seen to fall far behind that of the computer simulated mode. The lack of contextual orientation of the patterns causes performance here which is quite different from that expected for human character recognition. A modification of the model, wherein the computer uses each stimulus element only 60 per cent of the

time, gave the modified model curve which much more accurately approximates human performance. This leads to the possibility that perception or attention to the pattern elements is critical.

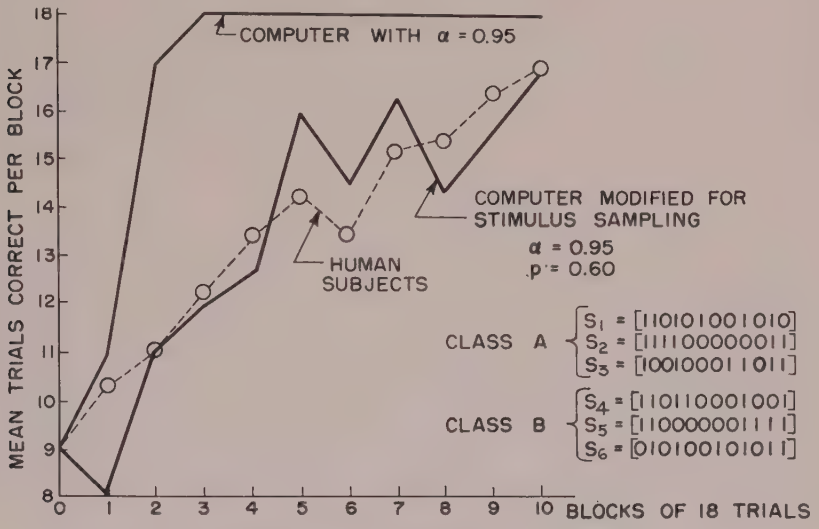


Fig. 6. Performance for Psychological Acquisition

CONCLUSION

An engineering model of learning based on psychological concepts has been shown to be useful for application to a variety of pattern recognition problems. The disparate results between humans and the model indicate modifications are necessary to describe human performance.

REFERENCES

1. Mowrer, O. H., *Learning Theory and Behavior*, John Wiley and Sons, Inc., New York, 1960.
2. Mowrer, O. H., *Learning Theory and the Symbolic Processes*, John Wiley and Sons, Inc., New York, 1960.
3. Bush, R. R., Mosteller, F., and Thompson, G. L., "A Formal Structure for Multiple Choice Situations," *Decision Processes*, Eds. R. M. Thrall, C. H. Coombs, and R. L. Davis, John Wiley and Sons, Inc., New York, 1954.
4. Sutton, O. G., *The Challenge of the Atmosphere*, Harper and Brothers, New York, 1961.

MARK W. CANNON

Aerospace Medical Research Laboratories

Wright-Patterson Air Force Base, Ohio

Mathematical Model of the Encoding Function of the Eighth Nerve Neuron

INTRODUCTION

A number of mathematical models of the spike activity of auditory neurons have been developed. Miranker[3] has used a simplified model of the ion transportation mechanism. His model shows many features observed in auditory neurons except for the randomicity of the interspike intervals. Weiss[5] started with the assumption of an exponential recovery of the neural threshold after a firing and used a noise signal superimposed on the input to give the random interspike intervals. However, this model does not show rate adaptation. A more sophisticated model is proposed by Gerstein and Mandelbrot.[1] They propose the random walk of a "state point" between threshold and some lower reflecting boundary. By changing two parameters they can model the interspike interval histograms of a large number of spontaneously firing neurons.

The model proposed in this paper is an approach to the problem from a functional viewpoint. An interspike interval density is postulated which is a function of the magnitude and rate of change of a generating potential. This generating potential is a function of both basilar membrane displacement and average firing rate of the model neuron. The model produces pulse occurrence histograms which compare very well with recorded data and also matches curves of average firing rate vs. input intensity from a number of units recorded in both the guinea pig and monkey.

BASIC ASSUMPTIONS

Many spontaneously firing eighth nerve neurons exhibit interspike interval histograms of the kind shown in Figure 1(a) which are skewed to the left and have zero amplitude over some absolute refractory period r.

When interspike interval histograms are computed from stimulated neurons, the entire curve becomes more skewed to the left, as shown in Figure 1(b).

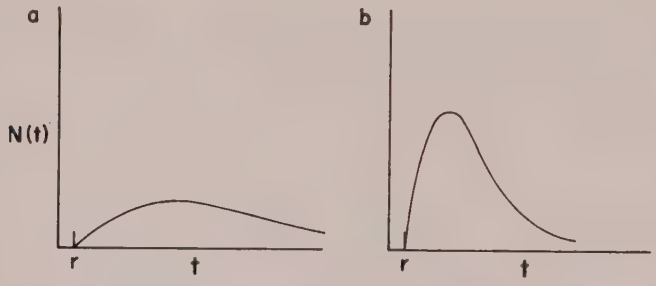


Fig. 1. Typical Interspike Interval Histograms.

A similar behavior can be produced if the interspike intervals are chosen from the population

$$p(t, \alpha) = \begin{cases} \alpha^2 (t - r) e^{-\alpha(t-r)} & t \geqq r \\ 0 & 0 \leqq t < r \end{cases}$$

where the average interval is $r + 2/\alpha$.

If α is assumed to be a generating potential and is made some function of a stimulus intensity, the basis for a simple model of neural firing has been achieved.

The average interval for spontaneous firing is set by choosing α equal to some α_0 and a threshold is introduced by not permitting any α's less than α_0 to have an effect on the model.

With these ideas in mind we will proceed to a dynamic model by making some simple assumptions about the behavior of the interspike interval density function when the stimulus changes with time.

TRANSIENT INPUT EFFECTS

Consider the integral of $p(t, \alpha)$ from 0 to t

$$\int_0^t p(t, \alpha)\, dt = A(t, \alpha) = 1 - e^{-\alpha(t-r)} [1 + \alpha(t - r)]$$

This function gives the probability that a firing will have occurred by time t after the previous firing.

Two such curves are shown below in Figure 2 for $\alpha = \alpha_0$ and $\alpha = \alpha_1$ Suppose that at some time τ, α jumps from α_0 to α_1, producing a new function, $A(t, \alpha(t))$, which is shown in Figure 3.

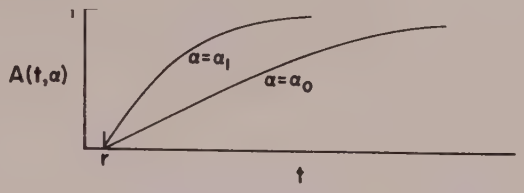


Fig. 2. Interspike Interval Distribution Functions.

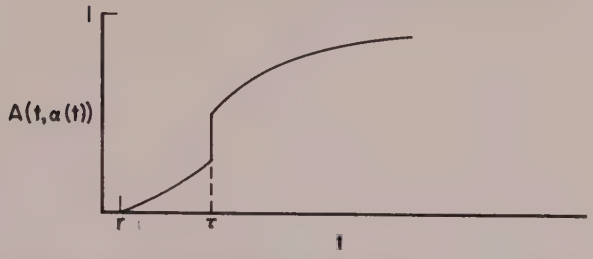


Fig. 3. Interspike Interval Distribution Function for a Step Increase in α

The interspike interval density function, $p(t, \alpha(t))$, for this case is shown in Figure 4.

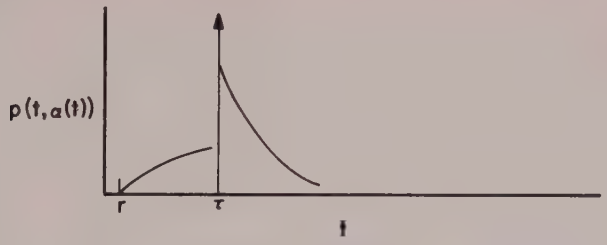


Fig. 4. Interspike Interval Density Function for a Step Increase in α

$$p(t, \alpha(t)) = \begin{cases} p(t, \alpha_0) & 0 \leqq t < \tau \\ p(t, \alpha_1) & \tau < t \\ c\delta(\tau) & \tau = t \end{cases}$$

$$c = 1 - \int_0^{\tau} p(t, \alpha_0)\, dt - \int_{\tau}^{\infty} p(t, \alpha_1)\, dt$$

Physically, of course, such a case is impossible since $\alpha(t)$ can never be a true step function. $A(t, \alpha(t))$ will, therefore, never have an infinite slope. However, the simple result in Figure 4 leads to a method of computing the interspike interval density function for αs which are continuous. This method is given in detail in the Appendix and yields a function, $p(t, \alpha(t))$, which is good for any $\alpha(t)$.

$$p(t, \alpha(t)) = [t - r]\,\alpha^2(t)\,e^{-[t-r]\alpha(t)}\left([t - r]\frac{\alpha'(t)}{\alpha(t)} + 1\right)$$

$$\alpha'(t) = \frac{d\alpha(t)}{dt}$$

Figure 5 shows a typical $\alpha(t)$ and corresponding interval density, $p(t, \alpha(t))$.

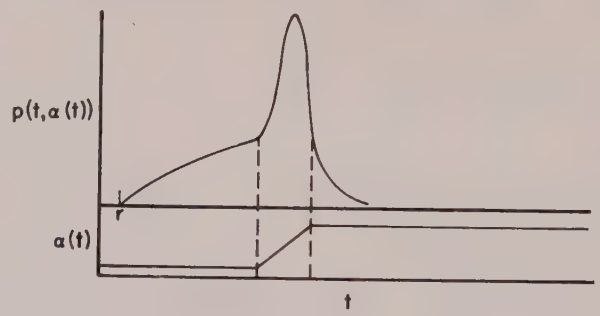


Fig. 5. Interspike Interval Density Function for Linearly Increasing α

The result predicts quite reasonably that the model is more likely to fire in a region of increasing stimulus than anywhere else. This is in agreement with real behavior where firing synchronization of a population of neurons is observed at the onset of a stimulus.

The case of decreasing α presents a slight problem, however, which can be overcome by introducing a simple rule. First, we will demonstrate

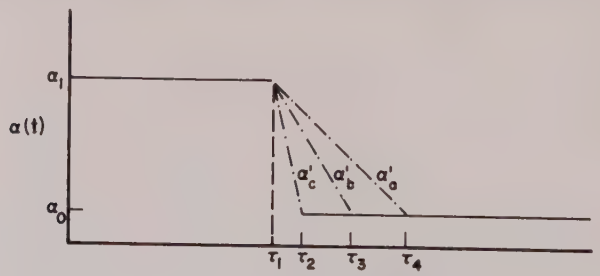


Fig. 6. Three Typical α Transitions

the problem. Let $\alpha(t)$ undergo a transition from α_1 to some smaller value α_0 and consider the three paths for this transition which are shown in Figure 6.

In this example $\alpha_a' > \alpha_b' > \alpha_c'$ since all are negative. Now we look at the curves, $A(t, \alpha)$, produced by these three transitions (Fig. 7).

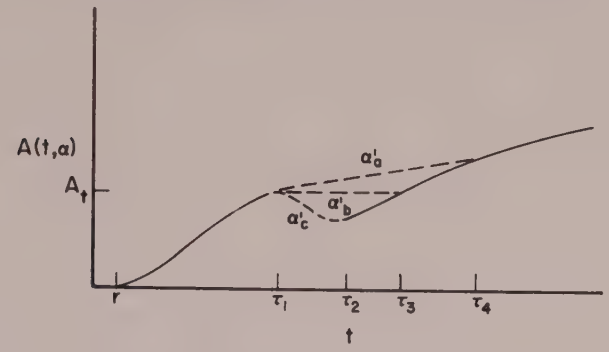


Fig. 7. $A(t, \alpha)$ curves for the Three α Transitions of Fig. 6.

We see that α_b' is such as to produce a curve, $A(t, \alpha)$ with zero slope in the region $\tau_1 < t < \tau_3$. Any α' greater than α_b', therefore, will give a $p(t, \alpha(t))$ which is positive everywhere and is a true interspike interval density function. However, when $\alpha'(t) = \alpha_c'$ during transition, we have the absurd prediction that the probability that the model will have fired by some time $\tau_1 < t < \tau_3$ is less than the probability that it will have fired by time τ_1. The correction that must be made then, is simply to invoke a rule which states that once the amplitude of $A(t, \alpha)$ has reached some arbitrary value (in this case A) it can never drop below that value, regardless of the value of $\alpha'(t)$. This means that all transitions with α' less than α_b' follow the curve labeled by α_b' in Figure 7. The interspike

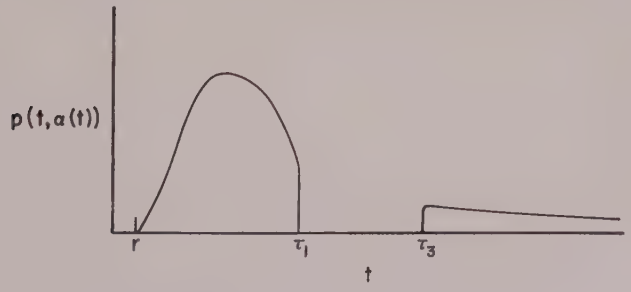


Fig. 8. Interspike Interval Density for Decreasing α

interval density corresponding to this $A(t, \alpha(t))$ curve is shown in Figure 8. This result (quite necessary if we wish to keep the interval density curve from producing absurd negative values) produces the depression in firing, after removal of the stimulus, which is so often observed in eighth nerve neurons.

We now have a model which, even before the refinements to come in the next section, predicts the synchronization of firing at the onset of a stimulus and the depression of spontaneous activity after the stimulus is removed.

GENERATING POTENTIAL AND ADAPTATION

The input to the generating potential transfer function of the model is given by the following diagram

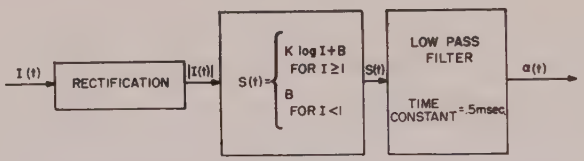


Fig. 9. Block Diagram of Generating Potential Transfer Function

$I(t)$ is the input amplitude representing the displacement of the basilar membrane (assumed to be linear). The next three blocks represent the combined effect of hair cell, dendrites and nerve cell body in producing a function, $\alpha(t)$, which determines the firing rate under the statistical rules set forth in the previous sections. B is set equal to some α_0 to determine the spontaneous firing rate and k controls the dynamic range, with small ks giving large dynamic ranges.

Since much of the experimental work on guinea pigs at the Aerospace Medical Research Laboratories was done with tone bursts above 1 kHz, it seemed adequate to simulate tone bursts of frequencies in this region with the envelope of the rectified signal. This approximation proved to be justified by the results.

After considerable experimentation, a representation of the so-called adaptation effect (a reduction in average firing rate after the onset of a sustained stimulus) was found. The effect was modelled by subtracting a function, $he^{-t/\tau}$ from the generating potential α after each firing when α was greater than α_0. This is a sort of negative feedback which causes

the average interval between pulses to lengthen beyond the initial average value attained during stimulus onset. It was soon found, after more experimentation, that the coefficient h must depend in some way on the input amplitude.

Values of h large enough to give proper adaptation at high input amplitudes would wipe out responses for low inputs immediately after the onset. Small h values produced almost no effect at high input levels. A function, $Gh(\alpha)$, was derived impirically which was found to be linear at low amplitudes and logarithmic for α's greater than 800. This function is shown in Figure 10 for $G = 1$ and $G = 2$. This dependence of h on α

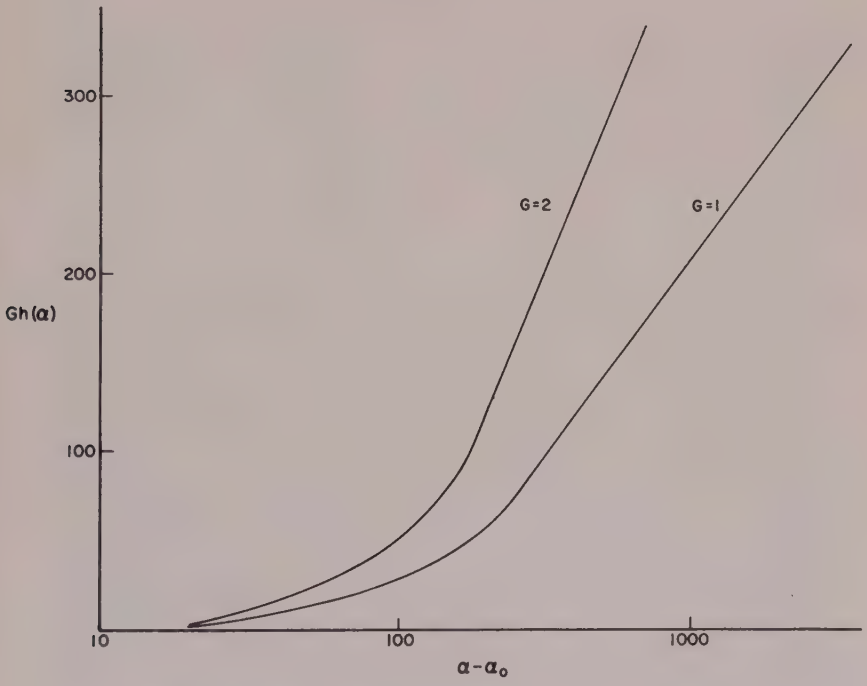


Fig. 10. Feedback Gain as a Function of α for Two Values of R

is a rather interesting feature. The fact that the amplitude of these exponentials is dependent on the amplitude of the input makes this system quite different from standard feedback systems. This can be made clearer by considering what goes on in the model. If the firing rate goes up, the number of exponentials per unit time subtracted from $\alpha(t)$ goes up.

If the average rate is reasonably high, this has the effect of reducing α by an amount proportional to $h\tau$ times the average firing rate R. If h and time constant τ are fixed, the input level α is reduced an amount proportional to the output firing rate. However, if h is also a function of α we have a variable gain feedback which is proportional to α. Consequently, the function $\alpha(t)$ is now reduced by $Gh(\alpha)\,\tau R$.

METHOD OF SIMULATION

The input function, $I(t)$, is generated and stored in the computer. $\alpha(t)$ is then computed through the transfer function indicated in Figure 9. This function is specified at 500 points on the time axis, each point representing a $\frac{1}{2}$ millisecond interval. A first firing density function is then computed which gives the density of intervals from the origin to the first pulse. A subprogram picks one interval at random from this population, determining the position of the first pulse on the time axis. After this an interval density function, $p(t, \alpha(t))$, is computed and a pulse interval

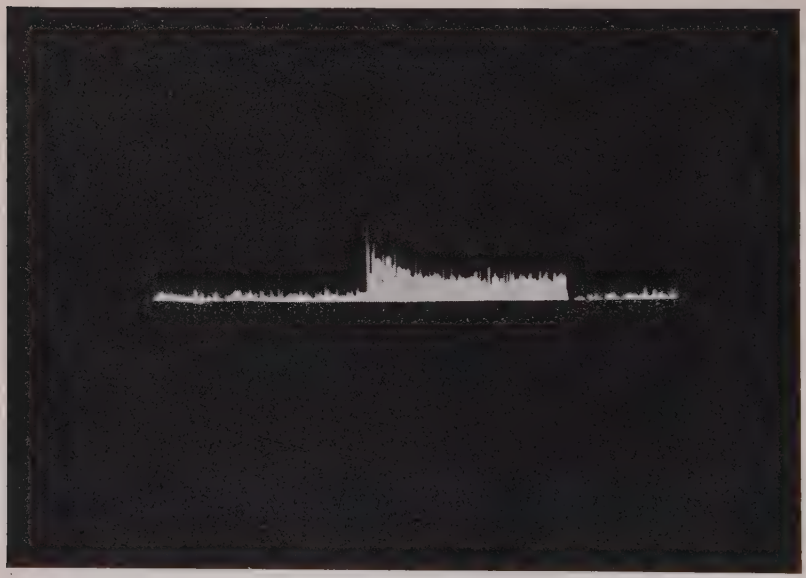
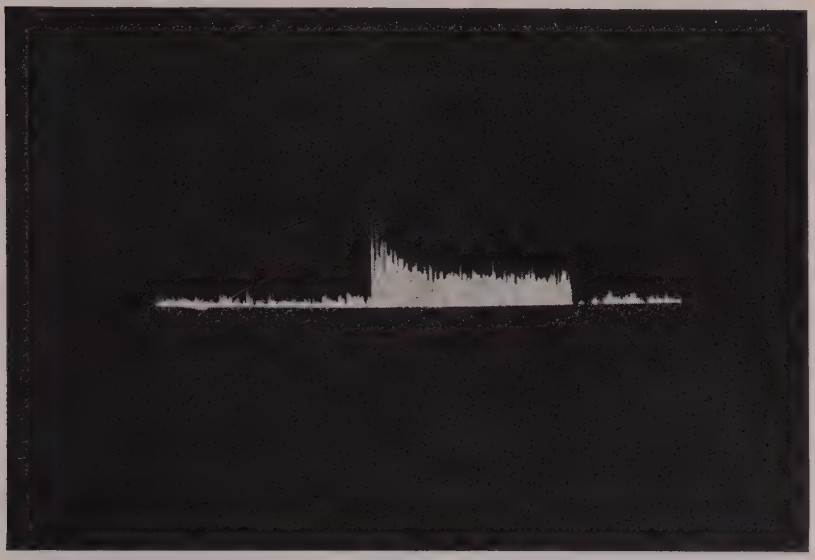


Fig. 11. Pulse Histograms Produced by the Model
a) 20 msec Rise Time $I_{max} = 1{,}000$

is picked from this population. The program continues in this way to compute the occurrence of pulses on the time axis until one pulse falls outside the 250 millisecond interval. This pulse train is then displayed on a computer controlled CRT for a short interval, after which the same routine is repeated as many times as desired. After each pass the pulse train produced is added point by point onto the sum of all previous pulse trains producing a pulse occurrence histogram, two of which are illustrated in Figure 11.

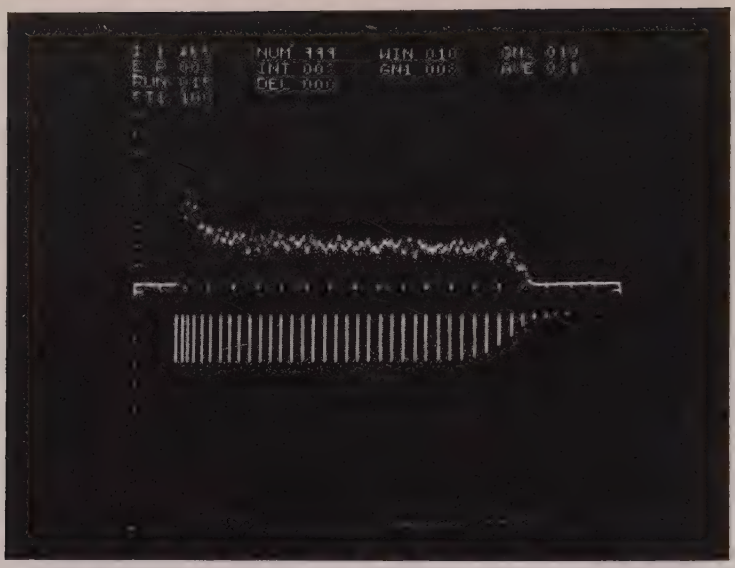


b) 20 msec Rise Time $I_{max} = 10{,}000$

Smoothing or filtering of these histograms is accomplished by performing a sliding window averaging operation to compare the results with experimental data which has been averaged in the same way. Most of the results presented in this paper have been averaged with a 2 millisecond window.

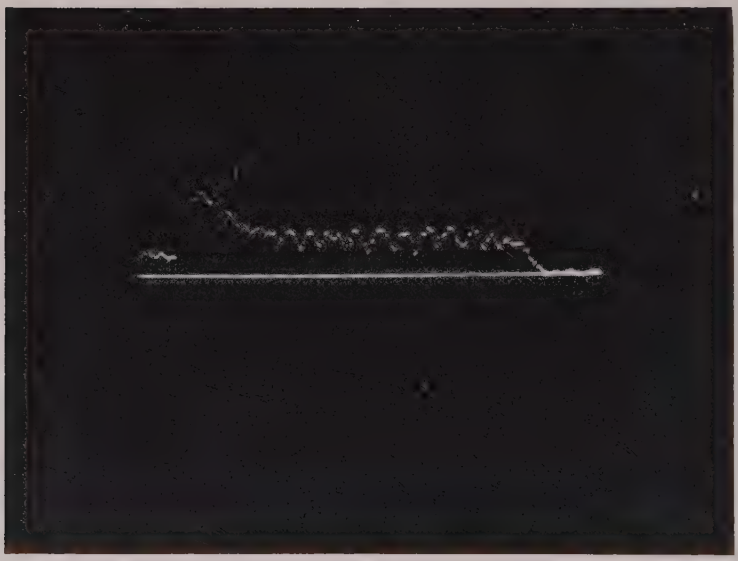
DISCUSSION

Figures 12 and 13 show model response histograms compared with those recorded from real neurons. In these cases the resemblance is very close. The model response in Figure 12 shows an initial peak of about the same duration in time as the real neuron and of the same relative amplitude

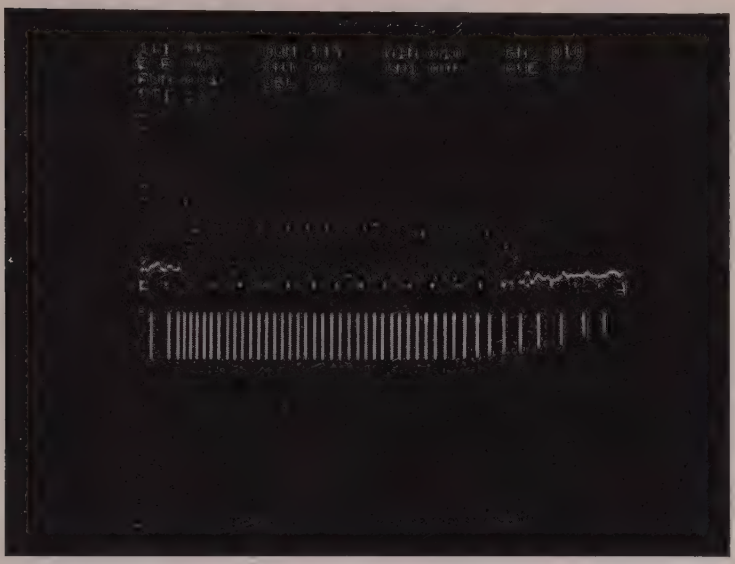


a) Real Neuron

Fig. 12. Averaged Pulse Occurrence Histograms
Input — 1000 Hz Tone Burst 10 msec Rise Time



b) Model



a) Real Neuron

Fig. 13. Averaged Pulse Occurrence Histograms
Input — 333 Hz Tone Burst 5 msec Rise Time

b) Model

with respect to the activity after adaptation. The adaptation of the real neuron is complete in about 16 msec while that of the model takes about twenty and the average slopes at the right end of both histograms are about the same. The real neuron was subject to a 1000 Hz signal of 100 msec duration. The envelope was 10 msec rise time, 80 msec sustained amplitude and 10 msec fall time. This will be referred to as a 10-80-10 signal. Since the frequency was 1000 Hz, only the envelope was presented to the model. In Figure 13 a signal of 320 Hz was presented to the real neuron and 333 Hz to the model. The signal envelope was a 5-90-5. This type of nonadapting response was found to be typical of experimentally observed low frequency sensitive neurons. The really interesting feature in this case is that the model is essentially the same as the one which showed such strong adaptation in Figure 12. The negative feedback function was still there, but it did not affect the result because of the low frequency of the input.

Leaving response histograms, let us look at the response in terms of average rate vs. intensity of the input signal. There are two well defined types of rate behavior usually observed in the eighth nerve. Katsuki[4], in particular, has designated them as "parallel ramp type" and "crossed ramp type." The parallel ramp type have a dynamic range of about 30 db,

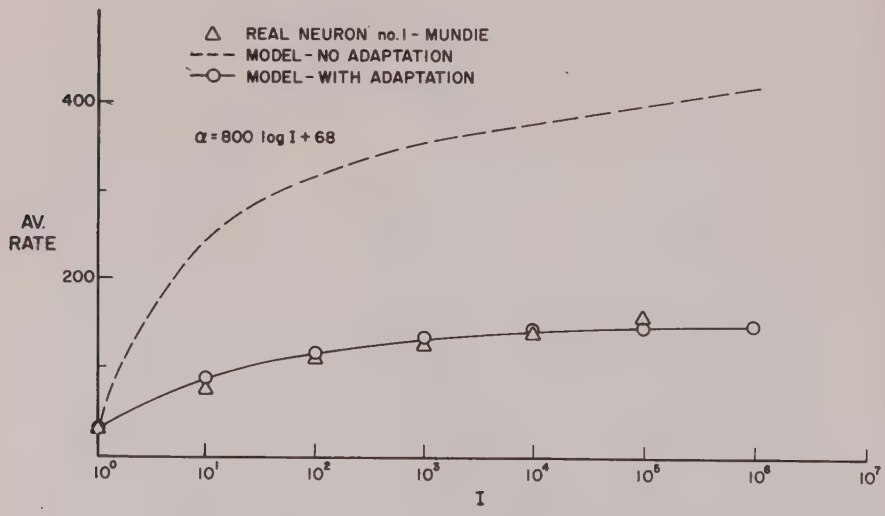


Fig. 14. Comparison of Model Firing Rate with an Eighth Nerve Neuron in the Guinea Pig.

while the crossed ramp type can have a dynamic range of about 100 db. The crossed ramp type are modelled by using the original transfer function $S(t) = k \log I + B$. Response curves for this type are shown in Figures 14, 15 and 16 and compared with actual data recorded by Dr. Mundie's group[5] from guinea pigs at the Aerospace Medical Research Laboratories and by Katsuki from the monkey eighth nerve.

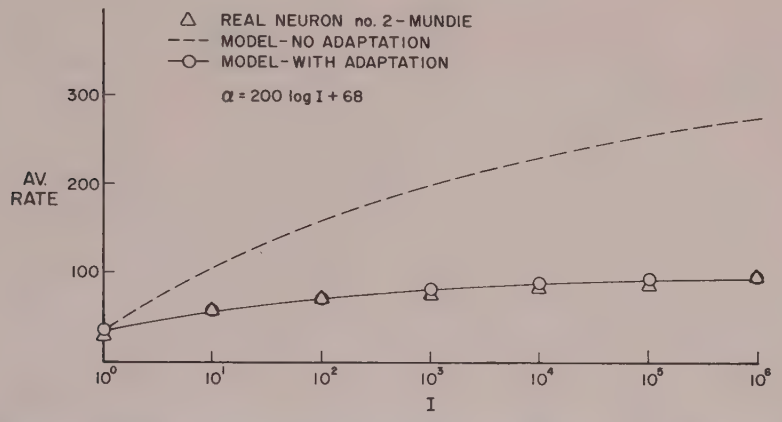


Fig. 15. Comparison of Model Firing Rate with an Eighth Nerve Neuron in the Guinea Pig.

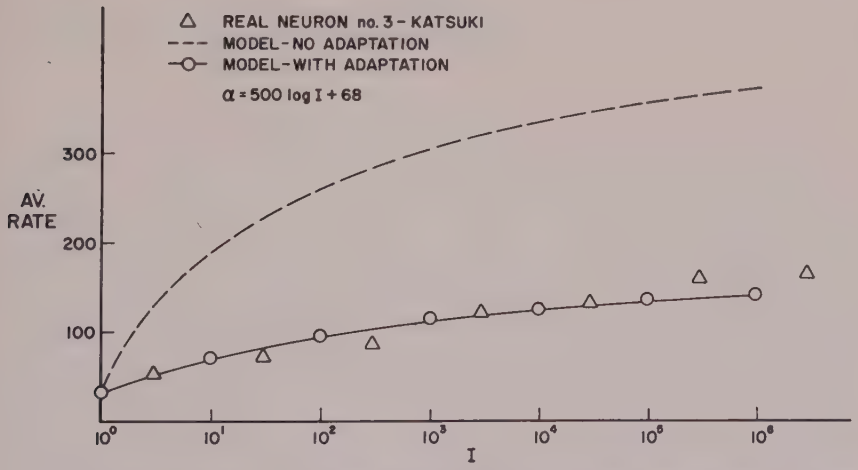


Fig. 16. Comparison of Model Firing Rate with an Eighth Nerve Neuron in the Monkey.

If the transfer function is changed from logarithmic to linear ($S(t) = kI + \alpha_0$) the curve of Figure 17 results and this is compared with three similar responses obtained by Katsuki from the monkey eighth nerve. These are the parallel ramp types, so-called because the thresholds of these

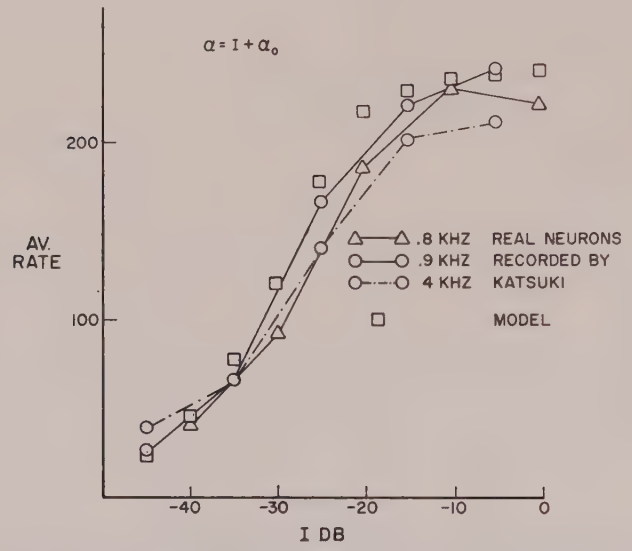


Fig. 17. Comparison of Linear Model with Responses from Parallel Ramp Neurons in the Eighth Nerve of the Monkey.

neurons are distributed over most of the input amplitude range. The threshold of the linear model can be similarly shifted by changing k_1 without altering the shape of the response curve on a semilog plot. Low values of k_1 give high threshold neurons. Curves showing the unadapted average rates vs. input amplitude for three values of k_1 are shown in Figure 18. The behavior is obviously of the parallel ramp type.

A number of real neurons show interspike interval densities for spontaneous firing that are essentially Gaussian and there are some with absolute refractory periods between .3 and .4 milliseconds. Neurons exhibiting the latter feature show response histograms with very little drop in amplitude after the first peak. Since it was certain that this depression in the model histogram was caused by the absolute refractory period, it is no surprise that this depression is not pronounced in real

neurons with very short absolute refractory periods. The model can be revised to use other spontaneous interspike interval densities and this will be done in the future to give the model the capability of matching a somewhat wider variety of responses than it is presently capable of doing.

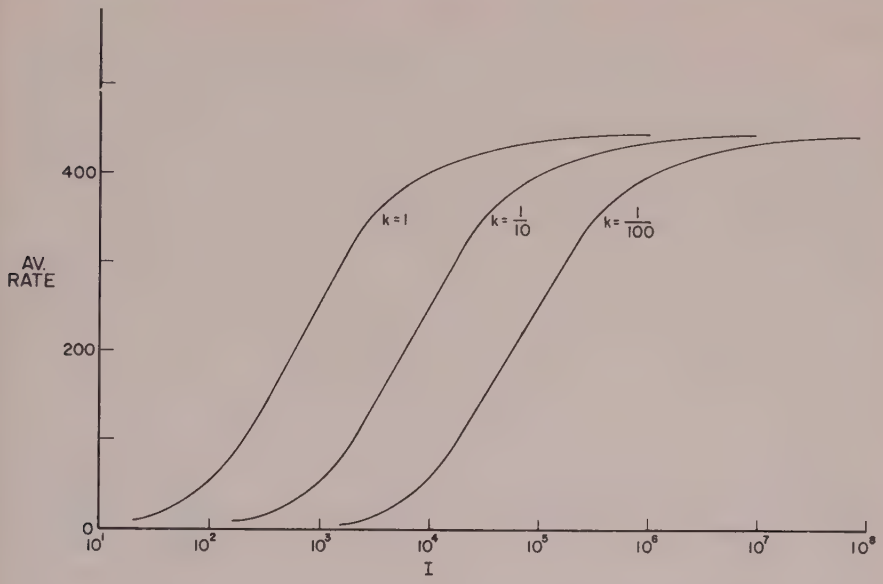


Fig. 18. Parallel Ramp Behavior of Unadapted Linear Model ($\alpha = kI + \alpha_0$).

APPENDIX

Let the interspike interval density for $\alpha = \alpha_0$ be $p(t, \alpha_0)$ and let the density for $\alpha = \alpha_1$ be $p(t, \alpha_1)$. Assume that α increases linearly from α_0 to α_1 in a time interval Δt starting to rise at time t_1 after the last firing. Assume also that the nature of the neuron is such that $p(t, \alpha) = p(t, \alpha_0)$ from $t = 0$ to $t = t_1$ and $p(t, \alpha) = p(t, \alpha_1)$ for $t \geqq t_1 + \Delta t$, this means that the average density in the interval $t_1, t_1 + \Delta t$ must be

$$p_{av}(t_1) = \frac{1}{\Delta t}\left[1 - \int_0^{t_1} p(t, \alpha_0)\, dt - \int_{t_1+\Delta t}^{\infty} p(t, \alpha_1)\, dt\right]$$

The density function then looks like the figure below.

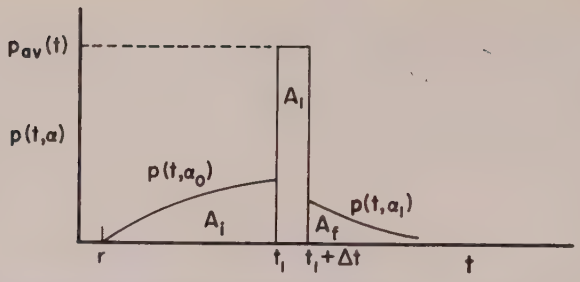


Fig. 19. Interspike Interval Density for an Increase in α from α_0 to α_1 Between t_1 and $t_1 + \Delta t$.

Let us now consider the areas under the curves.

$$A_i = \int_0^{t_1} p(t, \alpha_0)\, dt$$

$$A_f = \int_{t_1+\Delta t}^{\infty} p(t, \alpha_1)\, dt$$

$$A_1 = 1 - A_i - A_f$$

If we let α increase further from α_1 to α_2 in time $t_2 = (t_1 + \Delta t)$ to $t_2 + \Delta t$ the density curve changes again. A_i remains the same, but A_{f_2} changes.

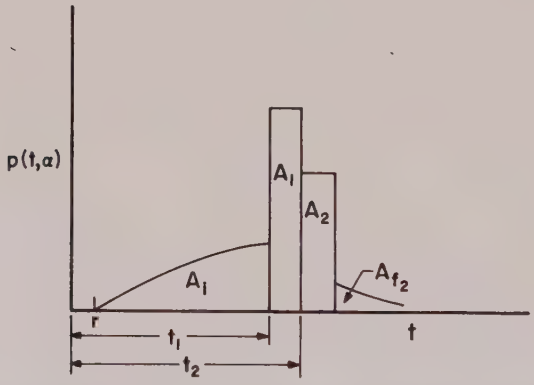


Fig. 20. Interspike Interval Density for Two Successive Changes in α.

Now $A_2 = 1 - A_i - A_{f_2} - A_1$
But $A_1 = 1 - A_i - A_{f_1}$
So $A_2 = A_{f_1} - A_{f_2}$

Likewise if α changes in three steps

$$A_3 = A_{f_2} - A_{f_3}$$

and in general for n steps

$$A_n = A_{f_{n-1}} - A_{f_n}$$

or

$$A_n = \int_{t_{n-1}}^{\infty} \alpha_{n-1}^2 t e^{-\alpha_{n-1} t}\, dt - \int_{t_n}^{\infty} \alpha_n^2 t e^{-\alpha_n t} dt$$

$$= e^{-\alpha_{n-1} t_{n-1}} [\alpha_{n-1} t_{n-1} + 1] - e^{-\alpha_n t_n} [\alpha_n t_n + 1]$$

If we specify that $\alpha(t)$ be a continuous function, and that all the derivatives exist, we may take the limit of $A_n/\Delta t$ as Δt goes to zero to get the density at t_n.

Let

$$\alpha_n = \alpha_{n-1} + \alpha'_{n-1}\Delta t + \frac{\alpha''_{n-1}}{2}(\Delta t)^2 + \ldots$$

for small Δt

$$\alpha_n \sim \alpha_{n-1} + \alpha'_{n-1}\Delta t$$

also let

$$t_n = t_{n-1} + \Delta t$$

then

$$A(t_n) = e^{-\alpha_{n-1} t_{n-1}} [\alpha_{n-1} t_{n-1} + 1] - e^{-(\alpha_{n-1} + \alpha'_{n-1}\Delta t) t_n - (\alpha_{n-1} + \alpha'_{n-1}\Delta t)\Delta t}$$

$$\times [(\alpha_{n-1} + \alpha'_{n-1}\Delta t)\, t + (\alpha_{n-1} + \alpha'_{n-1}\Delta t)\Delta t + 1]$$

Expanding the second exponential for $\Delta t \ll 1$

$$A(t_n) = e^{-\alpha_{n-1} t_{n-1}} [\alpha_{n-1} t_{n-1} + 1] [\alpha'_{n-1}\Delta t t_{n-1} + \alpha_{n-1}\Delta t]$$

$$- e^{-\alpha_{n-1} t_{n-1}} \{1 - [\alpha'_n \Delta t t_{n-1} + \alpha_{n-1}\Delta t]\} \{\alpha'_{n-1}\Delta t t_{n-1} + \alpha_{n-1}\Delta t\}$$

Now

$$\lim_{\Delta t \to 0} \frac{A(t_n)}{\Delta t} = e^{-\alpha_{n-1} t_{n-1}} [\alpha_{n-1} t_{n-1} + 1] [\alpha'_{n-1} t_{n-1} + \alpha_{n-1}]$$

$$- e^{-\alpha_{n-1} t_{n-1}} [\alpha'_{n-1} t_{n-1} + \alpha_{n-1}]$$

or since $t_n \to t$

$$p(t, \alpha) = e^{-\alpha t} [\alpha' t + \alpha] [\alpha t]$$

$$= \alpha^2 t e^{-\alpha t} \left[\frac{\alpha'}{\alpha} t + 1\right]$$

When $\alpha' = 0$ (that is, the stimulus is not changing) $p(t, \alpha) = \alpha^2 te^{-\alpha t}$ which is the original density function for spontaneous activity. This function has only been derived for increasing $\alpha(t)$. The cases for decreasing α are much simpler and are indicated in the text.

ACKNOWLEDGEMENT

The research reported in this paper was conducted at the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio. Further reproduction is authorized to satisfy needs of the U.S. Government.

REFERENCES

1. Miranker, W. L. "A Model of Pulse Generation in the Peripheral Nervous System," *Kybernetik, III*, No. 1, January 1966, 13–17.
2. Weiss, T. "A Model for Firing Patterns of Auditory Nerve Fibers," Ph. D. Thesis, Massachusetts Institute of Technology, May 24, 1963.
3. Gerstein, G. L., and Mandelbrot, B. "Random Walk Models for the Spike Activity of a Single Neuron," *Biophysics. J.* **4** (1964) 41–64.
4. Katsuki, Y., Suga, K., and Nomoto, M. "Peripheral Neural Mechanism of Hearing in the Monkey," Contributed Paper Preprints, Bionics Symposium 1963.
5. Mundie, J. R. Unpublished work done at Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base.

GORDON PASK

System Research Ltd.

Richmond, Surrey, England

*A Cybernetic Model for some Types of Learning and Mentation**

INTRODUCTORY NOTE

System theoretic and cybernetic concepts played a creditable part in unifying the Gestalt and the behaviouristic views of psychology, long before the jargon of system theory was well known. Intermediary constructs such as Tolman's[51, 89] "sign learning" and Bartlett's "schematisation" rely upon ideas of goal achievement and organisation and owe their strength to these notions. However, they are descriptive models (for example Bartlett's[84] schemata explain the odd ways in which stories, or patterns pictures are recalled and reorganised) and they commit us to an attitude about mentation, rather than a structure of what goes on in the brain.

Recent methods of physiological and psychological data gathering make it possible to consider structural (as against descriptive) models quite seriously and since they are highly predictive it is advantageous to do so. But, largely due to the influx of data and the greater depth of explanation that is expected, we are currently faced with a number of gaps between knowledge. They are just as broad as the gap that used to separate "Gestalt" psychology and "behaviourism", though they are nowhere near so contentious. The fact is, there are many sorts of models for mentation, each sort having its own separate experimental technique. Let us illustrate the point.

First, there are models for cognitive structures, exemplified by the work of Hovland[90], Hunt[91] and (in psycholinguistics) of Chomsky and his school. These are structural models for symbol systems, and, in the ab-

* The research reported in this paper has been supported by the Air Force Office for Scientific Research, under contract AF 61 (052)-640 through the European Office for Aerospace Research.

stract, they are tied to systems of artificial intelligence (such as those of Minsky[74], Selfridge,[92] Newell, Shaw and Simon[93]).

Next, there are maturational models, notably Piaget's[94] in which the hierarchical organisation of an effective symbol system interacts with the structure imposed by development (similar systems are encountered in embryology where an organisation persists as a function of genetic, epigenetic and metabolic processes; here, a cybernetic approach has been quite successful.[95])

Next, there are models in which the organism is viewed in close relation to its fellows or its environment; on the one hand, there are social and linguistic models like Vygotsky's[96] and (in a different field) Bateson's homeostatic structures;[97] on the other hand, the models of ethology.

Finally, there are structural models in neurology and physiology that talk about brains directly and take their support either from physiological measurements or detailed behavioural experiments.

Each sort of model and each sort of experiment is concerned with the same thing; namely, the behaviour and character of mind and organism. One is no better nor worse than the other, in principle, and there are few obvious contradictions. But it can certainly be argued that some unification is needed to pull the threads together. In this paper, we shall tackle the problem with a cybernetic "theory" or "class of related models" which is broad enough to comprehend symbol systems, languages, and homeostatic mechanisms and, in a tentative way, some of the physiological constructs also.

1. THE CYBERNETIC MODEL AND ITS ASSOCIATED OBJECT LANGUAGE

1.1. Introduction

Any scientific model that is proof against tautology*[12] must be reducible to basic units or indivisible elements from which models of the given class are assembled by the application of specific composition rules. This

* The issues of tautology and reduction have been discussed, in the milieu of psychology, by Deutsch.[12] The principal difficulty with models that are not reducible is that descriptive class properties such as the property of "learning," are readily confused with explanatory constructs, such as some "mechanism of learning." Reducible models are proof against tautology and preferred as vehicles for explanation because, within them, this confusion is illegal.

paper is concerned with cybernetic models, that is, models[9, 10] which are reducible to control units. We shall confine our attention to cybernetic models identified with systems that behave and learn; either in order to describe these systems (in the case of tentatively verified models) or to pose testable hypotheses about them. The main contention is that cybernetic models are particularly apt representations of psychological and biological events or structures; they are (in a sense we discuss later) natural representations, and they accommodate a greater number of unifying principles than the other sorts of model that may, alternatively, be adopted.

A fairly thorough account of the control units and composition rules of cybernetic models is attempted in section 1.3. But, before embarking upon this project, it will be useful to stand back and broadly review the framework in which models of this type are manipulated.

Let us first recall a distinction considered by Cherry[3] between an "observer's metalanguage" and an "object language" or set of object languages. The observer's metalanguage is used to describe the model, the experimental situation, and the identification established between attributes of the real system and variables in the model. Insofar as components of the model communicate, they are described as doing so within an object language which is part of the model and is completely specified in the metalanguage. The models we shall examine do, always, involve communication. There is always an organism that interacts in a symbolic, rather than a solely energetic, sense with the experimenter or its environment. This symbolic interaction takes place *in* and according to the syntatic rules *of* the object language. But we do not preclude symbolic interactions between several organisms, several parts of the same organism, or several regions in its brain. These modes of communication also take place in the object language.

When the physical system is man-made (for example, when it is a telegraph system or a data processing system) the form of object language is chiefly dependent upon the model maker or observer; he need only satisfy the canons of simplicity, consistency and good practice. In contrast, when dealing with living systems, the choice of an object language must be attuned to the existing and often bizarre modes of animal communication (and the experimental design must take account of these natural constraints, if the experiment is to yield sensible data).

In the simplest case a single sensory modality of an organism receives

an input from the environment. But the organism views a universe of discourse rather than the flux of physical events to which its receptors are sensitive. Consider, for example, the frog's visual system, which has been elucidated by Lettvin, Maturana, McCulloch, and Pitts[6] or the pigeon's visual system, as described by Maturana.[59] In each case there are perceptual filters responsive to properties of the real world that act as the coordinates for a universe of discourse. Particular subsets of points in this perceptual "space" are "signs" insofar as states or percepts within such a subset of points elicit cogent responses (jumping away from a shadow or eating up an insect) that serve to satisfy an organic goal and maintain the animal in dynamic equilibrium with its surroundings. Here, the alphabet of the object language is the set of signs for percepts and actions; its syntax consists in constraints upon usage (for example, the small moving object detector in the frog visual system is a natural device that connotes an object language expression). The experimenter needs to respect these constraints; for the animal is usually unresponsive to events that appear to be sign like in the domain of his descriptive metalanguage.

More generally, the alphabet of the object language for a given species is the pertinent collection of "releasers"[4] and sign stimuli and the patterns of stereotyped behaviour that are "released". The syntactic constraints are imposed by social convention or by an aggregate of individual control programmes of the sort we shall consider in section 1.4. For laboratory experiments with higher animals, this alphabet of innately determined or imprinted signs is often augmented by conditioned stimuli, which assume sign value as a result of the conditioning process (but the importance of the underlying innate structure of signs, even in this type of experiment, has recently been stressed by Konorski).[5]

The alphabet of an object language may thus be regarded as the set of signs that prove effective within the "specific" sensory motor system relevant to the experiment ("specific" in the sense of a sensory motor system selected by the animal's orienting reaction[7]). When more than one specific sensory motor system is relevant, for example in experiments entailing changes of attention, it may be necessary to introduce more than one object language for the constraints upon usage will, in general, differ amongst different specific sensory motor systems. Certainly, experiments with the attention directing mechanism involve at least a pair of object languages for, although the specific and non specific inputs of an

organism deal with many common signs, the syntactic constraints upon specific and non specific operation are utterly dissimilar.

In one respect, experiments with a human subject are curiously simple; the human orienting reaction habituates very rapidly against inputs that are not sign valued;[8] hence, in most experimental conditions, we can rest assured that only a specific operation is involved. A stimulus becomes a sign when we instruct the subject to attend to it. The experimental instructions thus determine, or are intended to determine, the current and restricted domain of the object language. But in other ways man presents the experimenter with peculiar difficulties. The subject, as well as the experimenter, may adjoin terms to the object language. Further, he can construct genuine abstractions by naming coherently associated groups of terms; in other words, the unrestricted object language for experiments with man is a self-referencing and open ended language like any natural language.* To avoid the ambiguities of reference and meaning (Gorn[19] calls them pragmatic ambiguities) that appear when the object language, L, approximates a natural language in its flexibility, it is necessary to stratify L into distinct levels of discourse so that $L = L^0, L^1, \ldots$ Indeed, as we argue later, L must be stratified if we aim to avoid these ambiguities of reference and meaning in experiments with rather lowly creatures.†

Stratification is introduced to avoid difficulties which, in their most pronounced form, give rise to logical paradoxes of the "class of all classes" sort exhibited by Russell (and which Russell[72] resolved by introducing a theory of "logical types"). To avoid these paradoxes we must somehow arrange to distinguish between talk about distinct entities such as "objects" and "classes of objects". Stratification yields exactly this sort of discrimination. In the present context, L^0 will be a level of discourse at which it is possible to evaluate variables, to designate states denoted by L^0 signs and to manipulate these entities; however, we cannot name variables

* It is evident from Dr. Lilly's lecture,[64] that the same comments apply to the object languages that are competent in experiments with dolphins. In fact the excellence of Man's linguistic ability is probably a matter of degree. In suitable conditions apes and other creatures may use self-referencing languages.

† The properties of L have no direct connection with the properties of verbal or written discourse. The communication system of a dolphin is constructive but not particularly manlike. Conversely, our comments refer equally well to experiments in which man, as the subject, is constrained to communicate with the experimenter in a mechanised symbol system that is neither verbal nor written.

nor can we determine a state description in L^0. Similarly, L^0 admits action statements "do a certain operation"; but statements of a goal, of a class of operations that bring about a desired state of affairs, are disallowed. We cannot say "learn French" or "go to Naples" in L^0 though, if these were primitive operations in L^0 we might say "attend the Technical College", then "enter the language laboratory" or "go to London Airport" and "take an aeroplane to Rome" and so on. In this paper I shall deliberately avoid the issue of what these primitives *should* be. For the present purpose, they *are* the elementary terms in the object language that is discovered to be apt for the experimental subject. It is evident that this definition entails relations between the subject and the experimenter and I have argued in other papers[44] that the denotation of an homogeneous object language (wherein the primitive entities are coherently related) is what Harre refers to as an "ontological class" (it is not at all accidental that an ontological class is denoted, again in the sense of Harre,[71] by a sequence of family continuous observations; indeed, the denotation of a homogeneous object language for organism 0 is precisely the set of primitive observations and actions that are "family continuous" when 0 is regarded as the observer or experimenter). It will be evident, later in the discussion, that the nature of the primitive entities depends upon other than logical considerations, namely upon properties of the fabric and the energetics of the organism and its environment. If, in this case, we take the nature of these primitives as discovered, then L^1 is defined as the level of discourse at which we can name variables, determine state descriptions, issue instructions like "learn French" or "go to Naples" and name goals. The stratification may, of course, be extended and later we shall redefine it by establishing an identity between levels of discourse in L and levels of organisation in an hierarchically organised control system.

Specification of a homogeneous object language is closely bound up with the notion of goal achievement and the closely related notion of control. At this point, we define a control unit, which is the elementary particle in a cybernetic model, as the least organisation able, in Gorn's[61] sense, to perform a complete constructive act in a given object language L; in other words, that is able to perform a genuine abstraction in L. The control unit uses L in its essential modes of prescription, description and instruction, or command; hence its specification involves at least L^0 and L^1 in L. The abstraction generated by a control unit with domain L^0 in L

is the achievement of an L^0 goal, named in L^1 as G. Conversely, an instruction in L^1 to aim for G prescribes a class of goal directed operations and a class of descriptions of L^0 properties that determine proximity to and achievement of the goal G. Parenthetically, if we are able to determine a set of primitive control units, for example, by resource to properties of energy or fabric, then the primitives in L are determined as the L^0 signs for objects in the disjunction of their domains, the L^1 signs are accepted as instructions, and so on.

At first sight, this seems to be an unduly tedious way of stating the obvious. But, in fact, the idea of control is more profound than it is often supposed to be; we are so used to meeting control systems in the familiar and solid context of engineering that their logical calibre is taken for granted. Similarly, the choice of a cybernetic model has profound implications. It is not just a matter of descriptive convenience.

Control units will be properly discussed in 1.3. For the moment, let us examine their character and the different garments in which they are dressed. A control unit is a TOTE unit (a "Test, Operate, Test. Exit Unit" in the sense of Miller, Gallanter and Pribram[21]); it is also a "Command Programme" (the basic unit in a proper logic of commands) in the sense of Rescher;[60] I have argued in other papers[1] that it operates in satisfying material analogy. Finally, if it acts upon a suitable symbolic domain, a control unit is a problem solver. In each guise the control unit is unreservedly an organisation or a programme; in contrast the control system or control mechanism is that which embodies such an organisation as a physical object; it effects the command, realises the TOTE operation, or secures the relation of material analogy.

1.2. The Organisation and the Fabric of Living Systems

A real animal is an organisation; but it is also a physical object. The question arises of whether or not there is an interaction between its organisation properties and its object properties. The reply is affirmative.

If the same question is asked about the interaction between the organisation "computer programme" and the computing machine in which it is embodied the reply is negative, unless we countenance some technical and logically irrelevant limitations. Computers are designed to secure this reply. The computer world, technicalities apart, is not concerned

with real time or the decay of real parts that goes on as time passes. Regarded as a data processing device, the machine may have unrestricted memory and may act as a Turing machine of arbitrary complexity. Regarded as an Artificial Intelligence, the paradigm computer is tarnished and concepts of the "cost" and the "efficiency" of computation merge with the pure organisation of the system;[35] but the broad features of the paradigm are still apparent. They are adumbrated by the comment that the physical fabric of the computing machine ("fabric" is used in the sense of Beer[63]) is irrelevant to the programme or organisation so that the physical stability of the control systems that embody the prescribed control units may be taken for granted.

But, when dealing with the "computing machine" that is a real brain, the fabric *is* relevant. The stability, permanence, and availability of the mechanism needed to embody an organisation cannot be guaranteed and the capabilities of the fabric determine whether the operations prescribed by a programme of control units can ever the embodied as physical operators in real control systems (hence, the original affirmative reply).

A complete statement of the constraints imposed upon the development of an organisation by the fabric of a brain would entail very detailed physiological data* which is often unavailable. However, it *is* possible to list a number of statistical properties of biological fabric which are, on the one hand, manifest in the activity of brains and which, on the other may be inferred from abstract models for fabric; models which demonstrate, for example, that any large and rather closely coupled system of physical parts will differentiate into subsytems and will exhibit the pheno-

* In order to make sense of the argument it is essential to maintain the separation between organisation properties and fabric properties and also to avoid the temptation to give primacy to one or other as we do in statements like "if a complete physiological description of the brain were available we should know its organisation." The epistemological issue of whether such a statement is true will not be debated. For the present purpose we have chosen a cybernetic and organisational model "because it allows us to predict or control the behaviour of some living system." The price paid for using this model in the milieu of biology is that organisations are not so unaffected by affairs as they are in the customary milieu of computing machines and, as a result of this, we are bound to countenance the interaction of organisation properties and object or fabric properties.

mena of adaptation and habituation.[62] It will be convenient to state these properties at once. They replace the assumptions that are usually made about computing machines. However, we shall use these properties sparingly and introduce them only when they are required in order to develop the argument. Until the listed properties are explicitly introduced, the usual assumptions about computing machines will be taken to apply. Thus, for example, we initially assume that physical structures (though limited in various ways) are indefinitely stable. The property of fabric which vitiates this assumption (namely that the fabric decays and must be repaired) is explicitly introduced when we consider the phenomena of learning. The fabric properties are designated by the letter *F* and a number

F.1. As a result of maturation and development, biological fabric becomes partitioned into discrete packages. One sort of package is a physically distinct individual; another sort of package is a functionally distinct part; such as a tissue or an organ or a region in the brain responsible for a specific type of computation. The minimal package of *fabric* is that which embodies the minimal component in the cybernetic model, namely the control system that embodies one control unit. All viable packages of fabric, that is, all packages that appear in an experimenter's metalinguistic account as "organism like", are *integral* multiples of "minimal packages" that embody one control unit. This property evidently imposes a principle of functional quantisation upon all physical organisms represented by cybernetic models. The functional quantisation implies and is implied by the observation that organisms are made up from goal directed parts like tissues and organs and control systems involving regions of the brain.

F.2. Within a given package of fabric, there is an activity restriction. Since discrete packages of fabric are identified with control systems, this restriction limits the rate at which physical operators may be applied to a certain maximum value and imposes a limited data processing capability of the sort that is known to exist in the specific sensory motor system of man.[65, 66]

F.3. Certain types of biological fabric also manifest the property that the rate of operator application has a definite minimum value. This sort of fabric embodies those control systems that are "active" control sytems. The increase of Thorpe's[32] "action specific potential" associated with particular releasable and goal directed acts is a special case of the mini-

mum activity restriction, but it is also manifest in the curiosity drive[39] of many animals and the minimum rate at which man must receive data from an input that occupies his attention.[45]

F.4. Biological fabric is abraded by haphazard perturbations, and the structures embodied in it will decay unless they are repaired. The rate of abrasion and the amount of work that needs to be expended in order to repair and maintain a structure depends upon an inherent physical stability of the sort introduced in F.6 below.

If the embodiment of an organisation is observed to persist, then it is legitimate to infer that this structure is maintained and reproduced by an active repair process. We shall later argue that the work expended in this process may be measured in terms of the common currency* of operator applications.

F.5. Biological fabric is malleable or adaptable in the sense of Ashby[62] and the stable configurations are (as in F.1) goal directed control systems.

F.6. Given the argument presaged in F.4, operator applications are required in order to maintain a stable configuration of fabric. However, configurations may be more or less inherently stable (for example, changes in messenger R.N.A. giving rise to the production of a template for a specific message sensitive and message selected protein at a neurone, as in the system tentatively mooted by Hyden,[69] would yield a highly stable configuration; synaptic changes at this neurone a less stable configuration and some sort of excitation loop involving this neurone a still less stable configuration). More stable physical configurations are less readily modified by relevant transformations as well as being, by definition, less readily modified by haphazard events, that is, more stable configurations have greater inertia.

F.7. Consider a control system which is a cybernetic model, *A*, embodied in one of the package of fabric of F.1, labelled *B*. The existence of this control system is equivalently stated as the persistence of the configuration in *B* that embodies *A*. From F.2, F.3, F.4, and F.6, this depends upon the conservation of the operator applications available in *B* over a given interval (or simply upon the conservation of the "effort" in B).

* To show that operator applications do provide a common currency it will be necessary to show, as in 1.4 and 1.6, that the cybernetic model is reducible and that learning, for example, may be conceived as the control of control systems.

The broad requirement is that "effort" should be applied to perform useful work; hence, that the available effort should be distributed in an economic way amongst the types of work that need to be done. However, the economic distribution of effort (which need not be unique) can only be specified when the organisation, A, is defined and when the inertial properties of B are limited. This condition is developed in section 1.7.

F.8. The final property of fabric is really a "principle of cooperation" which expresses the fact that the whole of a system is greater than the sum of its parts or subsystems. This principle has been developed most thoroughly by von Foerster.[36, 37] Let Z_1 and Z_2 be a pair of packages of fabric, such as a pair of individual organisms and let $P(Z)$ be an average payoff over Z, which is interpreted as a determinant of the survival of Z. If so, then $P(Z)$ is superadditive rather than additive so that

$$P(Z_1, Z_2) > P(Z_1) + P(Z_2) \text{ rather then } P(Z_1, Z_2) = P(Z_1) + P(Z_2).$$

Thus, in terms of survival, it pays Z_1 and Z_2 to cooperate. But, in order to establish cooperative interaction, Z_1 and Z_2 must maintain communication. In particular, if there is a condition of the form

$$\tfrac{1}{2} \cdot P(Z_1, Z_2) > P_0 > P(Z_1) = P(Z_2)$$

where P_0 is the least value of P associated with survival, then Z_1 and Z_2 must cooperate in order to survive and they are consequently forced, by the constraints upon their fabric, to communicate with one another.

1.3. The Notation

To avoid a great deal of cumbersome symbolism (in particular, the development of lengthy axiom lists for defining entities which are only slightly manipulated) we shall adopt a graphical notation for representing control units and their composition to form cybernetic models. Nothing is lost by using this expedient, providing we bear in mind that the graphical notation represents no more than the broad form of an organisation; it is accurate but it leaves many details to be filled in. A useful and testable cybernetic model is a specific structure in which these details have been filled in; its detail reflects either a definite experimental situation (such as the adaptively controlled man machine learning experiments discussed in 2.5, 2.6, 2.7, and in 2.8) or a set of definite biological constraints; for

example, in a useful model for the type of learning called "imprinting", definite assumptions are made about innate structures and the "imprinted" behaviour patterns are labelled as "maternal" or "sexual" as befits the animal represented by the model. The act of filling in the framework consists in (I) a proper specification of the alphabet and syntax of $L = L^0, L^1, \ldots$ and (II) the evaluation of the variables in any expressions derived from any of the properties of fabric assumptions that are involved in the model. In our argument, we suppose that these chores have been done.*

1.4. Control Units

Figure D.1 is a control unit in the object language L. It accepts an instruction I, in L^1, to aim for a goal, G, that is named in L^1. It performs an operation, Op, prescribed in L^0, upon a domain which is a state descrip-

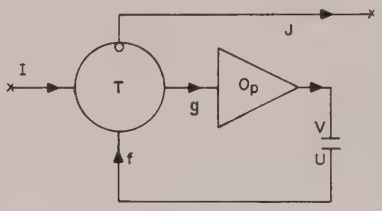


Fig. D.1

tion with elements (or states) denoted by part or all of the L^0 alphabet and having constraints that tally with the L^0 syntax. The domain is partitioned into a set of output states, V, modified by Op, and a set of input states, U, The states of the domain are members of the product set, U, V. In certain cases, U and V may be simple variables; if so, the input and output states are values of these variables (in terms of control engineering,[80] the embodiment of the control unit is acting upon an "observable" and "controllable" system; V is a control variable or free system parameter and the values of U are observations or measurements). The operation, Op, is applied by the action command sequence g, until a test, T, for the

* The chores (I) and (II) have been done in many existing cybernetic models; to cite a few cases, the models of Andrea,[11] Deutsch,[12] Feigenbaum and Simon,[13] Klix,[14] Mittelstädt,[15] Napalkov,[16] Young[17] and myself.[18]

value of a property or L^0 description of the domain, labelled f in D.1, indicates that G is satisfied; in some cases f may be no more than a value of U. When G is satisfied the control unit generates an L^1 statement, J, which usually engages some other control unit; in all cases, it modifies the goal that is aimed for.

To embody the control unit as a control system, we introduce F.1 and F.2 as specific assumptions. The resulting control system is an autonomous physical entity. Its descriptive feedback, f, is internal; this is a sequence of L^0 terms used to guide the application of Op, and g is a sequence of action commands in L^0. No feedback is generated in L^1, at the instructional level, until G is achieved and instruction I is carried out. Further until this moment, the I initiated process may only be interrupted in a traumatic fashion. In other words, F.2 imposes a temporal quantisation upon the activity of any cybernetic model embodied in real fabric, which is analogous to the functional quantisation of F.1.

The control systems act to achieve one goal after another, sequentially. Each of the actions in this temporal sequence occupies a definite interval, a quantum of activity that tallies (through assumption F.2) with so much "effort" (in the sense of F.7) or, loosely, with an L^0 "capacity" that may be evaluated whenever, the cybernetic unit is identified with reality.

If, for example, the control unit represents a pertinent specific sensory motor system of man the quantal interval is the "specious present" and the "capacity" is the span of apprehension or the number of "chunks"* of sensory data that exist in the working register of immediate memory and are simultaneously processed. Alternatively, the control unit may effect some equilibrial behaviour pattern, leading to goal G, and released by a sign stimulus, I. Less specifically, the control unit is, as mooted in 1.1., isomorphic to a TOTE unit (in the sense of Miller, Gallanter and Pribram[21, 25]) or a command programme in the sense of Rescher[60] for bringing about the state of affairs characterised by G on receipt of the command, I, or an analogy operator[1]; the control system, its embodiment, is a homeostat in the sense of Ashby.[20, 22]

We may, at this point, clear up a rather common confusion between feed forward systems and feedback systems. In an L^0 description the unitary control system is a feedback device; a homeostat. On the other hand, if viewed in L^1 it is a feed forward device, like the proprioceptively

* In the sense of Miller.[33]

stabilised leg aiming system in Mittelstädt's[15] model for the Mantis, which accepts an instruction I. In Mittelstädt's system, the instruction I, is delivered from the visual system of the insect and it defines the position of its prey. The internal model built into the "feed forward" system is the L^0 structure, U, V, f and the several "feed forward" systems are assembled into the embodiment of a larger cybernetic model by L^1 feedback paths of various sorts.

If I is single valued (in Rescher's sense, a state command) the control unit is aptly represented by the shorthand form D.2. If I is many valued,

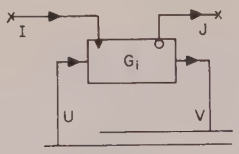


Fig. D.2

so that this instruction selects amongst several goals, the shorthand form is D.3.

At this point, we shall introduce assumption F.3. Given this assumption, an organism is a control system that cannot be turned off and its organi-

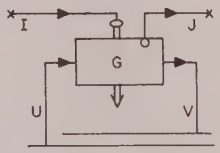


Fig. D.3

sation is an "active" control unit, wherein J must be accepted; in contrast, in a computer programme, J may lead to the instruction "terminate".

Instruction J may legitimately engage some other control unit in the cybernetic model; but if no connections are specified, J may only maintain

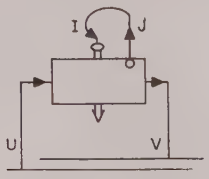


Fig. D.4

the trial making activity of the original control unit itself; as in D.4, which closely resembles a "link" in Deutsch's[12] learning model, the instruction returns to the control unit through an arbitrary list which is the surrogate for a programme. The embodiment of D.4 is an ultrastable homeostatic system. Just as D.2 images some equilibrating and autonomous activity released by an instructional sign, so D.4 images some equilibrating and autonomous activity that is associated with an "action specific potential"[32] and which, consequently, must be released. It also represents a degenerate sort of "curious" system, say a "restless" system, or, in wholistic terms, a precursive sort of "Neophyllic"[38] organism. The degeneracy is due to the fact that an "arbitrary" list is, at most, a device for selecting different ways of achieving the same goal, or members of a class operations that bring about the goal characterised state of affairs. This list has a name in L^1. It cannot contain the L^1 names for different goals.*

In Ashby's sense, the variety of D.4 depends upon the number of entries in the list and it is finite. On the other hand, any programme of control units, of which the list is a peculiar and limiting exemplar, has a variety that need not be finite. In particular, the list may be replaced by some L^1 effective procedure for constructing operations and tests over the domain of L^0. This effective procedure is, itself, an active control unit. But, as we argue later, it is an active control unit with a domain consisting of the goals of L^0 control units, (or the classes of control units that achieve these goals); this domain is L^1 and the constructive or selective control unit is an L^1 control unit.

1.5. Composition Rules

Control units may be composed to form a cybernetic model in parallel (or by "parallel composition") in sequence (by "sequential composition") or in an hierarchical fashion to yield an "hierarchically organised control system" (in the sense of Mesarovic[23] and Tarjan[24]) which is conveniently called a "control of control" construction. The majority of cybernetic models entail each mode of composition to some degree but we shall introduce the modes as separate entities.

The parallel composition of a pair of control units with goals G_1 and G_2 is a simultaneous coupling of these control units. When the coupling

* In Rescher's nomenclature the arbitrary list is a device for specifying the set of "action commands" that realise the "state command" asserted by a single goal.

is accomplished in L^0, through the domain of the units concerned, it is represented by D.5. This organisation images the interaction of a pair of real and distinct control systems, for example, geographically separate organisms or a pair of allosteric enzyme control systems separated by

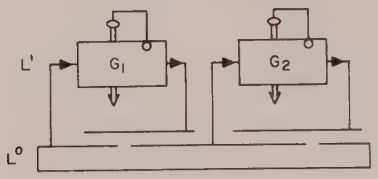


Fig. D.5

chemical specificity. The control systems act simultaneously to achieve the independent goals G_1 and G_2. At the other extreme, represented by D.6, the coupling occurs in L^1, through instructional signals. Here, the domains of the control units may be entirely separate, as shown in-D.6, and the interaction leads to the simultaneous but joint satisfaction of G_1 and G_2. Unlike D.5, the goals of the control units in D.6, are not

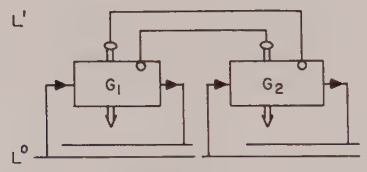


Fig. D.6

independent; indeed there is a sense in which the composition of G_1 and G_2 yields a compromise goal. This comment is cogent just when the control units are embodied to form control systems and when their domains do overlap; in this case the compromise goal may subserve the cooperative interactions assumed in F.8.

The sequential composition of control units leads to an uncontingent programme such as D.7. The coupling necessarily entails L^1 instructional statements and it leads to a succession of autonomous actions or "operation sequence". The successively satisfied goals (G_1, G_2, ... of D7) are. "successive subgoals" of the goal, G, that is satisfied by the entire operation sequence. To emphasise that the instruction I_G selects the entire

operation sequence (and that the uncontingent programme is a real, albeit a limiting entity) the shorthand notation is expanded, in D.8 (organisation D.8 is a TOTE hierarchy which is related to but not identical with an hierarchy of control or of organisation). The control systems

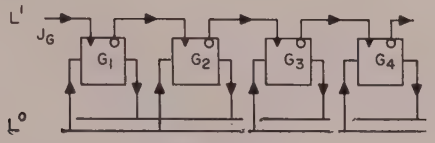


Fig. D. 7

obtained by embodying these figures may be isomorphic to chain reflexes, such as the jump reflex in a hypothallamic cat. On the other hand, if the several control operations are identified with coherent patterns of behaviour released by $I_1, I_2, \ldots$ the operation sequence may represent one of the behoviour sequences encountered in ethology.

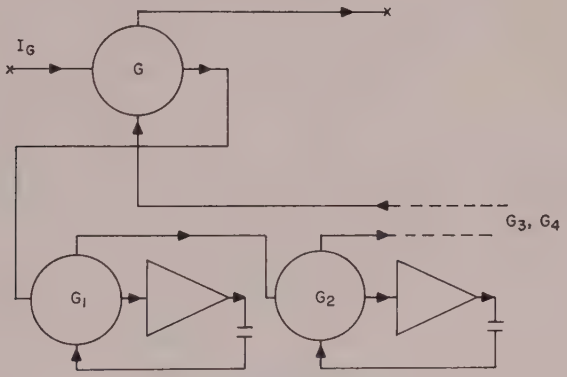


Fig. D. 8

Hybrid forms of parallel and sequential composition are important. D.9 is a commonly occurring organisation and it could, for example, be used to describe the system of interaction between the male and the female stickleback elucidated by Tinbergen[26]. The programmes α and β are sets of innately determined constraints selected in the male and female stickleback when these animals are in a condition for copulation; the domains x, y, and u, v, contain states of the same physical system and

thus overlap. The merit of D.9 is that it succinctly indicates the status of "symbolic goals" and "sign stimuli" with respect to the stimuli and responses of the animals.

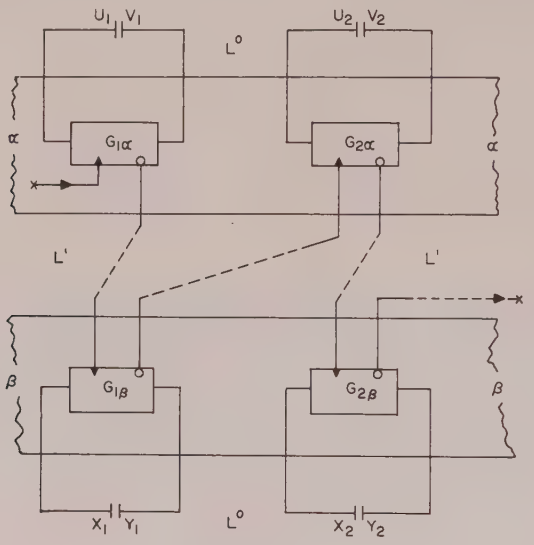


Fig. D.9

The "control of control" construction, an hierarchy of control, is D.10. In D.10, the domain of control unit A^1 is the set of goals and properties of the class of L^0 control units (of the form D.2) that satisfy goals pre-

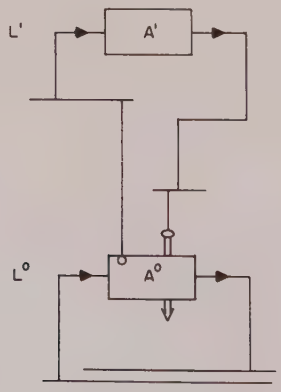


Fig. D.10

scribed by the output of A^1. These control units in A^0 will be labelled as a_i^0, so that A^0 is a possibly ordered collection of a_i^0 in $\{a_i^0\}$. Thus A^1 is an L^1 control unit; any a_i^0 is an L^0 control unit. The domain of A^1 is denoted by L^1 terms; the domain of any a_i^0 in A^0 is denoted by L^0 terms. In this way, we establish the correspondence, promised, in 1.1, between the levels of discourse $L^0, L^1, \ldots$ in L and the levels of organisation in an hierarchically organised control system. Further, since any a_i^0 has the form D.2 and since any control system is the embodiment, say b_i^0 of some a_i^0, the assumptions of F.1 and F.2 establish the denotation of L^0 (the alphabet of the homogeneous object language of the cybernetic model is the denotation of the disjunction of the domains of the b_i^0 it includes), which provides the set of "primitive L^0 terms" required in 1.1. The set of primitive L^1 terms is the disjunction of the domains of all control units of the form A^1 and of any further instructional terms that may be accepted by A^1 and specified in an experiment.

Organisation D.11 is a useful special case of D.10.

To obtain D.11 the generic goal property evaluator is replaced by a

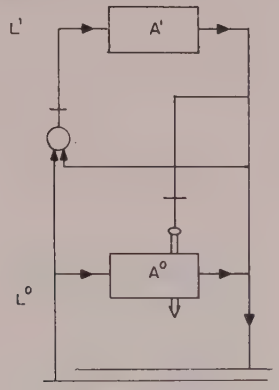


Fig. D.11

comparator that examines the product set U, V, to evaluate some function of the difference between the immediate state and the goal state. This representation is useful when describing experimental situations in which certain mechanical components serve as comparators and produce signals that are used by the experimenter or that are returned to the subject.

There are various embodiments of these figures. First, D.10 is the organisation of a system that is genuinely "curious" rather than merely

"restless" in its demeanour. Next, organisation D.10 may be identified with an hierarchically organised control system (one familiar physiological exemplar is the alpha and gamma neuromuscular control system). Alternatively, it san be identified, at a behavioural level, with Tinbergen's hierarchy of drives and programmes and behavioural sequences.[26] In psychology it images the attitude control of an organism. McCulloch and Pitts[27] model (for the mechanism that presents a visual form in standard position on the foveal region of the retina, and thus determines an organism's attitude with reference to a specific visual sensory motor system) is a particular case. McCulloch and Pitts have argued that the arrangement is broadly interpretable; for example, the system may orient problems in the standard fashion required for their solution. One version of their model appears in D.12 which is readily equated with D. 11.

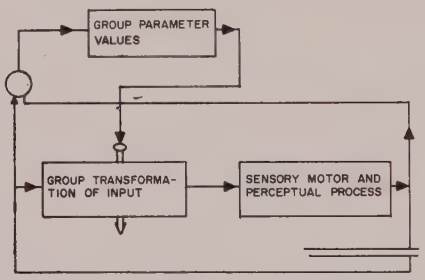


Fig. D. 12

Such devices are classifiers that extract invariants of groups of transformations of L^0 objects and assign L^1 names to the characterised class. The crucial point is that they classify within an homogeneous domain; denoted by L^0 signs. Neither a group of transformations, nor the looser structures that Weiner[28] has suggested may approximate a group in the calculus of a real animal, can be defined over a domain lacking the homogeneity of a single object language, in the sense of section 1.1. Consequently, a more elaborate model is needed in order to represent those facets of an organism which are involved in experiments with its attention directing and orienting mechanisms. In fact, a couple of problems must be distinguished.

First, each modality of an organism may require a separate homogeneous object language; for example, the octopus appears to handle primitive visual signs and primitive tactile signs in different relational

frameworks; further its abstractive mechanisms are separately organised (this beast does not appear to generalise from the abstraction "visual square" to the abstraction "tactile square", though the evidence for this assertion is incomplete). Supposing that this is so, the minimal cybernetic model for an octopus is D.13, wherein L_1^0 and L_2^0 are the object languages

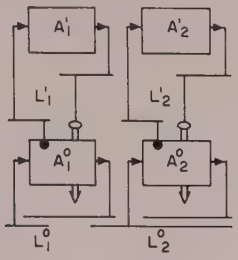


Fig. D. 13

pertinent to the visual and tactile modalities; L_1^1 and L_2^1 being similarly distinguished (the overlap of the L^0 domains is not accidental; the visually directed responses of an octopus *do* alter its tactile input; the tactile directed responses of an octopus *do* modify its visual input, but the control systems for each modality are separate). On the other hand, if the abstractions of an organism can be generalised, so that there is some common sign in L^1 for "square" or "circle", whether it is a visual or a tactile square or circle, then the minimal cybernetic model is D.14. Finally, the limiting case is D.10 which obtains if either we confine

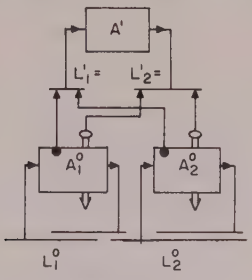


Fig. D. 14

our interest to a single specific sensory motor system or if the organism has an homogeneous object language with a denotation covering all of its modalities.

Perhaps this is so for the specific sensory motor systems of man, and perhaps this is what we mean when we say that man has a symbolic environment. But, even in the case of man, it is necessary to countenance the existence of non specific sensory motor systems and consequently to use cybernetic models such as D.15 when we are concerned with the interactions between the non specific and the specific signals. In particular, D.15 is needed to model the mechanism of the orienting reaction, its habituation and adaptation. In D.15, there are a pair of object languages,

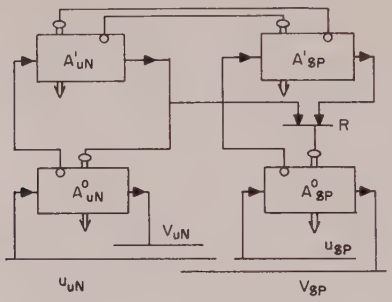


Fig. D.15

L^0_{Sp} and L^0_{Un}. Of these, L^0_{Sp} is a natural object language, in the sense of 1.1. though it is, of course specified in the experimenter's metalanguage; L^0_{Sp} is the object language of a specific sensory motor system (in man and a few other organisms, perhaps of all specific sensory motor systems). L^0_{Un} is no more than a construct invented by the experimenter; its alphabet denotes those discrete events or states which the experimenter believes the organism can appreciate and produce (cogent stimuli and cogent responses but not, in the sense of 1.1., necessarily signs). Both L^0_{Sp} and L^0_{Un} denote internal events as well as the events in the environment; both are involved in internal homeostasis as well as the control of the organism with respect to its surroundings. Further, any event or state denoted by L^0_{Sp} is also denoted by L^0_{Un}, at any rate if L^0_{Un} is fully specified, but L^0_{Sp} and L^0_{Un} are germane to separate control units A^0_{Sp} and A^0_{Un}.

With this preamble, we identify the domain U_{Sp}, V_{Sp}, in D.15 with the denotation of L^0_{Sp}; in it U_{Sp} is a set of specific or "cognitive" inputs and V_{Sp} is a set of "volitional" acts or responses; similarly, U_{Un}, V_{Un}, is identified with the denotation of L^0_{Un} where U_{Un} is a set of non specific stimuli and V_{Un} is a set of autonomic acts. The domain R is part of the

denotation of L^1 and it contains names for the basic modes of activity in whcich the organism may, at any instant, be engaged. In addition to the selective output from A^1_{Sp} and A^1_{Un} some organisms may accept external instructions that select from R.

D.15 appears to be the least cybernetic model that is able to embody the crucial features of the physiological mechanisms proposed by Sokolov,[29] Jouvet[30] and others.[31] Although D.15 is a functional model, we may, tentatively, relate A^0_{Sp} to feedback loops involving no higher than thallamic connections, A^1_{Sp} to feedback loops having some cortical components, A^0_{Un} to feedback paths traversing the core of the midbrain reticular formation and its lateral parts and A^1_{Un} to those involving the higher or thallamic reticular formation and the closely associated anatomical structures. By adjoining F.4, F.5, and F.6 as properties of the fabric of A^0_{Un} and A^1_{Un} (and choosing suitable parameters) we obtain embodiments of D.15 that habituate against non specific stimuli that are not novel; by interpreting A^1_{Sp} and A^0_{Sp} but not A^1_{Un} and A^0_{Un} as the image of a learning system (and presupposing the argument to be advanced in a moment) we obtain a model that abstracts certain stimuli as signs and thereafter responds to them.

It is apt, at this point, to recall the comments in 1.3. about the limits of the present notation; it exhibits form alone. Useful cybernetic models contain a great deal of detail which is compatible with this framework; strictly, the form is a homomorphism of the detailed model. One model containing a useful amount of detail has been presented by Blum, Craighill, Kilmer and McCulloch,[85] at the present symposium; its activities are homomorphic upon those of an embodiment of D.15 and its relative elaboration is a fair measure of the work that is needed to get from the form to the content of natural things.

1.6. Learning Systems

The organisation of learning is identically control of control. However, the selective operations that are performed by control unit A^1 in D.10 are interpreted as constructive operations that build L^0 control systems. Hence, A^0 depends upon a history of selective operations rather than the selection made at a particular instant of time; consequently, this organisation must be indexed by time, t (or, more conveniently, for

the present purpose, by trial number, n, where each trial in an experiment occupies an interval of ∇t so that $t = n \cdot \Delta t$). We thus write $A^0(n)$ in place of A^0 (and if the control of control construction is iterated $A^1(n)$ also).

The situation is formalised by invoking F.5 whereby $A^0(n)$ is embodied in the malleable fabric of the organism as a physical control system $B^0(n)$. This comment also applies to other levels; thus, A^1 is embodied by B^1. In each case, the control system may be reduced to the minimal components of F.1 and F.2 which embody elementary control systems. These will be designated b_i^0, embodying the control units a_i^0 in the case of the L^0 systems.

We have introduced a distinction between organisations or programmes and the physical structures that embody them or realise them as mechanisms. This distinction may, of course, be trivial; for example, if $A(n) = A$ and if some $B(n) = B$ exists to embody the organisation A, then A and B are isomorphic and need not be separated; this is essentially the situation so far assumed. On the other hand, the distinction is certainly not trivial if there is a contingency associated with embodiment, if there is a restrictive limitation upon the physical machinery, or if this machinery decays.

Within the compass of a cybernetic model, "learning" is "learning to control". Perhaps the simplest exemplar of this process is *D*.16. The L^1

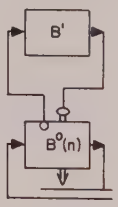


Fig. D.16

control system B^1 has an invariant structure. Thus A^1 and B^1 are isomorphic. B^1 is exploring the collection, $\{a_i^0\}$ of L^0 control units (or homeostatic organisations), a_i^0 from which it is able to select. Only those homeostatic organisations that lead to stability in the L^0 environment become embodied in physical control systems (actual homesstats) b_i^0 and the ensemble or "repertoire" of embodied control systems that is constructed at the n-th-trial in an experiment is $B^0(n)$. The process entails

some arrangement for testing the hypothesis that a particular a_i^0 will be stable before it is permanently embodied in b_i^0 and the arrangement may amount to a rapid decay of any b_i^0 that fails to achieve stability. The L^1 feedback of $B^0(n)$ properties needed to evaluate hypotheses and to guide the exploration strategy is provided by the internal L^1 input to B^1. In fact, it is difficult to interpret D.16 unless we explicitly introduce ideas of physical stability and decay; these ideas have already been suggested in connection with testing and in the assumption that B^1 is invariant. Let us formalise them by introducing F.4 and F.6. From F.4 the b_i^0 in $B^0(n)$ decay and, insofar as they are observed to persist, must be maintained by the reconstructive or literally reproductive activity of B^1 (memory in this sense is relearning); from F.6 we argue that L^1 structures are inertial and physically stable relative to the L^0 structures that decay.

Although this model is fairly elaborate it is open to the criticism that it represents "adaptation" rather than "learning" and this criticism has substance in the degenerate case when the a_i^0 are ordered along a single coordinate. If so, the exploration of B^1 merely adjusts the value of a parameter indexed by this coordinate until the process converges to a stable configuration in which $B^0(n)$ is a particular homeostat.

A rather more complex model that is not prone to this objection is shown in D.17 wherein the embodiment of a control unit is contingent

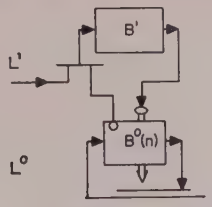


Fig. D.17

upon the value of an external "reinforcing" signal. It is important to notice that such a signal is a term in L^1, not in L^0, for the point is often neglected in the literature. A "reinforcing" signal acts, by definition, as an input to an L^1 control system and serves to guide its exploration strategy in exactly the same way as the internal L^1 feedback.* As in D.16 memory is conceived as a reconstruction or reproduction of the components of $B^0(n)$ within the physically stable framework of B^1. Since either

* This is much the same comment that Pribram[7] makes in connection with reinforcement and the content context relation.

D.16 or D.17 is reducible to control systems it is possible to represent the activity involved in L^0 control, learning or reproduction in terms of the common currency of operator applications or of "effort" in the sense of F.7.

We comment that the learning process in either model is the converse of the control effected by the L^0 homeostats; thus the homeostats act upon their environment to modify it until it is stable and in the converse process of learning the environment acts as the context in which a representative homeostat is modified (or in which selections are made from a set of putative homeostats) in order that stability may be achieved. Apart from the degenerate case, however, a higher order system, B^1, is involved in the converse process of learning. All the same it is reasonable to view the learning in D.16 as a differentiation of homeostats[75] (and in D.17 as an L^1 controlled form of differentiation).

These models for learning can be composed, like control units. Thus, Sutherland's[53, 67] recent learning model is a sequential composition of a pair of the entities in D.17. The first of these is responsible for learning to select a relevant "sensory analyser" (or "perceptual filter") that extracts relevant cues from the environment. The next D.17 unit, which accepts its L^0 input from the L^0 output of the first, represents the learning process whereby (in Sutherland's model) the organism learns to respond in a successful fashion using the cues provided by the already selected sensory analyser.

Further, the control units that appear in the reduction of a cybernetic learning model can be variously interpreted. If, for example, these control units are interpreted as problem solvers (which is one of the possibilities suggested in 1.1.) then learning is reducible to an hierarchy of problem solving processes, as proposed by Newell[34] and others.[35]

To illustrate this point and to comment more cogently upon the acts or construction and reproduction that were mooted a moment ago, it will be necessary to insert some details into the model. For this purpose, I shall use a cybernetic model in which human learning of a problem solving skill is conceived as the context dependent construction of operation sequences, starting from initial sequences that are able to solve, at the most, very simplified (partially or nearly completely solved) problems. We have computer simulated this[18] and other related models[76, 50] and have performed specially designed experiments on human learning to test our ideas.

1.7. Learning to Solve Problems in a Particular Situation†

In the sort of learning that concerns us, the subject knows how to set about solving the problems entailed in the performance of a skill before the experiment begins. His knowledge in this respect is partially innate and partially gleaned from the experimental instructions which, amongst other things, define problems in relation to their solutions (a solution is the result of applying certain L^0 legal operations to a problem).* This is consonant with the further assumption that an L^1 structure, B^1, exists to embody this knowledge and that its form is constant (so that $A^1(n) = A^1$ and $B^1(n) = B^1$); B^1 is also responsible for an initial and context independent construction of an initial L^0 structure $B^0(0)$. We identify $B^0(0)$ with an "immediate memory" mechanism or "span of apprehension" mechanism coupled to the problem domain, in other words, the mechanism whereby the subject attends to stimuli and response alternatives and interprets stimuli as problems and response alternatives as possible solutions. We shall be largely concerned with skills in which the subject is required to solve only one type of problem (the definition of type is circular; the domain U, V, contains problems U in U, of one type only if all of then can be solved by similar procedures; hence the subject's repertoire should contain the embodiments of similar procedures). However, this sort of model is readily extended to deal with the conjoint or alternate solution of several types of problem (and this extension is necessary to deal with issues such as interference between the subskills of a skill and the converse effects of positive transfer of training).

The context dependent part of the learning process is true to the paradigm in D.17. The context for learning is provided by a sequence of stimuli (denoting problems which may or may not be simplified) that is selected by the experimenter and on L^1 knowledge of results signal indicating at the n-th-trial whether or not the problem presented at the n-th-trial has been correctly solved. Problems in the sequence are spaced Δt apart and a solution may only be correct if it is produced within this interval Δt.

The context dependent learning is conceived as the construction and

* The system L^1, L^0, used in this model appears, in retrospect, to be a subsystem of the language used in computor programmed concept accquisition by Banerji.[100]

† The reader may find it helpful, when perusing this section, to refer to D.18. Figure D.18 is fully described in section 1.9.

embodiment of operation sequences in which each control unit is an L^0 problem solver. These operation sequences are variously identified according to the form of problem (if the skill entails the solution of classification problems they are test trees, if it entails transformation problems, they are sequences of object transformations). In any case, an operation sequence will be designated $A_i^0 = [a_{1i}^0 a_{2i}^0 \ldots]$ and its embodiment will be designated $B_i^0 = [b_{1i}^0 b_{2i}^0 \ldots]$. The repertoire at the n-th-trial, $B^0(n)$, is the ordered collection of B_i^0 available to the subject at the n-th-trial and $A^0(n)$ is (equivalent statements) the "L^0 code for $B^0(n)$" or "the organisation of $B^0(n)$".

Given $A^0(n)$, a set $\{a^0\}$ of L^0 operations, and a set of L^0 constructional rules (specified by A^1) we may derive the L^0 legal extensions of $A^0(n)$. Given $A^0(0)$ and a complete goal specification of the sort outlined below, we might even derive $\{A_i^0\}$, the set of all legal $A^0(n)$. But only some of the legally possible $A^0(n)$ in $\{A_i^0\}$ will be realised in the experiment, namely those that are L^0 codes for the $B^0(n)$ that are really built. It is evident that the novice, characterised by $B^0(0)$ or its organisation $A^0(0)$ is allowed a great deal of latitude in how he solves problems and learns to solve problems. His liberty is reduced as he becomes proficient by the repertoire $B^0(n)$, $n > 0$, that has been constructed up to the n-th-trial. Similarly, the possible extensions of $A^0(n)$ are increasingly restricted, as n increases in value.

The L^0 goal G^0, is satisfied by the conjoint satisfaction of goals G_i^0 defined for each unsimplified problem u_i in the set U. Each G_i^0 is itself decomposable into subgoals G_{1i}^0, G_{2i}^0 ... some of which are satisfied when either the subject partially and correctly solves an unsimplified problem u_i in U or, given a simplified (partially solved) problem u_i in U (derived from the original u_i) if he completely and correctly solves it. In each case, the criterion for correct solution involves transforming the given problem in a way that satisfies a stipulated relation between problems and solutions and doing so in an interval of Δt. To be correct, and thus to receive an affirmative knowledge of results signal, the subject must solve problems at a satisfactory rate.

The domain U, V, is highly restricted. It may be derived from a class U of unsimplified problems u_i and a subset of the ordered set $\{A_i^0\}$ such that G_i^0 is satisfied, given u_i, by the application of A_i^0. Partially solved problems are equivalence classes of states that are equivalent to the application of an applicable (and thus a maximum length) A_i^0 to u_i and G_{ij}^0

is the subgoal for a^0_{ij} in A^0_i. Finally, solutions, v in V, are completely solved problems. Within this structure, the experimenter is able to simplify any problem before he presents it (as though he possessed the computing machinery that the novice will acquire, by dint of learning, in this own brain).

The construction of $B^0(n)$ is brought about by a control system B^1 which embodies an organisation A^1. This organisation is partially innate and partially engendered by the experimental instructions (it is innate, insofar as the subject has the equipment to understand and interpret the instructions; but most of A^1 is literally determined by the acceptance of an instructional statement; before the experiment starts, this part of A^1, formally an L^1 statement, is embodied or "written" as the corresponding part of B^1).

The first component of A^1 is an L^1 description, A^1_d, of $A^0(0)$ and G^0 (thus, of all possible A^0_i in $A^0(n)$ that will satisfy G^0). Hence, A^1_d is an L^1 code* for those $B^0(n)$ that may be constructed. The embodiment of A^1_d is B^1_d, a physically stable "writing" of the code A^1_d. In detail, A^1_d is an ordered set of instructions of the form "achieve subgoal G^0_{ij}" and "achieve subgoal G^0_{ij+1}" iterated for each G^0_i entailed in G^0. From 1.1., a goal description is a description of a class of control units that will achieve the goal. This coincides with our initial specification of A^1_d (as determining possible A^0_i) and suggests that a subject who has accepted A^1_d knows how to set about solving the problems in U even though he may not have the physical computing equipment needed for this purpose (in other words, the problem class is defined and its definition is understood by the subject).

Next, A^1 contains prescriptions A^1_p for L^1 problem solving, the embodiment of which is an L^1 control system B^1_p with domain of $B^0(n)$ and a goal, G^1, that is to construct $B^0(n+m)$, $m > 1$, such that G^0 may be satisfied. It is equivalent and illuminating to regard B^1_p as a transducer that decodes A^1_d in the context of the problems posed for solution by the experimenter and in the context of $B^0(n)$ at the n-th-trial. This view of things lays emphasis on the fact that for $n > 0$ the form of $B^0(n)$ constructed by B^1_p depends upon the problem environment as well as the constructional rules.

* Similarly, $A^0(n)$ is the L^0 code for $B^0(n)$ though this assertion is rather trivial until we introduce the further assumptions F_4 and F_6.

The initial action performed by B_p^1 (we conceive it as taking place before the start of the experiment so that is is a context independent process) is the construction* of $B^0(0)$ from the A_d^1 code for $A^0(0)$. Given $B^0(0)$ the L^0 problem domain exists (it makes sense to say that the subject attends to the context of L^0 problems and tries to solve them). Apart from the trivial case when the subject is thereby rendered proficient, this construction also leads to the existence of L^1 problems, for G^1 is not satisfied unless $B^0(0)$ is a control system that is able to satisfy G^0. If it is not, B_p^1 acts to solve the L^1 problem engendered by the deficiencies of $B^0(0)$ or, in general, of $B^0(n)$. It operates upon $B^0(n)$ guided by an L^1 feedback indicating deficiences in the subject's repertoire and in such a way that $B^0(n + m)$.will be able to satisfy G^0.

In the special case we are considering, at least a pair of L^1 operations must be distinguished. These are the *concatenation* of further operators b_{ij}^0 embodying control units a_{ij}^0, at the end of strings of operators B_i^0 that embody some of the A_i^0 in $A^0(n)$, and the *substitution* of parts of existing strings. Of these L^1 operations, concatenation is guided by an external L^1 feedback or "knowledge of results" signal, as in D.17, which is provided by the experimenter. B_p^1 may attempt any legal concatenation; but only those a_{ij}^0 are embodied in b_{ij}^0 which, when applied to a problem as the terminal member of A_i^0, lead to a "favourable" or, possibly, a "reinforcing" knowledge of results signal. In contrast, the substitution process depends only upon an "internal feedback" of properties of $B^0(n)$. The control system B_p^1 aims to shorten the existing B_i^0 by substituting a single operator for a group of operators if it does the same job, and B_p^1 aims to generalise the structure of $B^0(n)$ by substituting a part, B_i†, of B_i^0 in B_k^0 if B_i^0 and B_k^0 perform analogous operations.

This is probably the simplest framework in which learning, the construction of $B^0(n)$, can be reduced to an hierarchy of problem solving.

* Or, as in D.18 and D.19, we may regard $B^0(0)$ as given with the perceptual and immediate memory capabilities of a man. The existing machinery is merely organised by B_p^1.

† As in 2.6. and 2.7. the onus is placed on the experimenter to provide problems that either can be solved or that can be solved after a small extension of $B^0(n)$. Usually, this implies that the experimenter must simplify or partially solve the problems and often it implies, as in 2.5 and 2.6. that the degree of simplification must be adjusted as a function of the learning that has occurred.

Even so, the simplicity is deceptive for we have yet do deal with the issues that arise (I) due to the coincidence of the L^1 problem solving (of the learning process) and the L^0 problem solving (manifest in the subject's behaviour) and (II) due to the decay of the physical operators b^0_{ij} and the consequent need to repair or reproduce the B^0_i.

So far as (I) is concerned, it is usual to assign a priority to L^0 problem solving. In the experimental situation the subject receives problems at a rate of $1/\Delta t$ and if he can solve them (if there is a B^0_i in $B^0(n)$ that is applicable to $u(n)$, the problem presented at the n-th-trial) then he must make the attempt. This priority is formalised as a further instruction A^1_m in A^1 and is certainly not peculiar to the experimental situation (an organism learns in order to control a larger part of its environment but, if it is to survive long enough to learn, its behaviour must control the immediate environment by solving the problems that are immediately posed).

To deal with (II), we invoke F.4, F.5, F.6 and F.7. The initial assumption that A^1 and B^1 are invariant implies that the embodiment of B^1 is physically stable (in the sense of F.6) at any rate relative to $B^0(n)$. It is $B^0(n)$ that decays and it is the B^0_i that must be repaired or reproduced to counter this decay.

However, we have already introduced the operation required for reproducing an L^0 operator b^0_{ij}; to reproduce b^0_{ij} is identically to re-embody or to re-substitute b^0_{ij}; hence, the L^1 substitution process is (in a special case) reproduction.

In the computer programme, we have used to simulate this model each b^0_{ji} in each B^0_i in $B^0(n)$ is associated with a pair of variables "activation" and "availability". The activation value of b^0_{ij} increases whenever this operator is applied in a way that leads to reinforcement (thus, its value increases whenever B^0_i is applied to a problem and solves it successfully); otherwise the value decreases to 0. The availability value of b^0_{ij} increases whenever the operator is reproduced, otherwise it decreases to 0; this decrement represents the decay of the physical structure. The L^1 substitution process (which may reproduce b^0_{ij} by substituting it by itself) is controlled by an L^1 feedback indicating the average difference between activation value and availability value and the operation is applied to the various L^0 operators in $B^0(n)$ in an attempt to match the prevailing availability to the prevailing activation, which is the same as matching availability to successful utilisation. Other schemes can be invented, some

of them far more elegant than this scheme, which have the same flavour and would fit the facts at least as well. The present arrangement is probably the simplest that will suffice.*

Because the cybernetic learning model is reducible, the work done in L^0 problem solving, learning (or L^1 problem solving) and reproduction of the L^0 operators can be expressed in the common currency of operator applications or "effort". To go further, we need some conservation rule for "effort" of the sort briefly indicated in F.7 in Section 1.6.

Probably the simplest conservation rule is derived in this fashion. From F.2 we may write $\lambda_{max} > \lambda$, where λ is the effort expended in Δt and λ_{max} is the maximum limit in F.2. From F.3 we may write $\lambda > \lambda_{min}$, where λ_{min} is the minimum limit. Let these limits approach one another and let Λ be their average value. We shall assume that if the attention of a subject is fully occupied by the experimental data, then, on average $\lambda \to \Lambda$.† Further, if λ_1, λ_r, λ_a represent the effort entailed in learning, reproduction and L^0 application or problem solving, then $\lambda = \lambda_1 + \lambda_r + \lambda_a$.

Not all distributions of effort lead to a stable form of $B^0(n)$. The system can run into a number of interesting and lifelike difficulties either due to the exhaustion of effort or (when the activation values are uniform) a lack of direction for the reproduction process which gives rise to a decay of certain often crucial parts of $B^0(n)$. For the present purpose, we shall only consider one sort of difficulty engendered either by uneconomically long strings of operators, B_i^0 in $B^0(n)$ or by the construction of many different and moderately lengthy strings, each concerned with a different problem in the class U. In either case, the system is unable to maintain all of the structure it has constructed and bits of it fall apart.

* The rule is ubiquitous in biology and the social sciences. We regard the repertoire as a population of b_i^0 and impose the regulations of a homeostatic population control system. The resulting dynamics are identical with those of Wynne Edwards[42] symbolic control systems which regulate the number and type of animals in a local population to a figure that is compatible with the local resources, or their analogues in anthropology, as described by Rappoport.[43]

† One essential feature of the simulated model is that each basic process is autonomous, once it has been initiated; for example if the system "applies" an M membered string of operators then M operator applications are used up, one for each member of this string; if the system "concatenates," a certain number of operator applications are used up. Hence, in any particular interval Δt, the system may use more than Λ applications. If so, it is subsequently starved of them. This is the interpretation given to equality on average or $\lambda \to \Lambda$.

1.8. The Need to Make and Retain Abstract Descriptions

It is evident that the situation we have just considered is likely to occur whenever the subject (or a cybernetic model) is required to learn about an environment that is fairly complex. The difficulty may be ameliorated by the condensation of strings of operators (which takes place as a result of L^1 substitution) but there are various reasons why this process, alone, cannot avoid the basic dilemma. As Minsky[74] and others[40] have pointed out, the genuine resolutions of the difficulty entail an abstract description; the organism must construct and retain a concise and abstract description of its problem environment or its own state, rather than retaining the computing machinery that it uses to solve problems.

In the present framework, an L^1 code is an abstract and usually less complex representation than an L^0 code; in order to resolve its difficulties the model should construct an L^1 description of the organisation $A^0(n)$ which is the L^0 code for $B^0(n)$. This L^1 code, say A_b^1, will have the same form as A_a^1 (the L^1 code for $A^0(0)$ and G^0) and it will be adjoined to A_a^1.

The restriction that $A^1 = A^1(n)$ and $B^1 = B^1(n)$ obviously prohibits such an abstractive process (for if A_b^1 is built and adjoined to A_a^1 then A^1 is not invariant as we assumed). However, this restriction was imposed for the sake of expository and programmatic simplicity; if it is relaxed to allow for an L^1 description of $A^0(n)$ then the actual construction of A_b^1 does not involve any essentially novel process.

To show this point, we notice that the construction of $B^0(n)$ is precisely the production of a first order abstraction which is the L^0 description, $A^0(n)$, of the problem environment; exactly, it is a description in terms of what the organism must do to control its environment; that is, to solve problems. Similarly, the L^0 problem sequence and the L^1 reinforcement signal jointly provide an ostensive definition of $A^0(n)$ and $B^0(n)$ is the particular subset of possible G^0 satisfying repertoires that are jointly realisable in the subject (in the sense of satisfying the various requirements cited in section 1. 7.) and relevant to his experience of the environment throughout the learning process.

The construction of $B^0(n)$ has been discussed in some detail. If a construction of the same type is iterated by a control system B^2 analogous to B_p^1 over the domain of L^0 descriptions (one member of which is $A^0(n)$) it yields the embodiment B_b^1 the code for which, A_b^1, is an abstract de-

scription. In the computer programme that simulates our particular cybernetic model, abstraction is carried out by a special class of the analogy operators we have already mentioned in connection with substitution.* I have argued, in other papers, that systems such as $B_p^1 \times B_b^1$ or $B_p^1 \times B_d^1$ are L^1 concepts[87, 88, 9] the former being a concept acquired by learning); that A_p^1 and A_d^1 are conceptual descriptions (or in a limiting case, concept classes in a cognitive structure); and that B_b^1 and B_d^1 are the written or embodied or descriptively memorised forms of A_b^1 and A_d^1. Similar definitions apply to L^0 concepts and the structures they describe.

On the plausible assumption that the embodiment of L^1 structures is more stable, in the sense of *F*.6, than the embodiment of L^0 structures, the abstractive process is a device whereby an organism can retain what it has learned, or can learn in stages (rather than all at one) or whereby it can learn the broad outline of a situation and fill in the details later. It is true, of course, that *some* of the L^1 structure must be innate (the genetic programme must give the organism the codes it needs for "understanding" signs and for constructing and reproducing structures). In some organisms *all* of the L^1 structure is innately determined. The salient

* Let a, b, c and d be objects with properties $P(a, b)$ and $P(d, c)$. Let a and b belong to one universe of discourse (one homogeneous domain, in the sense of 1.1.) and let d and c belong to one (usually different) universe of discourse. Let f be a systemic or dispositional relation $a(f)\,b$; similarly g in $d(g)\,c$. Let R be a relation of similarity or dissimilarity between the property sets $P(a, b)$ and $P(d, c)$; in the sense of Hesse,[68] R is an analogical relation that assigns relevance, irrelevance, or "indifference" to these analogical properties and the entire structure is an analogy of the form "a is to b as d is to c." I have argued, in other papers,[70] than an analogy operator is a control unit that given $a(f)\,b$ and R, operates upon the domain of d and c until the goal $[d(g)\,c] = d[R(f)]c$ is satisfied.

Analogy operators of the sort involved in substitution act upon the domain of L^0 descriptions, which includes the universe of discourse of a and b as well as d and c; for example, in the case when a subject is learning to solve more than one type of problem such analogy operators may construct strings of operations applicable to type 1. problems that are analogous to type 2. strings that already exist. In contrast, the analogy operators responsible for abstraction are given $a(f)\,b$ and R, where a and b are control units embodied in operation sequences in the repertoire (with codes that are L^0 descriptions) and where R is a relation between L^0 descriptions and L^1 descriptions (for example, this relation may determine a homomorphism); d and c are L^1 objects and the analogy operator constructs the analogous and abstract representation $d(g)\,c$ which is an L^1 description of $a(f)\,b$. We comment that the decoding whereby B_d^1 is represented as $B(0)$ is the converse of this analogy operation. Thus B_p^1 is a converse analogy operator.

peculiarity of higher organisms, and in particular of man, is that other-than-innate codes such as A_d^1 and A_b^1 may either be constructed by abstraction from experience or they can be introduced and written into the system as L^1 statements; typically, experimental instructions. In our account of learning, A_d^1 was introduced in this fashion.

1.9. The Representation of Concept Learning and other Processes*

It will be useful to adopt the graphical convention that a shaded box is an uncommitted box. A shaded control unit means a control unit (at some position in a programme) that can be substituted by several different control operations (the actual substituent being embodied in a physical operator). If a complete box is shaded, like the L^0 box in D.18(I), this means that no relevant programme is specified. However, D.18 (I) is not to be interpreted as "the subject has no L^0 control system". It should be interpreted as "the subject has no relevant control system", that is, "he is attending to something that we and the experimenter regard as irrelevant". The further convention that dotted lines convey no signals is useful but inessential. In D.18 (I) for example, there is no L^1 "knowledge of results" signal, and no problems are delivered to the subject.

The cycle of learning processes may now be imaged graphically; we show a sequence of four stages in the cycle in D.18. The initial and context independent part of the process appears in D.18 (I) where the L^1 statement of A_d^1 and A_m^1 is introduced as an instruction. In D.18 (II) B_p^1 constructs an L^0 programme from the written embodiment B_d^1, B_m^1 of A_d^1, A_m^1. This L^0 programme is the L^0 expression of the construction and substitution rules and it is adjoined to the initial string of embodied operators in $B^0(0)$. In *D*.18 (III) at the n-th-trial, $n > 0$, some of the control operations have been substituted by legal substituents and these substituents have been embodied as physical operators. This is the context dependent part of learning. Without unduly elaborating the image it is difficult to show the abrasive effects of decay. In D.18 (IV) B^2 generates a description of $B^0(T)$ where T is some terminal value of n, and this description is embodied in B_b^1. The abstract description may be

* On the assumption, embodied in D.18 and D.19, that $B^0(0)$ is part of the innate perceptual machinery and that B_p exists. The latter, including the λ distribution process, may be tentatively identified with the limbic system in the brain of mammals or, in octopus, with the vertical, medial and lateral superior frontal lobes.[17]

subsequently retrieved, as an L^1 statement A_b^1 made by the subject, even though $B^0(T)$ has decayed.

If the subject is instructed to learn to solve a pair (or more) of different classes of problem then there must be a pair (or more) of the shaded boxes of D.18 (I) as in D.18 (V). The pairing is preserved throughout

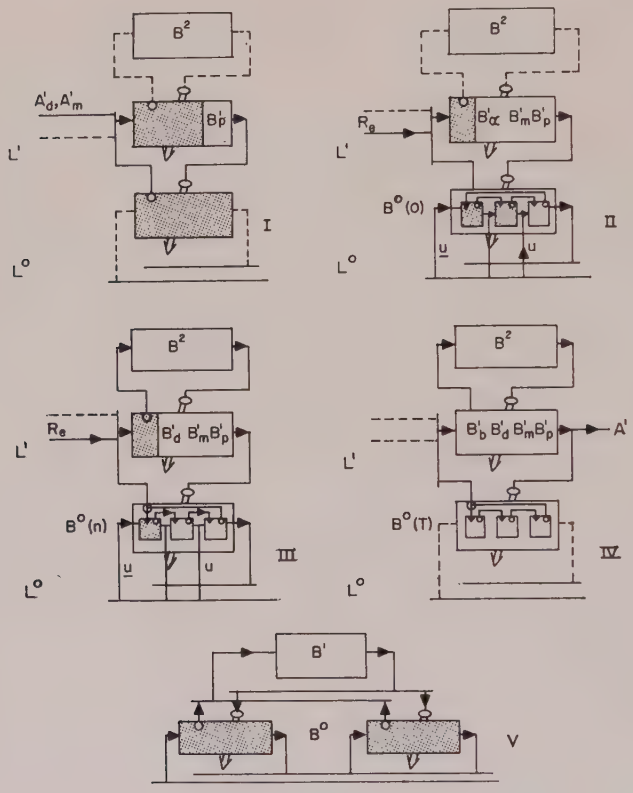


Fig. D.18

the cycle and the relevant box is selected by an L^1 signal designating problem type.

Some clarity is gained by splitting the boxes into a part concerned with the written embodiment of codes and another part that acts as the decoder or coder or transducer control system. The construction is illustrated by D.19 wherein the stages, D.19 (I), (II), (III) and (IV) replicate stages in the cycle considered a moment ago.

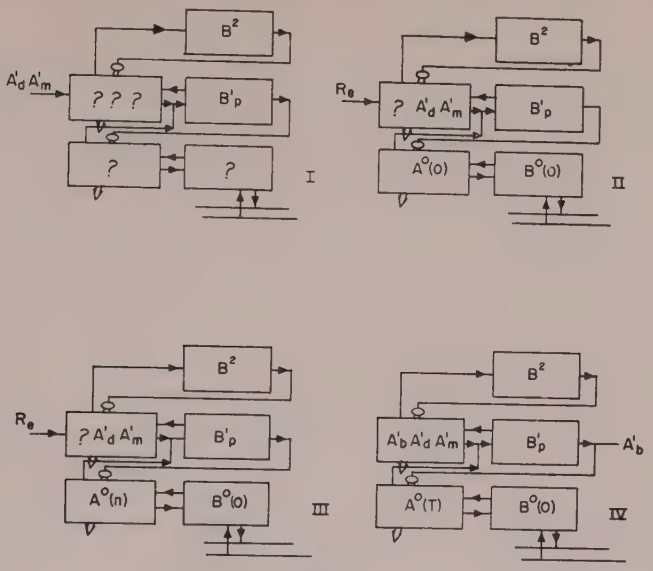


Fig. D.19

For organisms that cannot accept instructions, the act of learning an operation sequence is reduced to the paradigm of D.20. The embodied codes are innate. Latent learning and other processes that do not depend upon external "knowledge of results" may be represented by D.21. Here, the embodied codes are innate and the L^1 goal (often called a "symbolic goal") is simply to substitute the control units in the L^0 programme with operations that satisfy the G_i^0. Finally, there is the special case of "imprinting" (whereby most organisms acquire the basic, releaser dependent, control systems involved in sexual and maternal and gregarious behaviours); this sort of learning is imaged by D.22.

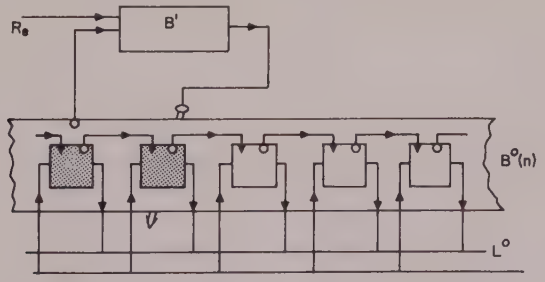


Fig. D.20

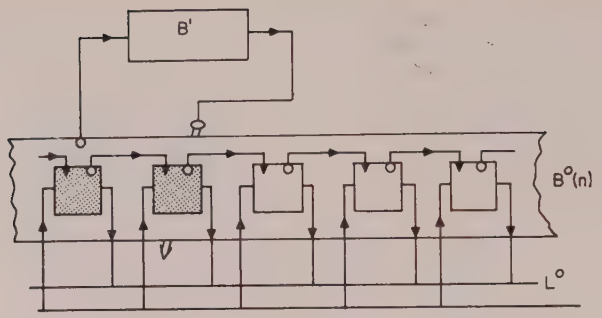


Fig. D.21

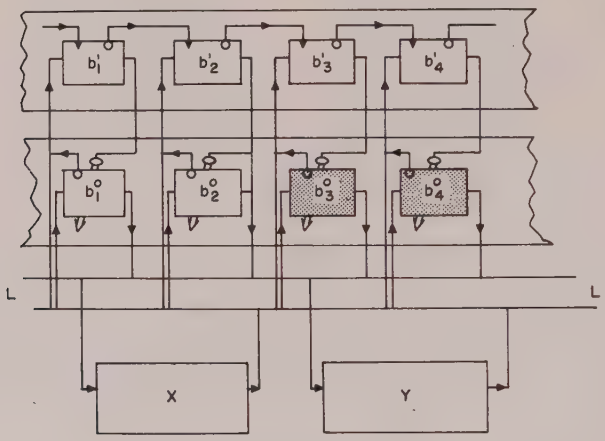


Fig. D.22

x, y, represent other members of a popolation of the same type.

We may characterise imprinting as a code substitution process that is dependent upon the context of an organism's early environment to which the listed conditions are applicable.

(1) The embodiment of all of the L^1 codes and some of the L^0 codes is innate. The L^0 embodiment, $B^0(0)$, consists in perceptual filters which, as in section 1.1. specify the coordinates of a universe of discourse and the organised motor activities such as "following" which are associated with sexually or maternally oriented behaviours. The embodied L^1 codes determine a set of L^0 programmes, each of which is represented in D.22 by a single "shaded" or imperfectly specified control unit subserving a sexual or maternal or gregarious goal G_i^0.

(2) The L^1 substitution operator, B_p^1, is an operation sequence with subgoals G_i^1 that are satisfied by the substitution (in the i-th-imperfectly specified control unit) of a control operation that satisfies the L^0 goal G_i^0. Hence, the $i + 1$-th-substitution cannot occur until the i-th-substitution has been accomplished.

(3) The embodiment of these substituents is physically stable (the L^0 operators are not prone to decay).

(4) The internal list of substituents, $\{a_j^0\}$, is either incomplete or nonexistent. Most or all of the substituents a_j^0 are provided as L^1 terms in $L = L^0, L^1$, which, in the sense of 1.1., is the natural object language of the species. In other words, the a_j^0 are ostensively defined by the significant appearance or significant behaviour of another member* of the species in relation to an activity, such as following, produced by the given member. The substitution establishes the releaser sign for the innate activity in respect to G_i^0.

(5) The L^1 goal, G^1, which is satisfied by the sequential achievement of the G_i^1 is a symbolic goal that is uninfluenced by external reinforcement signals. The instruction to achieve G^1 is an internal signal indicating a particular stage in the maturation of the organism.

This model accounts for most of the phenomena of imprinting, for example, most of the phenomena mentioned by Slukin.[48] It fails to account for the critical time at which a given type of imprinting can occur, though it does impose a critical and immutable order of imprinting (the order, in sequence, of the G_i^1 and thus of the substitutions subserving the G_i^0). This is not, perhaps, a very serious defect since the work of Guiton[49] suggests that the time in which a particular sort of imprinting can take place is widely variable (provided that the order of imprinting is preserved).

1.10. Cooperative Interaction and Development of Levels of Discourse

Let us introduce F.8 and suppose that a pair of organisms Z_1 and Z_2 are not only able to cooperate but are bound to cooperate in order to survive.

* There are some well known but pathological exceptions. Thus the appearance of an animal of a different species or even the appearance of the experimenter may satisfy the necessary relation and may be imprinted as a bizarre releaser. Similarly, behaviours may be simulated by changes in many perceptually filtered attributes. It is possible to imprint some birds so that they run after a flashing light.

In this case, Z_1 and Z_2 must also communicate in a fashion that allows for compromise between their possibly disparate goals. The least organisation is D.6 and the least really interesting organisation is a modified form of D.6 wherein L^0 as well as L^1 interaction takes place between Z_1 and Z_2. Given these premises, Z_1 and Z_2 must have an object language, L, that is stratified, at least to the extent that $L = L^0, L^1$ in D.23.

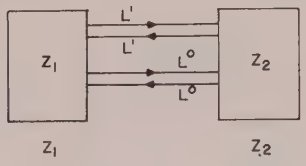


Fig. D.23

One exemplar or an essentially cooperative interaction that involves a stratified object language is the imprinting process of D.22 where any pair of organisms in a local population of the same species may act as Z_1 and Z_2 in D.23. Similar comments apply to many other types of learning including the case in which the demeanour of Z_1, an observing organism, (typically a pigeon) reinforces some behaviour of Z_2 (and possibly vice versa). Copying from an older or more experienced organism probably has the same character. So do various behaviours, most dramatically those involved in nesting, mating and sexual intercourse, which do not necessarily entail learning.[19, 41]

Z_1 and Z_2 may be forced to learn abstractions, in order to accommodate the flux of communication that is required to maintain their cooperative interaction. In these conditions, there will be higher levels of discourse because it becomes necessary to reach a compromise about higher and more abstract levels of goal.

Is it possible to avoid this proliferation of levels of discourse?

No, if the experimenter is anxious to maintain an unambiguous interpretation for object language expressions (when represented in his descriptive metalanguage). Hence, in laboratory experiments where such a condition is imposed (and where, in addition, the organism is forced to or inclined to learn and abstract) the object language must be stratified. We shall return to these stratified experimental systems later.

Yes, if the experimenter does not impose this condition or if he simply is not present. The fact is, higher organisms, and particularly man, com-

municate in an "unstratified" self referencing natural language. Their cooperative interactions constitute what Gorn calls "unstratified" control systems. The price paid for this arrangement is the appearance of Gorn's pragmatic ambiguity; an ambiguity of reference and interpretation which is experimentally intolerable but which Z_1 and Z_2 may readily resolve. Unstratified systems are perfectly efficient but unamenable to precise description.

For man, at any rate, an unstratified and self referencing language is important in development; it is the least environment in which the resolution of ambiguity can amount to creative activity and insight.*

2. THE CYBERNETIC MODEL IN RELATION TO EXPERIMENTAL METHODS

2.1. Experimental Situations

The adoption of any class of model imposes certain restrictions upon the experimental situations in which hypotheses derived from the model are tested. These restrictions may or may not be made explicit.

They usually are in specially designed studies, such as Bruner, Goodnow and Austin's[70] investigation of concept learning. Here, the experimental design (either in the selective or the receptive mode) is centered upon "externalising" the subject's strategy in a way that can be described within the framework of a cognitive structure (subsuming the universe of discourse of exemplars of classes) and can be explained by reference to limitations upon memory capacity. However, a study of this sort is exceptional, even in the field of concept learning where people are careful to examine underlying structures. For most experimental work, the restrictions imposed by a choice of model class are not made explicit; they are tactily assumed to exist by the body of observers and readers. There are lucid accounts, at a mathematical level, of the organisation of different sorts of experiment.[13] The experimenter is alive to the importance, for example, of the "experimental instructions". But there is often a

* The production of novel organisations is not the chief difficulty. This issue is dealt with, however, in (88), (44), (35), and by Apter[99] in his discussion of developmental systems.

good deal of disagreement about why they are important and how they are related to the basic model.

One virtue of a cybernetic model is that it commits the experimenter to a perfectly explicit experimental method; the subject is seen as part of a control loop completed in the experimental environment. The experimenter who designs this environment also designs a converse control system. In some cases the converse control system will help the subject to control his surroundings (acting in a cooperative fashion), in other cases it will aim to dominate his control strategy. But always, the experimental environment is active and, like any other control system, the system is designed on the basis of a model for the subject (including a special type of model, in which the subject is assumed to aim for the goal of controlling his surroundings).

At an organisational level, much the same comments should be applicable to many learning experiments; that is, much the same comments would be applicable if the tacit assumptions were made explicit. We shall illustrate the point by imaging a number of experimental learning situations in terms of the already developed conventions. With few exceptions, we shall not go beyond the issue of organisation.

The rates of responding, learning, or extinction measured in the majority of experimental situations are predicted within an organisational framework, but depend upon fabric assumptions like *F*. These assumptions are made in any fully developed cybernetic model. In the model of section 1.7. they are embedded in parameter assignments to the subroutines concerned with the reproduction and decay of physical operators; given the parameter values, it is possible to predict and test for learning or extinction rates, the effect of massed and spaced learning, and so on. The organisation of the experimental situation is a prerequisite for making these predictions but it does not, of course, lead to them. Hence, our account of experimental organisations is chiefly a discussion of the prerequisites for measurement.

2.2. Classical Experiments

Classical conditioning, that is, Pavlov's[51, 52] first type and Skinner's *S* type,[51, 8, 52] is represented by D.24. The left hand enclosure bounds the minimal organisation of a cybernetic model for the subject, namely D.16; the right hand enclosure bounds the organisation in the experi-

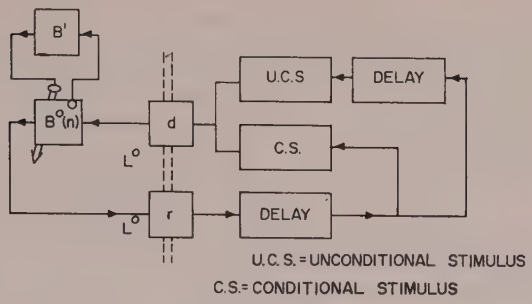


Fig. D.24

mental environment; letter "d" stands for a display and letter "r" stands for a response board or other facility. Interaction with the organism is confined to L^0 but, recalling the comments of Konorski, cited in 1.1., the experimenter must infer relationships betwenn the stimuli and B^1 as well as between the stimuli and $B^0(n)$.

2.3. Instrumental Procedures

Instrumental conditioning of the sort considered by Thorndike[51] and Hull[51] (as the instigators of a vast experimental project) involves the provision of a reinforcing L^1 signal through the auxiliary display "Re" in D.25. Of course, "Re" may also represent a mechanism for delivering food or some other drive reducing substance to an animal.

Apart from this, the main difference between D.24 and D.25 is that, in the latter case, the subject model D.16 is necessarily replaced by the subject model D.17. The box "Se" is an arbitrary stimulus sequence

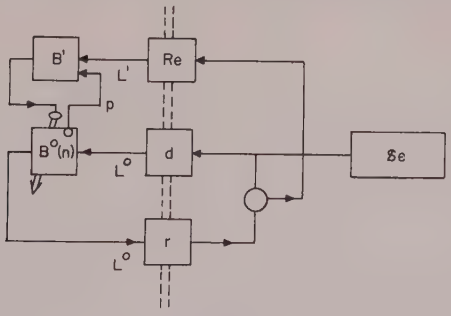


Fig. D.25

predetermined by the experimenter. "Se" may be degenerate (as when the animal subject is replaced in the same puzzle box or at the entrance to the same maze).

It is important to observe that the experimenter must ensure parity between the internal L^1 signal, "p", and the external reinforcing signal delivered through the box "Re". When the subject is a man, parity is achieved by care over the experimental instructions. When the subject is an animal, special inferences are needed to equate the reinforcing event and the reduction of an internal drive. Finally, we comment that D.25 is completely symmetrical. The subject model is based upon the paradigm D.10 whereas the experimental environment is derived from the equivalent image D.11 in which the comparator is explicitly stated. This is no more than a matter of convenience (the representation is reasonable because the comparator is a tangible bit of equipment). Hence, each side of the picture is equivalent apart from the stimulus sequence. But "Se" is balanced by an internal variation due to the fact that the subject is an active control system. Hence, each side is equivalent.

2.4. Operant Experiments

If the active L^0 control system in D.26 has no necessary input, then it is an operant (or a collection of operants) and this figure represents Skinner's type *R* or operant conditioning. If the reinforcement was always primary in calibre we might, in certain very simple cases, attach the output from "Re" directly to the parametric input of $B^0(n)$ and delete B^1. The paradigm for conditioning a discriminated operant is D.27. If we adopt the convention that "Re" is not only a display device but contains equip-

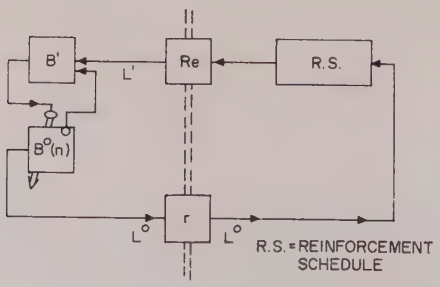


Fig. D.26

ment for elaborating the reinforcement schedule, then this paradigm becomes equivalent to D.25. The output from box *d* is the discriminating stimulus which, in Skinner's words, is an "occassion" for the response, rather than "that which produces the response".

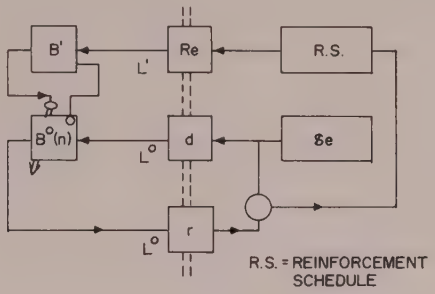


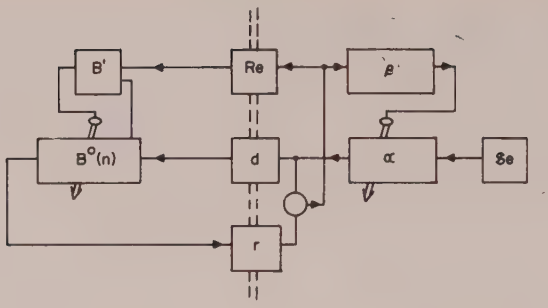
Fig. D.27

This is probably the least elaborate experimental situation in which we come up against Gorn's pragmatic ambiguity. Any discriminating stimulus is a term in L^0. However, such a stimulus sign becomes a secondary reinforcing sign, which is a term in L^1. Hence, there is likely to be an ambiguity regarding the level of discourse to which a sign refers, since secondary reinforcing signs are also stimulus signs.

2.5. Skill Learning

If an organism is taught to produce a chain of responses by a cueing procedure, the experimental situation is described by D.28, wherein the subject model is simply D.17. However, in order to attempt this procedure in the first place, the experimenter is bound to make a number of assumptions about the nature of a response "chain" and the nature of the organisation in the subject that responds to the "cues" and learns.

In fact, he needs much the same set of assumptions that we adopted in 1.7. and in 1.8. In particular, he needs to assume that $B^0(0)$ and B^1 exist, that "learning" is a "substitution" process in which the development of a chain is analogous to the development of an embodied operation sequence and that cueing is a simplification process. Given all this (and most of it must be given in order to make explicit all the constraints entailed by talking about chain learning) the paradigm D.29 is more in-



α =Adjustment of level of simplification or partial solution of problems posed by Se.
β =Strategy for adjusting simplification as a function of measure of average proficiency.

Fig. D.28

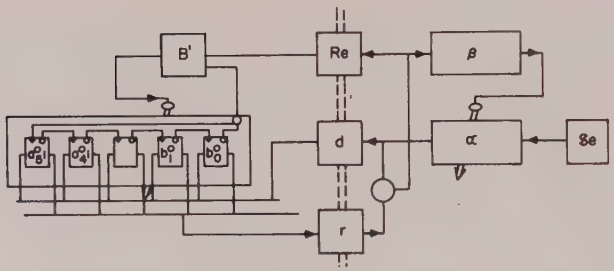


Fig. D.29

formative than D.28 (the subject model D.17 is replaced by D.18 or D.19 section 1.8).

The function of "simplification"* is to transform a problem u_i into a simplified or partially solved problem u_{ij}, simplified to the degree j, but derived from u_i. In other words, according to the degree of simplification, L^0 problems are delivered to different inputs in $B^0(n)$; at b_0^0, b_1^0, and so on, (these problems can be solved because embodied operation processes exist); or at code locations like a_5^0 or a_4^0 that have not been substituted at the n-th-trial, (in which case the problem cannot be solved).

* We assume a unidimensional simplification procedure. In fact, the number of dimensions is the same as the number of error factors, in the sense of Harlow,[98] entailed by the skill as specified and the degree of simplification is a vector with this number of entries.

This situation corresponds, in my own work, to the instruction of an homogeneous subskill of a structured skill.[18, 54] Here, the simplification strategy is also chosen to satisfy the loading requirements derived from 1.7. There is a class of simplification strategies which adjust the difficulty of the sequentially presented problems so that the subject is neither underloaded nor overloaded, but maximally loaded. Intuitively, these strategies are intended to maintain the subject's attention concentrated upon the relevant problem sequence (which is much the same as Lewin's[82] concept of maintaining the level of difficulty within an individual's capacity; only, in this case, the pertinent capacity is modified as the subject learns). For man and for particular skills it has been shown that strategies* of this class maximise the rate of learning[54, 56] (as we might predict from the model in 1.7. and in 1.8.).

2.6. Higher Levels

There are a number of experiments in which an animal such as a pigeon or a rat is required to make an L^1 response with the calibre of a judgement or choice. Thus, Blough[77] uses an ingenious technique to allow pigeons

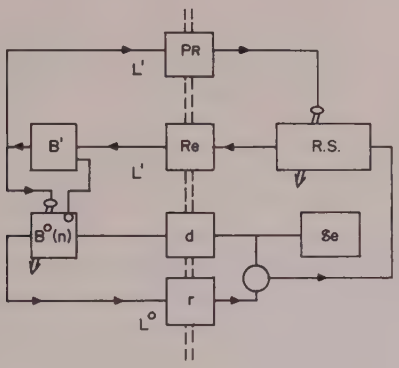


Fig. D.30

to make psychophysical judgements as a result of which it is possible to determine their visual threshold and to plot adaptation curves. Verhave[78] has trained these animals to signal their choice of one reinforcement

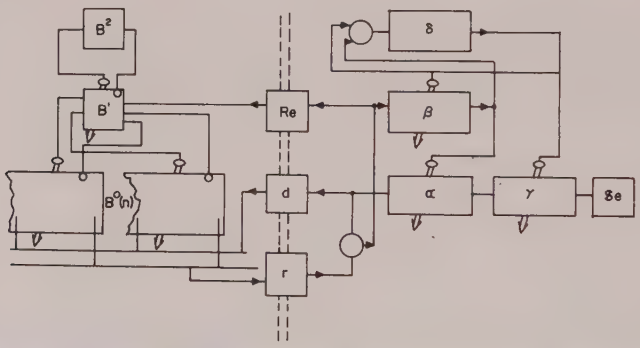
* In principle, the experimenter might perform these adjustments, but in practice, the job is delegated to an adaptive machine.

schedule rather than another, and, by a compensatory method, he is able to determine, amongst other things, a degree of relative preference. The paradigm for Verhave's experiment is D.30.

The immediately important point about D.30 is that it represents one of the simplest experimental situations that is characterised by a pair of control loops, one in L^0 and one in L^1.

2.7. Concept Learning

Consider the sort of concept learning that goes on when a man is required to combine a pair or more of subskills (either by alternate rehearsal or some other method). In these conditions, the acquisition of one subskill typically interferes with the acquisition of the other until the subject produces an L^1 organisation (a concept in the sense of 1.7. and 1.8.) that allows him to deal with the subskills as a whole and to wed them into a common pattern. The basic experimental situation closely resembles D.31



γ = Selection of subskill or, equivalently, of type of problem. The selection made by Se is thus restricted to problems of the given type.
δ = Strategy for alternating subskills or the rehearsal of different types of problems, simplification being determined separately for each type.

Fig. D.31

which depicts a system used in my own laboratory for investigations of concept learning. Here, the alternation strategy that selects between a pair of subsystems carrying out separate simplification strategies, is designed to maximise the product of the difficulty levels* and thus to equalise these levels.[55, 51, 58] But a less specific form of D.31 is involved in studies

* "Difficulty" is used as the converse of "simplification." The level of difficulty is the numerical converse of the degree of simplification.

of the "transfer of training", at any rate in Osgood's[83] sense of that phrase, and in studies of proactive and retroactive inhibition. Essentially the same sort of organisation must be embodied in experiments that deal with the dynamic features of "concept learning", even if it has nothing to do with skilled activity. But the basic organisation is often obscured, of necessity, by relatively uncontrollable modes of data presentation.

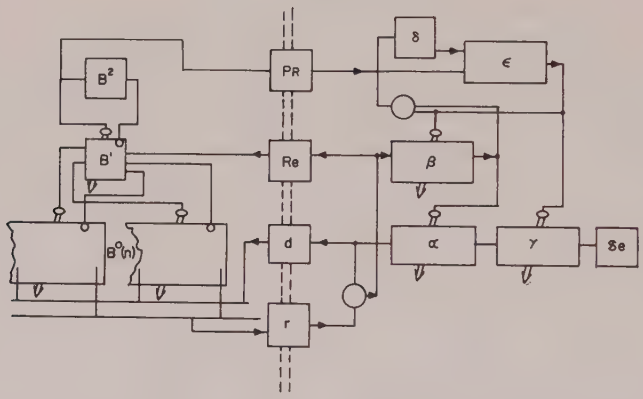
2.8. Participant Interaction

If the discourse is confined to L^0, it soon becomes impossible to realise the strategies intended to secure the conditions required by 1.7. and 1.8. The coupled man machine system, which is initially stable in the fairly strict sense of Lyapunov stable, exhibits signs of instability when the subject becomes proficient. The instability stems from a peculiar misuse of L^0 (terms in which ought to denote problems and attempts at their solution). We call this activity "participant interaction". The subject makes deliberately false response in such a way as to modify the characteristics of the machine that is controlling the experiment (he aims to discuss and modify the mode of instruction and L^0 contains no legal constructions for this purpose). Introspecting, after the experiment, the subject often agrees that he was trying to play with the system like this.

Of course, the experimenter can readily prohibit the byplay of "participant interaction" by interrupting the correlative information that the subject needs in order to determine the rules in his mechanised experimental environment. But if he does so, the man machine system becomes instable; its continued stability *depends* upon participant interaction. We might go further and suggest that participant interaction is a symptom of what Bartlett[84] calls the "search for meaning"; men, and a few other creatures like the dolphin,[64] are not only "curious" in the sense of D.10 (rather than merely "restless" in the sense of D.4), they also seek to control their surroundings and indulge in cooperative interaction with other organisms. As we argued in 1.10. this property would be sufficient to account for the "search for meaning" and the tendency to construct ever increasing levels of discourse.

Hence, participant interaction is something the experimenter may be anxious to *investigate* rather than inhibit. It only offends him because, as a result of it, well defined L^0 expressions becomes ambiguous. To avoid

the pragmatic ambiguity without preventing the participant interaction, it is possible to use an "adaptive metasystem",[9, 12] one form of which is D.32. The subject is provided with an L^1 facility (in addition to the L^0 facility) whereby he receives evaluative statements about his performance (in the case of learning a skill, or relevant property statements in other cases). In L^1 he is also able to issue preference statements, "Pr", (for example, that he prefers to rehearse one subskill rather than another). The weight associated with the subject's preference statements, "Pr",



ε = Device for computing the value of $\theta(n)$ and for selection the subject's choice or the machine's choice as that which determines the rehearsal to be undertaken at the n-th-trial.

Fig. D.32

and thus the amount of L^1 control that he is able to exert, is $\theta(n)$ at the n-th-trial. The machine, which previously controlled the situation, also has a "preference" statement at the n-th-trial. The machine "preference" is given a weight of $1 - \theta(n)$. Finally, $\theta(n)$ is defined as a variable in the interval between 1 and 0 which indicates the extent to which the subject has satisfied the conjoint L^0 goal G^0.

Briefly, the subject is allowed a degree of control in pursuit of G^1 that depends upon his average satisfaction of G^0. In a teaching system, for example, the novice has no L^1 control, for $\theta(0) = 0$; a moderately proficient subject has some L^1 control; a proficient subject has complete L^1 control for at $n = T$, when the terminal criterion is achieved, $\theta(n) = \theta(T) = 1$.

The construction may be iterated to yield an arbitrary level of strati-

fication. It is not restricted to teaching systems. Rather similar adaptive metasystems allow the subject a θ dependent control over a "miniature economy" wherein he can purchase test facilities, external memory facilities or computation; hence, the arrangement can be used for all manner of investigations of concept learning. The real restriction is that the language of the discourse is, and must be, stratified. Insightful learning, if it takes place, appears as a mysterious discontinuity.

As we argued in section 1.10, the insightful facet of man is bound up with the cooperative actions mediated by unstratified control. It seems likely that experiments on insightful learning necessarily entail the interpretation of unstratified discourse and involve genuine conversations. The present arrangement, though conversation like, is insufficient.

REFERENCES

1. Pask, G. "The use of analogy and parable in Cybernetics, with emphasis upon analogues for learning and creativity" Dialectica. **17.** 1963.
2. Pask, G. *An Approach to Cybernetics*, Hutchinson 1961.
3. Cherry, C. *Human Communication*, M.I.T. Press and Wiley 1957.
4. Lorenz, K. "The comparative method in studying innate behaviour patterns" Symp. Socy. Exp. Biol. **4.** 1950.
5. Konorski, J. "The Role of Central Factors in Differentiation" in *Information Processing in the Nervous System*, I.U.S.P. Leiden, 1962.
6. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. R. "What the frog's eye tells the frog's brain" in W. S. McCulloch, *Embodiments of Mind*, M.I.T. Press 1965.
7. Anohkin, P. K. *The role of the orienting reaction in conditioning*. Acad. Pedagogical Sciences. Moscow. 1958.
8. Sokolov, E. N. "Perception and the conditioned reflex" Pergamon.1963.
9. Pask, G. "Comments on the organisation of men, machines and concepts in *Education for Information Science*. Edited by Heilprin, Markusson and Goodman. Spartan Press and MacMillan. 1965.
10. Pask, G. "Teaching as a control engineering process" Control. January, February, March, April 1965.
11. Andrea, R. C. "Stella. A scheme for a learning machine" Proc. 2nd IFAC Conference, Basle, 1963, Butterworths 1965.
12. Deutsch, A. *The structural basis for behaviour*. Cambridge. 1960.
13. Feigenbaum, E. Paper at Symposium on "Information processing and memory" at American Psychological Association, Los Angeles, Sept. 1964.
14. Klix, F., Neumann, J., Seeber, H., and Sydow, H. "Die algorithmische Beschreibung des Lösungprinzips einer Denkanforderung" *Z. für Psychologie*" **168,** 1 and 2 (1963).

15. Mittelstaedt, H. "Control systems of orientation in insects" *Annual Rev. Ent.* **7** (1962).
16. Napalkov, A. "A study of the laws of development of complex reflex systems" Vest. Nik. 2. Moscow University 1958.
17. Young, J. Z. *A Model for the Brain*, Oxford, 1965.
18. Pask, G. "Man as a system that needs to learn" in Automata and Learning Theory Edited by Beer, George and Stewart, Academic Press. 1967
19. Gorn, S. "The treatment of ambiguity and paradox in mechanical languages" Proc. Symp. in Pure Mathematics. Vol. 5. American Mathematical Society 1962.
20. Ashby, W. R. *Design for a Brain.* Chapman and Hall 1952.
21. Miller, G. A., Gallanter J. E., and Pribram, K. *Plans and the Structure of Behaviour.* Holt Dryden. 1960.
22. Ashby, W. R. *An Introduction to Cybernetics*, Chapman and Hall, 1957.
23. Mesarovic, M. D. "Self Organising Control Systems" Applications and Industry, I.E.E.E. Sept. 1964.
24. Tarjan, R. "Problems of stability in adaptive control systems" in *Optimising and Adaptive Control*, Edited by Bollinger, Truxall and Minar. Instrument Society of America. IFAC Rome Symposium 1961.
25. Miller, G. A., and Chomsky, N. "Finitary Models of Language Users" in *Handbook of Mathematical Psychology* Vol. 2. McGraw Hill. 1964.
26. Tinbergen, N. *Social Behaviour in Animals*, Methuen. 1953.
27. McCulloch, W. S., and Pitts, W. R. "How we know universals. The perception of auditory and visual forms." In *Embodiments of Mind*, M.I.T. Press 1965.
28. Wiener, N. *Cybernetics*, 2nd Edition. Wiley 1961.
29. Sokolov, E. N. "Neuronal models and the orienting reflex" in *The Central Nervous System and Behaviour*, Edited by M. A. Brazier. J. Macy Foundation 1960.
30. Jouvet, M. "Recherches sur les mechanismes neurophysiologiques du sommeil et de l'apprentissage negatif" in *Brain Mechanisms and Learning*, Edited by Albe Fessard, Oxford. 1961.
31. Lynn, R. *Attention, Arousal, and the Orientation Reaction.* Pergamon, 1966.
32. Thorpe, W. H. *Learning and Instinct in Animals*, Methuen. 1956.
33. Miller, G. A. "The magic number seven plus or minus one," Psychological Review. **63**. 1956.
34. Newall, A. "Learning, generality, and problem solving" in *Information Processing*, Proc. of IFIP Congress. Edited by C. E. Popplewell, North Holland, 1962.
35. Pask, G. "A discussion of artificial intelligence and self organisation" in *Advances in Computors*, Vol. 5. Edited by M. Rubinoff. Academic Press, 1964.
36. Von Foerster, H., and Pask, G. "A Predictive Model for Self-Organising Systems," *Cybernetica* **1**. 1960 and **1**. 1961.
37. Von Foerster, H. "Biologic" in *Biological Prototypes and Synthetic Systems*, Edited by E. E. Bernard and M. R. Kare, Plenum Press, 1962.
38. Morris, D. *The Biology of Art*, Methuen, 1962.
39. Berlyne, C. *Conflict Arousal and Curiosity*, McGraw Hill, 1960.
40. Feigenbaum, E., Feldman, J. Editors. *Computers and Thought*, McGraw Hill, 1964.

41. Lurman, D. S. "The reproductive behaviour of ring doves," Scientific American, November 1964.
42. Wynne Edwards, W. C. *Animal Dispersion* Oliver and Boyd, 1964.
43. Rappaport, R. A. "Ritual Regulation of the Environmental Relations of the Tsombega" Columbia University. In Press.
44. Pask, G. "The Cybernetics of Ethical, Sociological and Psychological Systems" in *Advances in Biocybernetics* Vol. 4. Elsevier Press 1966.
45. Pask, G. "Simulation of Learning" in *Aspects of Artificial Intelligence* Edited by C. A. Muses. Plenum Press. 1963.
46. Pask, G. "Machines a Enseigner," Cegos, Paris, 1962.
47. Feigenbaum, E., and Simon, H. "Elementary perceiving and memorising machine" in *Information Processing*. Proc. 2nd IFIP Symposium. Edited by C. E. Popplewell, North Holland, 1962.
48. Sluckin, W. *Imprinting*. Methuen, 1965.
49. Guiton. P. "Socialisation and imprinting in brown leghorn chicks," *Animal Behaviour*, **7** (1959).
50. Pask, G. "Tests for some features of a Cybernetic Model," *Z. für Psychologie*. **171**. Kybernetic Sonderband, 1965.
51. Hilgard, E. R. *Theories of Learning*. Methuen, 1958.
52. Woodworth, E. S. and Schlosberg, H. *Experimental Psychology*. Methuen, 1954.
53. Sutherland, N. S. "The Learning of Discrimination by Animals," *Endeavour* **XXIII**, No. 90, Sept. 1964.
54. Pask, G. "Man Machine Interaction in adaptively controlled conditions," *Bull. Math. Biophysics* **27** (1965).
55. Pask, G. "Some results from adaptively controlled experiments on learning and teaching" Bionics Symposium 1966. Collection of "Short" papers. Wright-Patterson Air Force Base, Ohio 1966.
56. Pask, G. "The teaching machine as a control mechanism," Trans. Soc. Inst. Technology, June 1960.
57. Pask, G. "Electronic Keyboard Teaching Machines," Inl. Natn. Association for Education. July 1958. Reprinted in *Teaching Machines and Programmed Learning*. Vol. 1. Edited by Lumsdaine and Glasser. Nat. Educ. Association 1960.
58. Lewis, B. N., and Pask, G. "The theory and practice of adaptive teaching systems" in *Teaching Machines and Programmed Learning*, Vol. 2. Edited by R. Glasser, Nat. Educ. Association, 1965.
59. Maturana, H. R. "The functional organisation of the pigeon retina" in *Information processing in the nervous system*, I.U.S.P. Leiden, 1962.
60. Rescher, N. *The logic of commands*. Routledge and Kegan Paul. 1966.
61. Gorn, S. "The individual and political life of information systems" in *Education for Information Science*. Edited by Heilprin, Markusson and Goodman, Spartan Press and MacMillan. 1965.
62. Ashby, W. R. "The Mechanism of Habituation" in N.P.L. *Symposium* **10**. *The Mechanisation of Thought Processes*, Her Maj. Stat. Office, London, 1959.
63. Beer, S. *Cybernetics and Management*. English Universities Press 1959.
64. Lilly, J. C. Lecture at the present meeting. "Research with the Bottlenose Dolphin."

65. Miller, J. G. "Information Input Overload" in *Self Organising Systems*. Edited by Yovits, Jacobi and Cameron, Spartan Press, 1962.
66. Augenstein, L. "Information processing and control in humans," AGARD Lecture Series Paris, 1965.
67. Sutherland, N. S. "Shape and Size Discrimination in Octopus," *J. Genetic Psychology* **106** (1965).
68. Hesse, M. *Models and Analogies in Science*, Sheed and Ward, 1963.
69. Hyden, H. Contribution to *The Cell*, Edited by Brachet and Mirsky, Academic Press, 1960.
70. Bruner, J., Goodenow, J., and Austin, G. A. *A Study of Thinking*, Wiley, 1956.
71. Harre, R. *Theories and Things*, Sheed and Ward, 1962.
72. Russel, B., and Whitehead, A. N. *Principia Mathematica*, Cambridge University Press, 1927.
73. Suppes, P., and Zinnes, R. "Basic Measurement Theory" in Handbook of Mathematical Psychology. Edited by Luce, Bush and Gallanter, Wiley, 1963.
74. Minsky, M. "Some methods of heuristic programming and artificial intelligence" *The Mechanisation of Thought Processes*, Her Maj. Stat. Office, London, 1959.
75. Pribram, K. "Reinforcement Revisited. A Structural View." Nebraska Symposium on Motivation. University of Nebraska Press 1963.
76. Pask, G., Lewis, B. N., Mallen, G., Feldmann, R., and McKinnon Wood, T. R. Technical Reports, Contract, AF 61.052.640.1963, 1964, 1965 and 1966.
77. Blough, D. S. "Experiments in Animal Psychophysics." *Scientific American* **205** (July 1961) No. 1.
78. Verhave, T. "Towards an Empirical Calculus of Reinforcement Value," *J. Expt. analysis of behaviour* **6**, No. 4.
79. Etkin, J. *Social Behavior and Organisation amongst Vertebrates*. Chicago, 1965.
80. Dorf, R. C. *Time Domain Analysis and Design of Control Systems*. Addison Wesley, 1966.
81. Skinner, B. F. *The Behaviour of Organisms. An Experimental Analysis*. Appleton, Century, Crofts, 1936.
82. Lewin, K. *A Dynamic Theory of Personality*. McGraw Hill, 1935.
83. Osgood, C. E. "The similarity paradox in human learning. A Resolution," *Psych. Review* **56** (1949).
84. Bartlett, F. C. *Thinking*. Cambridge University Press, 1932.
85. Kilmer, W. L., McCulloch, W. S. *et al.* "On a Cybernetic Theory of the Reticular Formation," Bionics Symposium 1966. This volume pp.431–479.
86. Hull, C. L. *A Behaviour System*, Yale University Press, 1952.
87. Pask, G. "The Cybernetics of evolutionary processes and selforganising systems." Proc. 3rd. Congress. Int. Assn. for Cybernetics. Namur. 1961.
88. Pask, G. "Physical analogues for the growth of a concept" in N.P.L. *The Mechanisation of Thought Processes*, Her Maj. Stat. Office, London, 1959.
89. Tolman, E. C. "A Psychological Model" in *Toward a General Theory of Action*. Edited by Parsons and Shils. Harvard University Press, 1951.
90. Hovland, C. I. "A Communication Analysis for Concept Learning," *Psychological Review* **59** (1961).

91. Hunt, E. C. *Concept Learning*, Wiley, 1962.
92. Selfridge, O. "Pandemonium. A paradigm for learning" in *The Mechanisation of Thought Processes*, Her Maj. Stat. Office, London, 1959.
93. Newell, A. "Intelligent Learning in a General Problem Solver" in *Self Organising Systems, Edited by Yovitts and Cameron*, Pergamon Press, 1960.
94. Flavell, J. H. *The Developmental Psychology of Jean Piaget*, Van Nostrand, 1963.
95. Waddington, C. H. *The Strategy of the Genes*, George Allen and Unwin, 1957.
96. Vygotsky, L. *Thought and Language*, M.I.T. Press, 1964.
97. Bateson, G. "Cultural Problems posed by a Study of the Schizophrenic Process" in *Symposium on Schizophrenia*, Edited by Auerbach, Ronald Press Company, 1959.
98. Harlow, H. F. "Learning Set and Error Factors theory" in *Psychology. A Study of a Science*, Vol. 1. Edited by Koch, McGraw Hill, 1959.
99. Apter, M. *Cybernetics and Development*, Pergamon Press, 1966.
100. Banerji, R. C. "Computer Programmes for the Generation of New Concepts from Old Ones" in *Neuere Ergebnisse der Kybernetik*, Edited by Steinbuch and Wagner, Oldenburg, 1963.

V. V. GRIFFITH

J. A. DAVIS

R. H. KAUSE

Goodyear Aerospace Corporation,

Akron, Ohio

Learning of the Exclusive-Or Logic Function in Rats

ABSTRACT

A search by the authors of the literature on conditioning of lower creatures yielded a startling fact. In every conditioning experiment which was examined by the authors, the decision criterion could be expressed as a linearly separable logic function.

The present paper presents the results of a reinforcement conditioning experiment in which the decision criterion is the "exclusive-or" function—the simplest logic function which is not linearly separable.

The learning curves obtained are presented. It is shown that they resemble those which might be obtained from a particular two-level network of threshold elements.

Data on the extinction of the conditioned responses are also presented.

INTRODUCTION

Since Pitts and McCulloch[1] established that some aspects of neural behavior are described by boolean algebra, many researchers have attempted to establish some isomorphism between creature learning and changes in weights or thresholds in networks of threshold logic elements. Reference 2 describes some attempts, and lists a bibliography.

Hopefully, the learning curves of living creatures in making decisions might provide clues to the organization of the networks of neurons that actually perform such tasks. A fairly extensive search of the literature by one of the authors to find such learning curves yielded that in all reported experiments which were examined the animals were required only to make decisions which were linearly separable logic functions.

Although quite "complicated" decisions are made by experimental animals, "complexity" has been obtained by requiring that the animal's decision at any instant be mediated by events of the recent past, e.g., double alternation experiments.[3] Decision problems involving the past history of inputs are studied in automata theory, and can be instrumented by a combinational logic network plus a set of unit time delay elements which constitute the "memory" of the system for recent past events.

An experiment was conducted in which laboratory rats were to make a decision which was not linearly separable. The logic function chosen was the "exclusive-or" function of two variables, the simplest such logic function.

METHOD

1. Subjects

The *S*s, twelve albino rats purchased from Sprague-Dawley, Inc., Madison, Wisconsin, ranged from 3–5 months in age at the start of experimentation.

2. Apparatus

A single choice-point *Y*-maze (see Fig. 1) was used; all sections of the maze were five inches wide and five inches high (inside measurements). A gray starting box, nine inches long, was located at the base of the stem of the *Y* which was also gray. A guillotine door separated it from the stem. The distance from the door to the choice point was 20 inches. Guillotine doors were placed 1.0 inch in from the choice point on each side. The *Y* portion of the maze contained goal boxes, separated from the corridors by guillotine doors. Each goal box was 15.5 inches long and had at its terminus a glass coaster dish, 1.75 inches in diameter, mounted on a plywood platform 4 inches long and 2.5 inches wide. The right hand goal box was white; the left hand goal box was black. The maze overheading was made of $^1/_4$ inch wire mesh.

Two small flashlight bulbs (3V) were mounted vertically at the choice point. The frame and sockets for these lights were painted the same gray as the stem of the *Y*-maze. Each bulb was operated individually by a toggle switch attached to the side of the start box. Flashlight batteries for the lights were outside one delay box.

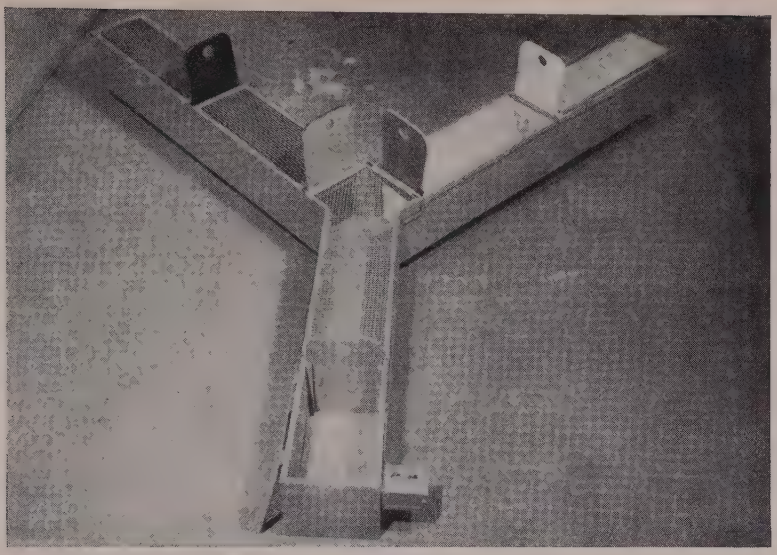


Fig. 1

3. Procedure

*S*s were prepared for the experiment by 48 hours food deprivation followed by a schedule of limited feedings every 24 hours for ten days. This reduced each *S* to approximately 80% of *ad libitum* body weight. They were kept at this weight throughout the experiment. The food given the *S*s following each day's training was Purina Lab chow. Quantities depended upon the size or weight of the animal. After 20% weight reduction the *S*s ranged from 256–326 gms. The weight of each *S* was relatively constant; $\pm$1–3 gms from the calculated 20 per cent weight reduction.

*S*s were allowed to explore the maze for 15 minutes on each of two successive days before training to reduce the novelty of the maze situation and to provide the maze with a familiar smell. The procedure on the first day was to remove all guillotine doors and place *S* in the start box from which it could wander about the maze. Since most *S*s preferred the black arm of the maze, on the second day *S*s were placed in the white goal box. No signal lights or food were presented to *S*s at this time.

After familiarization the *S*s were divided into two squads of six animals each. *Ss* received fifteen (15) training trials per day for twenty (20) days.

Table I illustrates the relationships of food reward location and signal lights, which *S* was required to learn. A counterbalancing procedure controlled for any inherent side preference.

Table I. Lights as Related to Location of Food Reward

Squad 1	Squad 2
Food reward—left goal box	Food reward—right goal box
Bottom light only	Bottom light only
Top light only	Top light only
Food reward—right goal box	Food reward—left goal box
Both lights	Both lights
No light	No light

Each reward was four round-shaped food pellets purchased commercially from the P. J. Noyes Company, Lancaster, New Hampshire.

During each trial, *S* was kept in the starting box for 15 seconds, and in the goal box until food was consumed. The *S* remained in the goal box for at least 30 seconds regardless of choice. Only on the first day was additional time required for *S* to consume the food. Upon completion of the trial, *S* was removed from the maze and replaced in its cage until its next trial. Water was available to *S* between trials. Approximately 25–30 minutes separated the trials for each *S*.

After a twenty day training period, the 12 *S*s were randomly divided into two groups. Both groups were run an additional eight days under the same training conditions, except that for group one only 50 per cent of the correct responses were reinforced while for the other group, reinforcement was completely omitted (an extinction procedure). After the eight day interval, group one was dropped from 50 per cent reinforcement to 0 per cent and run an additional eight days.

RESULTS

Figure 2 presents the per cent of "correct" (reinforced) responses for all subjects on each successive day. The curve shows a steady increase to about 90 per cent after an unsteady start.

As shown in Figure 2 the discrimination performance of both groups of animals fell off on the first day following the change in reinforcing conditions. Performance of those animals run on the extinction procedures continued to decrease to a chance level where it stabilized. On the

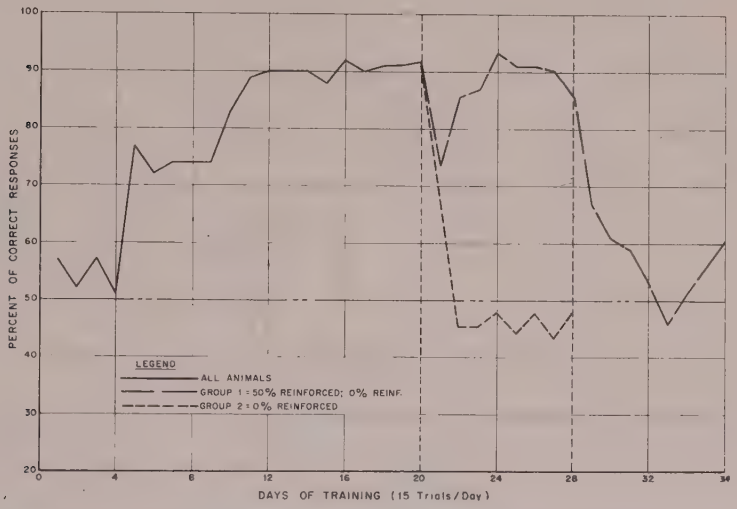


Fig. 2. The per cent of correct responses for all *S*s on all trials for each day of training and extinction.

other hand, performance of those animals on 50 per cent reinforcement rose again to the same asymptotic level at day five. When reinforcement was eliminated for these animals after the eighth day, discrimination performance again dropped sharply. The rate of descent of the curve after day one of extinction was less for those animals who had previously undergone partial reinforcement. After dropping to a chance level on day five, the curve unaccountably rose again. A comparison of the performance of the two groups on the last six days of the extinction trials was made using a Mann-Whitney "*U*" statistic. A difference between the two groups significant beyond the .01 level of confidence was obtained.

Figure 3 presents separately the per cent of correct responses to each discriminable light condition: no light on, both lights on, top light on with bottom light off, and top light off with bottom light on. As can be seen Ss rapidly learned the correct response to a single light and only

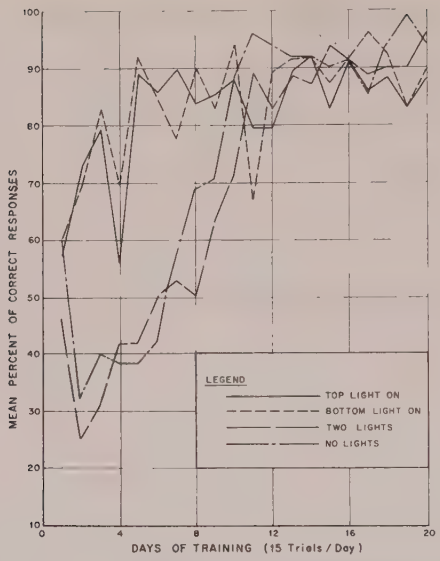


Fig. 3. The per cent of correct responses to each discriminable light condition for each day of training.

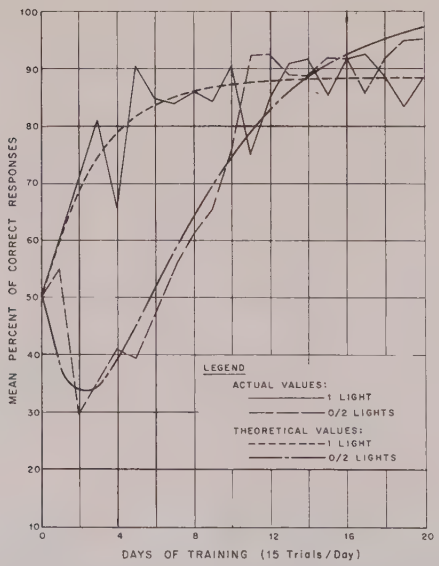


Fig. 4. The per cent of correct responses for the 1 light and the 0/2 light conditions with a theoretical curve fit for each condition for all days of training.

gradually learned the correct response to the zero light/two light condition. Figure 4 shows an averaging of responses to the one light condition and to the zero light/two light condition, plus theoretical curve-fits discussed in the next section. Performance on the zero light/two light condition rapidly deteriorates from a chance level during the first few days indicating that some interference is taking place, perhaps as a result of transfer from the discrimination learning of the one-light condition. This interference contributed to the unsteadiness of the learning curve in Figure 2. The sharp upswing in the total curve occurred when learning began to take place to the zero light/two light problem after about six days. The hypothesis that learning on the one light problem interfered with learning on the zero light/two light problem was tested in the following manner:

ANALYSIS

The learning curves of Figure 4 were fitted by use of a standard nonlinear multiple regression computer program. It was postulated that the learning of the correct response to one illuminated lamp (or, conversely, to one lamp *not* being illuminated) was a simple exponential learning curve of the form

$$\dot{p} = k_1(\hat{p} - p); \quad p(0) = 0.5$$

where $\hat{p}$ is the asymptotic value, p is the probability of a correct response, and k_1 is the inverse time constant. Values obtained were $p = 0.88$, $k_1 = 0.35$/day.

A fairly good fit for the 0, 2-light curve was obtained by use of the equation

$$\dot{q} = -fk_1(\hat{p} - p) + k_2(\hat{q} - q), \quad q(0) = 0.5$$

where f is a constant, k_2 is another inverse time constant, and $\hat{q}$ is the asymptote. Values obtained were $q = 1.009$, $f = 2.18$, $k_2 = .205$.

As pointed out by Hald[4], the standard tests for "goodness-of-fit" are not properly applicable to growth curves, since any particular point on such a curve is dependent on previous points. The *residuals* are not independent. An "eyeball test" of Figure 4 indicates, however, that the theoretical curves fit the experimental data reasonably well.

DISCUSSION

A logic function which is not linearly separable can be generated with threshold elements in at least two ways, as shown on Figure 5. It is readily seen that any logic function can be generated by a network of at most two layers. One way is for each minterm of the disjunctive normal form

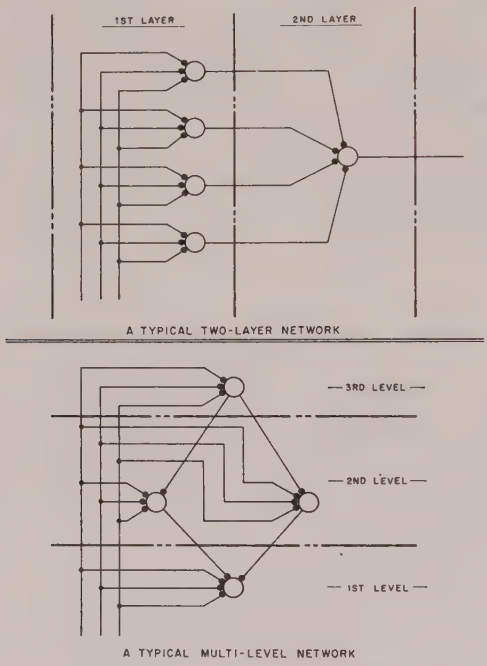


Fig. 5. Two Methods of Constructing a Logic Function which is not Linearly Separable.

of the function to be generated by one of the elements in the first layer. Each of these first layer elements is then sufficiently excitatory on the second layer element to cause it to produce an output whenever at least one first layer element has its threshold exceeded.

Another method (Fig. 5b) requires a number of levels determined by the exact form of the logic function. Intuitively, the second form seems more consistent with the interpretations of creature behavior as resulting from activity of higher centers mediating the activities of lower centers. If both higher and lower centers are learning simultaneously, and are learning opposing responses to the same clues (in different combinations)

from the environment, the lower center learning could possibly tend to negate the effects of higher center changes. Such an hypothesis does seem to fit the observed facts, but too little data is available to draw any firm conclusions.

The experiment does show an unusual effect that should be further investigated.

REFERENCES

1. McCulloch, W. S., and Pitts, W. "A Logical Calculus of the Ideas Immanent in Nervous Activity," Bull. math. Biophys., **5**, 155–133, 1943.
2. Hawkins, J. K., "Self-Organizing Systems—A Review and Commentary," *Proceedings of the IRE* **49**, No. 1, 31–48, January 1961.
3. Munn, N. L., *Handbook of Psychological Research on the Rat*, Houghton Mifflin Company, Boston, 1950, pp. 207–212.
4. Hald, A., *Statistical Theory with Engineering Applications*, J. Wiley and Sons, New York, 1952, pp. 667.

HANS BREMERMANN

University of California

Berkeley, California

*Numerical Optimization Procedures Derived from Biological Evolution Processes**

GENES, FITNESS, OPTIMIZATION

Biological evolution has for at least two billion years developed ever more sophisticated and complex organisms. The basic mechanisms and processes that bring about evolution are believed to be known: The genetic instructions are encoded in the sequence of nucleotide pairs of DNA. The carrier substance of genetic information is the same for all living organisms (with the exception of RNA viruses). Evolution comes about by mutation, recombination (mating, transduction, etc.) selection of the fittest and changes in the environment of a species.

Could one not utilize a similar method for an organized evolution of knowledge and technology (under well-defined circumstances) and for numerical optimization problems? There are numerous optimization problems that are difficult even for the largest and fastest computers. Evolution through selection, mutation and recombination seems to be nature's own principal optimization method. What can one learn from it for "software" design? And what can one learn about biology at the same time?

One might think that the mathematics of evolution has essentially been worked out. In fact, the opposite is the case. Dobzhansky,[10] in his address to the 5th Berkeley Symposium on Statistics and Probability pointed out that most of mathematical genetics makes assumptions that simplify the mathematics to a point where it becomes tractable, but only at the price of oversimplification. Dobzhansky writes[10]: ". . . The classical model of genetic population structure has until recently received the lion's share of attention. It has the advantage of simplicity but the disadvantage of misrepresenting reality. It is not entirely played out, and probably never will be, since it does contain a grain of truth, for some

* This work was supported by ONR Contracts Nonr 3656(08) and Nonr 222(85).

genes and for some environments its simplifying assumptions are satisfactory as approximations. But the biological reality is different, and if I may say so, more interesting than the classical model suggests." In the same article Dobzhansky further writes: "The difficulty stems from the premises and the assumptions. Most exasperating is the habit of certain mathematical geneticists who make their assumptions implicit rather that explicit, on the grounds that to them the truth of their assumptions seems self-evident."

Dobzhansky lists in his paper the assumptions that underlie the classical model and he criticizes in particular the assumption of a "constant environment" as unrealistic. In this author's studies another frequently made assumption appears most restrictive: independence of individual genes. This assumption reduces the problem of finding a genotype of optimal fitness from 2^n potential cases to $2n$ cases (for n genes with two alleles each). For $n =$ (10,000 a rough estimate of the number of genes in Drosophila and man) $2^{10,000}$ is larger that the number of particles in the universe, while in comparison 20,000 is a very manageable number.

Even when interaction between genes is introduced it is usually done in a way (e.g. by making specific assumptions about pairwise interactions and neglecting all "higher order interactions") that makes the model hardly more realistic. There are many phenotypic features that depend necessarily upon entire blocks of interacting genes (cf. Bremermann[3,5]),

If a realistic model of the evolution of a species under mutation and selection is mathematically intractable, computer simulation may give some insight.

Several authors have dealt with evolution processes experimentally. None of the simulations mentioned in the following, however, are pure genetic models. They are biologically inspired experiments oriented towards potential practical applications. Nevertheless they *are* at the same time simulations of evolution processes.

In mathematical genetics one ascribes to each gene combination a fitness value. In an evolving population those genotypes that have the highest fitness value are selected (or enjoy a "selection advantage"). In a *simulated evolution* one can apply selection to *parameter combinations* that give the highest value for some *given function*. In analogy to genetics we call this again the "fitness function."

Friedberg[14] was among the first experimenters who tried to apply mutation and selection to a fitness function that depends upon its vari-

ables (genes) in a complicated way: He tried to let a computer evolve a short program, the variables (genes) being the code bits of machine commands. He found that in some cases evolution took 1000 times longer than straight search through all possible variable combinations would have taken.

Friedberg's fitness function is highly "discontinuous"—or, since we have discrete values of the variables we should rather say: The values of the fitness function, in general, are not close whenever the Hamming distance between genotypes is small.

Samuel's checkers programs[19] are another example of an evolution scheme featuring selection of the "fittest" parameter combinations. These programs evolve optimal *heuristics* through mutation and selection. The game of checkers, like many other games, is finite but has too many cases for explicit evaluation. Thus board positions are evaluated according to a collection of "rules of thumb" and a weighted sum of the scores of these various rules is used to decide the next move. This weighted sum is the *heuristic*. The weight factors of the heuristic are the variables which are modified through the evolution process. Samuel's program is one of the most effective game playing schemes to date. On the other hand, the "fitness function" is not known explicitly and thus it is hard to assess the efficiency of evolution.

Of a similar nature as Samuel's program is Findler's[11] machine that generates and optimizes its strategy for the game "Go-Moku."

Fogel[12,13] and co-workers have evolved finite state automata that "predict" prime numbers. This is a very interesting piece of work, though the meaning of the results is difficult to interpret for the purpose of understanding optimization through evolution. Again, the fitness functions is not explicitly known, and thus it is hard to determine whether an evolved automaton is optimal or not.

SUMMARY OF EXPERIMENTS

This author, in cooperation with S. Salaff, M. Rogson, H. de Grasse and J. Goguen has experimented with evolution processes involving mutation, recombination and "selection of the fittest." In contrast to the experiments cited in section 1 we studied only problems where the

fitness function was explicitly known. Thus, in each experiment we knew exactly where the optimum was and we could concentrate on the evolution process *per se*. In the spirit of "Bionics" we experimented with processes that were inspired by biology, that, however, might hopefully lead to useful technologies—"software" in this case.

The early work was concerned with systems of linear equations: given a non-singular square matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$, find $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \boldsymbol{b}$. Beginning with a random vector $\boldsymbol{x}_0$ we compute $\|\boldsymbol{Ax}_0 - \boldsymbol{b}\|$, the length of the vector $\boldsymbol{Ax}_0 - \boldsymbol{b}$, measured either in the Euclidian norm or in some other norm. Then "mutant" vectors are generated by changing individual components of $\boldsymbol{x}_0$. If a mutant has smaller length in comparison with its progenitor it is taken to substitute the progenitor, otherwise it is discarded and another mutant is tried.

One would expect that through repetition of this process the length of $\boldsymbol{Ax} - \boldsymbol{b}$ is reduced close to zero which implies that $\boldsymbol{x}$ is close to the solution vector. This, in general, is not the case. It turns out that in many cases a point is reached where *more than one component* of the vector $\boldsymbol{x}$ has to be changed in order to reduce the value of $\|\boldsymbol{Ax} - \boldsymbol{b}\|$ further while all changes of a single component $\boldsymbol{x}$ produce an *increase* in $\|\boldsymbol{Ax} - \boldsymbol{b}\|$.

The probability of finding among the many possible combinations of components one of those *particular combinations* of components of $\boldsymbol{x}$ whose mutation reduces $\|\boldsymbol{Ax} - \boldsymbol{b}\|$ in many cases becomes so small, that improvement does not occur in thousands, millions, or even millions of millions of trials. In this case the process has reached a *stagnation point*. We found this phenomenon to be extremely persistent, no matter how we varied the details of the process.

Mating helped little to alleviate the situation. An extensive analysis of these phenomena and of reasons for their occurrence is contained in a recent paper: Bremermann, Rogson and Salaff,[7] while [6] contains a summary of results.

On the basis of this work I venture to conjecture that many biological species are at a *genetic stagnation point* rather than at an optimum[7].

If one would be exclusively interested in biological evolution then one might have concentrated on investigating stagnation phenomena. With an eye on software technology, however, we concentrated on how stagnation could be overcome.

To this end, we switched from linear systems to linear programming

problems. Linear programming is a widely used optimization technique. A linear function on Euclidean space is to be optimized subject to a number of linear constraints. A point satisfying the constraints is called a *feasible solution.* Since the constraint set (set of points satisfying the inequalities) is convex, the average of the coordinate vectors of two feasible solutions is again feasible. Thus we were able to experiment extensively with "mating" (averaging). The formal structure of linear programming is well understood, thus our fitness function and its optimum were known and we could concentrate on the solution process *per se.*

Initially very vexing stagnation phenomena were encountered. Then we modified the process: Instead of mutating (adding increments to) the components of the coordinate vector of a feasible solution we added increments to *differences* of feasible solutions. The new method proved much more effective and the most successful program eventually alternated between asexual and sexual cycles. When asexual evolution (which is simpler and requires less data processing) stagnated, the program switched automatically to a sexual cycle which is more time consuming but which would tend to break the stagnation. The details of this work are described in references 7 and 18.

Similar methods proved effective in the case of *convex programming problems* which differ from linear programming by having an arbitrary convex set as constraint set rather than a convex polyhedron. A summary is reported in ref. 6 and the details in Rogson[18] (copies available upon request from this author).

Numerical methods in practical use often utilize gradients. If the function to be optimized is not explicitly known and if sampling is noisy, then the computation of a precise gradient may be difficult. We thus tried to combine an evolution process with a subroutine that determines an approximate gradient. When the gradient is established, the function in question is *optimized on the ray in the direction of the gradient*, then the program switches back to an evolution cycle. Various variations with large and small families, sexual and asexual evolution, more or less accurate approximation of the gradient, were tried. The details (flowcharts, results) are being reported in a recent Technical Report (de Grasse[9]) which may be obtained from this author upon request.

The method was slightly more effective than Rogson's methods. From the experiments it appeared that a *precise computation of the gradient was disadvantageous:* The computation time required in many cases was

greater than the time saved from knowing the gradient with greater precision.

Throughout our experiments we mutated essentially individual components—in analogy to genetic mutations. This amounts to a highly non-uniform biased sample of possible directions in $\mathbf{E}^n$. Some stagnation phenomena may be traced to this sampling method (Bremermann, Rogson, Salaff[7]). Mating has the effect of giving a less restricted sample and thus is sometimes useful in overcoming stagnation.

Unless there are substantial savings in computation time, it seems that uniform sampling of directions is preferable. *The non-uniform sampling of biological genotypes may be due to biological constraints rather than to superior performance of the method.*

De Grasse's experiments[9] suggest the following method for optimizing a function: Instead of computing a gradient, sample directions until one is found for which the function decreases (or increases). Then optimize the function on the ray in this direction (this is a one-dimensional optimization which is an easy problem). Then iterate the process.

In the following section we will show that this scheme which was conceived as a result of our experiments is

a) an efficient procedure (possibly competitive with conventional methods of solving linear equations)
b) can be analyzed
c) can serve as prototype of results that might be obtained in a systematic, though as yet largely non-developed, *theory of search.*

ANALYSIS OF OPTIMIZATION ALONG RANDOM RAYS

Near an optimum the level surfaces of a sufficiently smooth function look like concentric ellipsoids. Thus the level surfaces of $\|\boldsymbol{Ax} - \boldsymbol{b}\|_e$ are typical for all sufficiently smooth functions and thus an investigation of the effectiveness of search procedures for this class of functions should give approximate results for all sufficiently smooth functions.

We call the procedure that was described at the end of the previous section *optimization along random rays.*

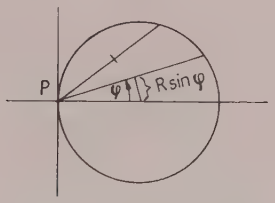
We will in the following always consider uniformly distributed random directions: By this we mean the following: Represent the random directions by unit vectors. Then the endpoints of these vectors lie on a unit

sphere. We say that the random directions are *uniformly distributed* if the endpoints are uniformly distributed on the sphere (constant probability density). In some cases we will consider random directions uniformly distributed on a hemisphere. We (speak of a "sphere" in any dimension rather that of "circle," "sphere" and "hypersphere".)

1. Search for the Center of a sphere

Given a sphere of radius R and a point P on the sphere. Half of the random directions at P intersect the sphere and half of them do not. Thus the probability of finding a "good direction" (a direction intersecting the sphere) is $\frac{1}{2}$, independent of the dimension. If a direction is "not good," then the opposite is "good."

We consider the case of dimension 2 (circle). We draw a diagram by putting P at the origin and the center of the circle at $(R, 0)$.



The equation of the circle, in polar coordinates, is $r = 2R\cos\varphi$. The minimum distance of a ray through P, making an angle φ, is obviously at the point $(\varphi, R\cos\varphi)$. The distance from the center is $\sqrt{(R^2 - R^2\cos^2\varphi)} = R\sin\varphi$.

Because of the uniform distribution of the random directions the angle φ between a "good direction" and the normal P is uniformly distributed between 0 and $\pi/2$. At each application of the process the distance from the origin is reduced by a factor $\sin\varphi$. Since successive reductions are multiplicative, the expected average reduction factor f_{exp} is

$$f_{\text{exp}} = e^{\frac{2}{\pi}\int_0^{\pi/2} \log_e \sin\varphi \, d\varphi}$$

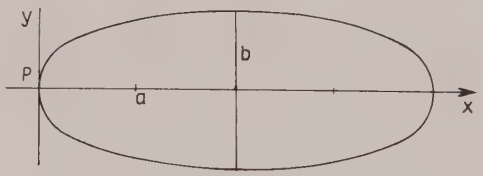
We have $$\int_0^{\pi/2} \log_e \sin\varphi \, d\varphi = -\frac{\pi}{2}\log_e 2.$$

Thus $$f_{\exp} = \frac{1}{2}.$$

Thus the expected number of optimizations along a random ray (counting only "good directions") that is necessary to reduce the distance of P from the center of the sphere by a factor $(\frac{1}{2})^m$ is simply m, or about 10/3 per factor 10.

2. Search for the Center of an Ellipsoid

We employ the same scheme as in the case of the sphere. Before dealing with the general case we consider the case of dimension 2: an ellipse in the plane.



We have drawn the ellipse with center at $(a, 0)$. It has the equation $\left(\frac{x-a}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$ in Cartesian coordinates.

At any point P on the ellipse the "good directions" lie in a half plane bounded by the tangent through P. We minimize the function $F(x, y) = \left(\frac{x-a}{a}\right)^2 + \left(\frac{y}{b}\right)^2$ along the rays through P. It is intuitively clear that the expected reduction of the minimum of $F(x, y)$ is the smallest when P is an endpoint of the major axis. We will estimate the expected reduction factor for P in this position. Without loss of generality we may assume P to be at the origin.

The "good directions" lie in the half plane $-\pi/2 < \varphi < \pi/2$ and the angle φ is uniformly distributed. The equation of the ray in direction φ is $x = \lambda \cos\varphi$, $y = \lambda \sin\varphi$, where λ is a parameter. The minimum of $F(x, y) = \left(\frac{x-a}{a}\right)^2 + \left(\frac{y}{b}\right)^2$ on the ray may be obtained by substituting the

equation of the ray and setting the derivative with respect to λ equal to zero:
We obtain

$$F(x(\lambda),\ y(\lambda)) = \frac{\lambda^2 \cos^2 \varphi}{a^2} - \frac{2\lambda \cos \varphi}{a} + 1 + \frac{\lambda^2 \sin^2 \varphi}{b^2},$$

and

$$F'(x(\lambda),\ y(\lambda)) = \frac{2\lambda \cos^2 \varphi}{a^2} - \frac{2 \cos^2 \varphi}{a} + \frac{2\lambda^2 \sin^2 \varphi}{b^2}.$$

Let λ_0 be the solution of $F' = 0$. We have

$$\lambda_0 = \frac{\cos \varphi}{a} \frac{1}{\frac{\cos^2 \varphi}{a^2} + \frac{\sin^2 \varphi}{b^2}}.$$

A short computation gives

$$F(x(\lambda_0),\ y(\lambda_0)) = \frac{1}{\left(\frac{b}{a} \cot \varphi\right)^2 + 1}$$

For $a = b$ we have a sphere and $\frac{1}{(\cot \varphi)^2 + 1} = \sin^2 \varphi$, which is consistent with our previous result. (Note that for a sphere $F(x, y)$ is the square of the distance of the point (x, y) from the center, divided by a).

We denote $F(x(\lambda_0), y(\lambda_0))$ by $R^2(\varphi)$. In order to compute the expected reduction of R^2 we have to compute $e^{\frac{2}{\pi}\int_0^{\pi/2} \log R^2(\varphi)\, d\varphi}$. In order to evaluate the integral $-2/\pi \int_0^{\pi/2} \log [(b/a \cot \varphi)^2 + 1]\, d\varphi$ approximately we observe that the integrand is small when $b/a \cot \varphi$ is small. For simplicity's sake we neglect the integrand for all values of φ for which $b/a \cot \varphi < 1$, which is equivalent to $b/a < \tan \varphi$ (which has an obvious geometric interpretation).

For $b \approx a$ the ellipse is approximately a sphere, and the previously computed estimate applies approximately.

Leaving aside the intermediate case we assume $b \ll a$. In this case $\tan\varphi \leqq b/a$ implies $\tan\varphi \approx \varphi$. Thus, letting $c = b/a$, we have

$$-\frac{2}{\pi}\int_0^{\pi/2} \log\left[\left(\frac{b}{a}\cot\varphi\right)^2 + 1\right] d\varphi \approx \frac{2}{\pi}\int_0^{c} \log\frac{\varphi^2}{c^2+\varphi^2}\, d\varphi$$

$$= \frac{2}{\pi}\left[2\varphi(\log\varphi - 1) - \varphi(\log(\varphi^2 + c^2) - 2) - 2c\arctan\frac{\varphi}{c}\right]_0^c$$

$$= \frac{2}{\pi}\left[2c(\log c - 1) - c(\log^2 c^2 - 2) - \frac{\pi}{2}c\right]$$

$$= -c\left(1 + \frac{2}{\pi}\log 2\right)$$

Thus the expected reduction factor is

$$\left(2^{\frac{2}{\pi}}e\right)^{-\frac{b}{a}} \approx (4.23)^{-\frac{b}{a}} > \left(\frac{1}{4}\right)^{\frac{b}{a}}$$

(Note that this factor corresponds to the *square* of the factor obtained in the previous case).

3. Higher Dimensions

For higher dimension the integrals that give the average improvement factors are more complicated than the ones for $n = 2$. Efforts to evaluate these integrals numerically have not yet been successful. Instead experiments with a computational realization of the search process itself have given statistical data about the behavior at higher dimensions. These data (unpublished, Technical Report in preparation) indicate that: 1) The number of iterations required to reduce R^2 by a given factor is proportional to the axis ratio a/b (for $a/b \gg 1$) not only for dimension 2 but *for any dimension* (Experiments were carried out with ellipsoids of rotation, one semiaxis of length a, all other semiaxes of length b). 2) The number of iterations required to reduce R^2 by a given factor, for fixed ratio a/b, *depends linearly upon dimension.*

The cases tested so far constitute a limited sample (dimension n between 2 and 100, ratio a/b between 2 and 256), however, the trends

found appear to be significant. A direct theoretical derivation of these expectation values, so far, is lacking and constitutes an interesting challenge.

SOLVING SYSTEMS OF LINEAR EQUATIONS THROUGH OPTIMIZATION ALONG RANDOM RAYS

We recall that $\|Ax - b\|^2$ is an ellipsoid (for Euclidean norm). Suppose we have already evolved or initially chosen a vector $\boldsymbol{x}_0$ and we have $\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}\|^2 = R^2$. Then we choose a random direction $\boldsymbol{u}$ (uniformly distributed) and minimize $R^2(\lambda) = \|\boldsymbol{A}\boldsymbol{x}_0 + \lambda\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}\|^2$. We find λ_0 by differentiating and setting the derivative zero. Using "$\cdot$" to denote the scalar product we obtain:

$$R^2(\lambda) = (\boldsymbol{A}\boldsymbol{x}_0 + \lambda\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) \cdot (\boldsymbol{A}\boldsymbol{x}_0 + \lambda\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b})$$

$$R^2(\lambda_0)' = 2(\boldsymbol{A}\boldsymbol{x}_0 + \lambda_0\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}) \cdot (\boldsymbol{A}\boldsymbol{u}) = 0$$

implies

$$\lambda_0 = \frac{\boldsymbol{b} \cdot \boldsymbol{A}\boldsymbol{u} - (\boldsymbol{A}\boldsymbol{x}_0) \cdot (\boldsymbol{A}\boldsymbol{u})}{\|\boldsymbol{A}\boldsymbol{u}\|^2}$$

Then we substitute $\boldsymbol{x}_1 = \boldsymbol{x}_0 + \lambda_0\boldsymbol{u}$ for $\boldsymbol{x}_0$ and iterate the process. The computation of $\boldsymbol{A}\boldsymbol{u}$ requires n^2 multiplications and n^2 additions while the dot products require n multiplications and n additions each. $\boldsymbol{A}\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{x}_0 + \lambda_0\boldsymbol{A}\boldsymbol{u}$ requires n multiplications and n additions. Thus one iteration cycle, aside from generating the random direction, requires $n^2 + 4n$ multiplications and $n^2 + 4n$ additions plus one division and subtraction.

We recall that the ratio of the largest to the smallest axes of $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 =$ Constant is the condition number P of the system. In section 3 we estimated for $n = 2$ that the expected number of iterations that is required to reduce R^2 by a given factor is proportional to P (for large P). While we were unable to estimate the expected number of iterations theoretically for $n > 2$, computer experimentation with the method just described indicates that the linear dependence upon P remains valid for higher dimension also, while the dependence upon n (for fixed P) is also linear (compare the comments under "Higher Dimension" in section 3).

Thus the computational effort required to achieve a reduction of R^2 by a given factor appears to be proportional to $n^3 + \text{const} \times n^2$. The time required for solving a linear system by conventional routines (elimination) also increases like n^3 plus a constant times n^2.

In our method we can avoid the accumulation of round-off errors by computing $\boldsymbol{A}\boldsymbol{x}_1$ anew (instead of from $\boldsymbol{A}\boldsymbol{x}_0 + \lambda_0 \boldsymbol{A}\boldsymbol{u}$). This adds n^2 multiplications and additions. In exchange for this extra effort any round-off errors that may have accumulated will only have the effect of "re-setting" the problem to a new R^2. *By iterating the process sufficiently often any desired precision of the solution vector can be obtained.*

1. Solution of Very Large Systems

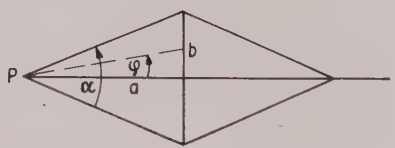
If we can multiply in one microsecond (10^{-6} sec), an iteration for $n = 10^3$ will take of the order of a second, but for $n = 10^4$ already 100 seconds. For $n = 10^4$ one would need 10^8 memory locations for the matrix —4 billion bits—thus the practical limits of this method seem to lie somewhere between $n = 10^3$ and $n = 10^4$.

Perceptrons and learning matrices (Compare Steinbuch and Schmitt[20] in these Proceedings) may have their limitations when used for pattern recognition, but they are good for computing scalar products $\boldsymbol{w} \cdot \boldsymbol{x}$, where $\boldsymbol{w}$ is a "weight vector." Our process requires essentially nothing but the repeated computation of dot products. Thus learning matrices conceivably could be utilized as special purpose computers for linear systems.

2. Sup Norm

In our experiments with linear systems we also worked with the sup norm instead of the euclidian norm (Bremermann-Rogson-Salaff[7]).

In the plane, the figure of constant sup norm of $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$ is a translated and rotated rhomb



Let the semi-axes be a and b. It is intuitively clear that the most difficult spot for making progress is either endpoint of the larger axis. The probability of finding a "good direction" at P is $\alpha/2\pi$. The improvement factor for a "good direction" at P is $\alpha/2\pi$. The improvement factor for a "good direction" that makes an angle φ with the major axis is $a/b \tan \varphi$ (see diagram).

The expected improvement factor is thus

$$f_{\text{exp}} = e^{\frac{2}{\pi}\int_0^{a/2} \log\left(\frac{a}{b}\tan\varphi\right) d\varphi}$$

For $b \ll a$ we have $\tan \varphi \approx \varphi, \frac{\alpha}{2} \approx \frac{b}{a}$. Thus we obtain

$$\int_0^{a/2} \log\left(\frac{a}{b}\tan\varphi\right) d\varphi \approx \int_0^{b/a} \log\left(\frac{a}{b}\varphi\right) d\varphi = -\frac{b}{a}$$

Thus $f_{\text{exp}} = e^{-(2/\pi)(b/a)}$. (We have assumed here as before, that in the case of a "bad direction" we automatically pass to the opposite. If not, we have to divide the exponent by 2.) Note the f_{exp} applies to $\| \ \|_{\text{sup}}$ rather than $\| \ \|^2_{\text{sup}}$—as was the case for the Euclidean norm, where we found a factor of about $4.23^{-b/a}$. Taking this into account we have to compare $e^{-4/\pi}$, which is approximately 3.56, with 4.23.

Thus the sup norm behaves very similarly to the euclidean norm—even though the surfaces of constant sup norm have non-differentiable edges.

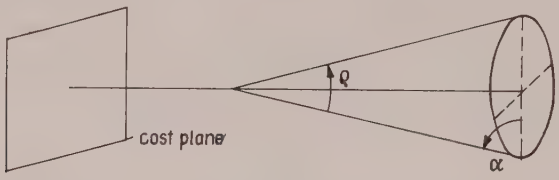
ANALYSIS OF A SEARCH PROCEDURE FOR LINEAR AND CONVEX PROGRAMMING PROBLEMS

1. Goguen's Example

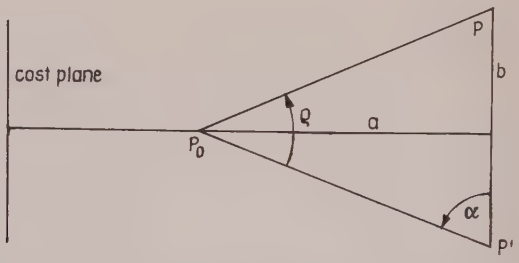
The following example and its analysis (with a slightly different approximation) is due to Goguen.[16]

Consider a right circular cone with axis perpendicular to a plane not intersecting the cone. Consider the convex programming problem with the cone as constraint set and the plane as cost plane. The task is to find

the point in the cone closest to the cost plane. This point we know to be the cone's vertex.



The cone may be of any dimension $n \geqq 2$. For $n = 2$ we have the following picture:

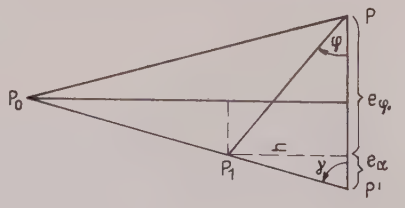


Note that $\alpha = \dfrac{\pi - \varrho}{2}$

The search method considered is as follows: Starting at a point P on the boundary (we know that the optimum point lies somewhere on the boundary), choose directions (pointing towards the cost plane) at random. When a direction is found that lies in the cone follow it until the boundary is intersected again, obtaining a new point P_1. Then iterate the process from the new point P_1.

We assume *uniformly distributed* random directions. In this detail our method differs from a similar search method employed in[18]. There the random directions are generated by mutation and are not uniformly distributed. A non-uniform distribution would be harder to analyze.

We first consider the case $n = 2$. Suppose we have found a "good direction." What is the reduction in "cost" associated with it?



Suppose the angle between $\boldsymbol{PP}_1$ and $\boldsymbol{PP}'$ is φ. The altitude h in the triangle $P'P_1P$ through P_1 satisfies the equations: $h = l_\varphi \tan\varphi$ and $h = l_\alpha \tan\alpha$ (see diagram). We also have $l_\alpha + l_\varphi = 2b$, $b/a = \cot\alpha$. Thus

$$h(\cot\alpha + \cot\alpha) = 2a\cot\alpha,$$

$$h = \frac{2a\cot\alpha}{\cot\alpha + \cot\varphi}$$

Thus the distance a is reduced to

$$a\left(1 - \frac{2\cot\alpha}{\cot\alpha + \cot\varphi}\right) = a\frac{\tan\alpha - \tan\varphi}{\tan\alpha + \tan\varphi}$$

Thus the *expected reduction factor* is

$$e^{\frac{1}{\alpha}\int_0^\alpha \log\left(\frac{\tan\alpha - \tan\varphi}{\tan\alpha + \tan\varphi}\right)d\varphi}$$

(Note that $0 < \varphi < \alpha$, otherwise P_1 does not exist or P_1 has a larger cost than P).

In order to estimate the integral we change variable: $u = \tan\varphi$ and obtain

$$\int_0^\alpha \log\left(\frac{\tan\alpha - \tan\varphi}{\tan\alpha + \tan\varphi}\right)d\varphi = \int_0^{\tan\alpha} \log\left(\frac{\tan\alpha - u}{\tan\alpha + u}\right)\frac{du}{1+u^2},$$

for $0 \leqq \varphi \leqq \frac{\pi}{4}$ we have $1 \geqq \frac{1}{1+u^2} \geqq \frac{1}{2}$. Thus we estimate the integral by evaluating (letting $c = \tan\alpha$):

$$\int_0^c \log\left(\frac{c-u}{c+u}\right)du = [(u-c)\log(c-u) - 1]_0^c$$

$$-(c+u)[\log(c+u) - 1]_0^c = -2c\log 2.$$

Thus for $0 \leqq \alpha \leqq \pi/4$ the expected value of the reduction factor satisfies the inequality

$$\left(\frac{1}{4}\right)^{\frac{\tan\alpha}{\alpha}} \leqq f_{\exp} \leqq \left(\frac{1}{2}\right)^{\frac{\tan\alpha}{\alpha}}$$

For $\alpha \to 0$ we have $f_{\exp} \to 1/4$.

The probability of finding a "good direction" is α/π. Thus the expected reduction factor $\tilde{f}_{\exp}$ of the cost of P, $C(P)$, minus the cost of P_0, $C(P_0)$, *per trial* satisfies the inequality

$$\left(\frac{1}{4}\right)^{\frac{\tan\alpha}{\pi}} \leqq \tilde{f}_{\exp} \leqq \left(\frac{1}{2}\right)^{\frac{\tan\alpha}{\pi}}$$

for $0 \leqq \alpha \leqq \pi/4$.
Thus the *expected number N of trials* to reduce $C(P) - C(P_0)$ by a factor 1/4 satisfies

$$\frac{\pi}{\tan\alpha} \leqq N \leqq \frac{2\pi}{\tan\alpha}$$

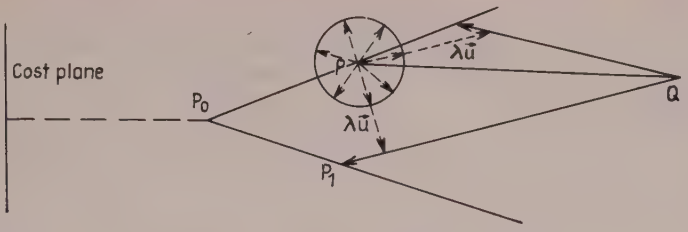
Note that $N \to \infty$ for $\alpha \to 0$. Thus for a *very flat* cone the process is unpractical. General convex sets with a smooth boundary locally look very much like a flat cone. Thus this particular process is not very suitable for solving convex and linear programming problems unless the optimum is at a point such that the constraint set is contained in a cone with vertex of the optimum such that $\alpha \neq 0$ and not too small. In the following we will discuss a modified search method that does not suffer from this disadvantage.

A MODIFIED SEARCH METHOD FOR CONVEX AND LINEAR PROGRAMMING

In the previous method (section 5) the probability of finding a "good direction" is α/π and the number of trials necessary to reduce the cost by a factor of 1/4 (or any other factor less than one) tends to ∞ for $\alpha \to 0$.

We now describe a modified process that avoids this difficulty: We select a point Q in the cone which remains fixed throughout the process. To the vector $\boldsymbol{QP}$ we add a random unit vector $\boldsymbol{u}$, multiplied by a factor λ. Let the unit vectors be uniformly distributed over the sphere. Follow the direction $\boldsymbol{OP} + \lambda\boldsymbol{u}$ from P until the boundary of the cone is intersected. Call this point P_1. If P_1 has a cost smaller than P iterate the process from P_1, otherwise try another direction $\boldsymbol{u}$.

It is clear that for λ small enough the probability of $\boldsymbol{OP} + \lambda\boldsymbol{u}$ being a good direction is 1/2.



It is clear the expected improvement approaches zero for $\lambda \to 0$. On the other hand the probability of $\boldsymbol{QP} + \lambda \boldsymbol{u}$ being a "good direction" approaches zero for fixed λ as P approaches the vertex P_0. Thus λ has to be modified in the course of the process. It has to be kept "between Scylla and Charybdis." This can be done through a feedback loop that reduces λ whenever the frequency of finding good directions falls below a threshold.

The search procedure sketched here is almost identical to the highly successful "buckshot method" of Bremermann-Rogson[8] and Rogson,[18] also Bremermann-Rogson-Salaff.[7] λ corresponds to the "cone factor" of the "buckshot method." There is one difference, however: in[7, 8] and[18] the directions $\boldsymbol{u}$ were generated through mutation and mating and the resulting distribution of $\boldsymbol{u}$ is not uniform but biased in favor of the coordinate directions. It seems that this non-uniformity is responsible for certain stagnation phenomena that were found. An explicit analysis of this process for non-uniformly distributed directions $\boldsymbol{u}$ also seems harder than for uniformly distributed $\boldsymbol{u}$.

It seems that the method described in this section may be economical for large linear and convex programming problems. This author intends to analyze the process fully in a further paper.

THEORY OF SEARCH

Searching for optima is the fundamental occupation of many private, public and military enterprises. The stockholder wants to maximize profit, the taxpayer wants to minimize the cost of public works and services, the manager wants to maximize plant utilization and efficiency, the manager wants to maximize plant utilization and efficiency, the strategist wants to maximize "pay-off," etc. The computational

tasks of optimizing even such a comparatively modest thing as a chess strategy seems to transcend the capability of even the most powerful existing computers.

Blessed with the sudden riches of bigger and better computers every year, where only two decades ago there was nothing but desk calculators and slide rules, one might be tempted to sit back and wait till computer technology catches up with the problems.

Ashby[1] has pointed out that we may wait in vain. There seems to be a definite ceiling beyond which data processing cannot be pushed (Bremermann[2,4,5]).

Thus estimates on the number of steps that are required to carry out a process, to solve a given problem are important now and will remain so in the future. Yet little work has been done in this direction.

Gelfand and Tsetlin[15] have pointed out that theoretical methods of optimization (e.g. "compute the partial derivatives with respect to the variables and solve for the common zeros") are frequently utterly impractical.

Much thought has gone into finding *algorithms* to solve problems rather than search procedures. However, algorithms are not always the most efficient way to solve problems. For example, Cramer's method (using determinants) is *not* the method to solve large systems of linear equations numerically.

On page 607 we have outlined a search procedure for systems of linear equations that seems superior to most algorithms. If a search procedure can beat algorithms on such a classical and simple problem as linear equations, then we should be encouraged to try search procedures on more complex problems, where no efficient algorithms are known (e.g. searching for strategies, optimizing "weights" in a multilayer neural net, etc.)

To guide our way, however, we need some insight, some theory. Thus there is a need for a *theory of search procedures*. The examples that we have outlined involve *feedback* (one could call these procedures cybernetic). Thus our procedures are different from merely *iterative* procedures, that apply the same operation to a set of data over and over again irrespective of the results of the operation.

The basic problems to be investigated by a theory of search are these: Given a problem (e.g. optimization) and a search method. One would like to have an estimate of how many steps it takes to reach the solution

(or an approximate solution) and how much actual data processing the procedure involves. In many cases such estimates seem hard to obtain and there is not much in the literature that is concerned with this question (an exception: Hook and Jeeves[17] and Wild[21]). Beginning with very simple examples, it seems possible, however, to obtain some results in this direction, and we hope to have made a very modest contribution toward this end.

REFERENCES

1. Ashby, R. Some consequences of Bremermann's limit for information processing systems. This volume pp. 69–76.
2. Bremermann, H. J. Optimization through evolution and recombination, in "*Self-Organizing Systems* **1962**," Yovits, Jacobi, Goldstein, editors, Spartan Books, Wash. D. C. 1962.
3. Bremermann, H. J. Limits of genetic control, *I.E.E.E. Transactions Military Electronics*, Vol. MIL-7, (Butsch, Oestreicher, eds.) 1963, pp. 200–205.
4. Bremermann, H. J. Quantum noise and information. *Proceedings of* **5***th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 15–22 (1967).
5. Bremermann, H. J. Quantitative aspects of goal-seeking self-organizing systems: *Progress in Theoretical Biology*, Vol. 1, 59–77 (1967).
6. Bremermann, H. J., Rogson, M, and Salaff, S. Search by evolution, in *Biophysics and Cybernetic Systems*, Maxfield, Callahan, and Fogel, editors, Spartan Books, Wash. D. C., 1965, pp. 157–167.
7. Bremermann, H. J., Rogson, M., and Salaff, S. Global properties of evolution processes in *Natural Automata and Useful Simulations*, Edelsack, Fein, Pattee and Callahan, editors, Spartan Books, Wash. D. C., 1966, pp. 3–41.
8. Bremermann, H. J., and Rogson. M. An evolution-type search method for convex sets, Technical Report, ONR Contracts Nonr 3656(08) and Nonr 222(85), University of California, May 1964.
9. Bremermann, H. J. and de Grasse, H.: A quasi-gradient method in convex programming *Technical Report*, same contracts as ref 8. June 1966.
10. Dobzhansky, Theodosius. Genetic diversity of environments, *Proceedings of the* **5***th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, 295–304 (1967).
11. Findler, N. V. A machine that generates and optimizes its strategy. Bionics symposium 1966. Short paper preprints. Wright-Patterson Air Force Base, Ohio, 1966.
12. Fogel, L. J., Owens, A. J., and Walsh, M. J. Artificial intelligence through the simulation of evolution, in *Biophysics and Cybernetic Systems*, Maxfield, Callahan, and Fogel, editors, Spartan Books, Wash. D. C., 1965, pp. 157–167.
13. Fogel, L. J. *Artificial intelligence through evolutionary programming*, John Wiley and Sons, New York, 1966.

14. Friedberg, R. M. A learning machine, IBM *Journal, Research and Development*, Part I: **2**, 2–13, January 1958, Part II: **3**, 282–287, July 1959.
15. Gelfand, I. M., and Tsetlin, M. L. Some methods of control for complex systems, *Russian Mathemat. Surveys*, **17**. no. 1, 1961.
16. Goguen, J. A. Some considerations on evolutionary algorithms, *Technical Report*, same contracts as 8), February 1966.
17. Hook R., and Jeeves, T. "Direct search" solution of numerical and statistical problems, *J. of the A.C.M.*, April 1961.
18. Rogson, M. A search method in convex programming, *Technical Report*, same contracts as 8), February 1965.
19. Samuel, A. L. "Some studies in machine learning using the game of checkers," E. A. Feigenbaum and Julian Feldmann, editors, *Computers and Thought*, pp.71 to 105, McGraw Hill, New York, 1963.
20. Steinbuch, K. and Schmitt, E. Adaptive systems using learning matrices. *This volume* pp. 751–768.
21. Wilde, D. J. *Optimum seeking methods*, Prentice Hall, Englewood Cliffs N.J., 1964

E. M. HARTH

Department of Physics

Syracuse University

Syracuse, New York

Time Dependencies in Memory

INTRODUCTION

Much has been learned in recent years about the functioning of the central nervous system, its microstructure, as well as its organization. However, very little progress has been made toward a solution of the central problem, that of identifying the physical changes that must underlie the acquisition, retention, and recall of sensory experience.

A number of experiments have demonstrated that ribonucleic acids are at least by-products of neuronal activity, and may well play a dominant role in memory and learning.* Among these are the well-known studies by Hydén and coworkers who have shown recently that marked increases in RNA content occur during the learning process in selected cortical neurons (Hydén, 1965). But the mechanism by which RNA stores information—if that is its function— is totally unknown, apart from a demonstration of hysteresis effects in polynucleotides, which make them potential memory elements (Katchalsky, 1965). Alternatively, it is believed by many, that the RNA merely serves to initiate the synthesis of proteins, which in turn produce structural or physiological changes in the neurons, and thus the memory trace (Dingman and Sporn, 1964). The regions believed to be most sensitive to small morphological changes, are the synaptic junctions between neurons, and *synaptic facilitation* is still one of the most plausible ways for the brain to acquire a memory. It should be emphasized, though, that to date few persistent changes have been observed, and none have definitely been shown to be responsible for learning.

The absence of long-term morphological changes has led to the postulate of a dynamic memory, i.e. one which is sustained by continuous activity in the CNS, rather than being deposited as a latent trace. Such a memory, however, would be extremely vulnerable and should be just

about destroyed by such conditions as deep anesthesia, electroshock, or epileptic seizure. It is therefore generally agreed, that a purely dynamic memory can at best account for short-term retention only, and that more permanent mechanisms are required for long-term memory.

On the other hand, the absence of observed long-term changes should not be taken too seriously, since slow changes in the morphology of cortical neurons would be difficult to establish, and even more difficult to correlate with neuronal activity.

The time dependencies of memory provide some fascinating clues to the whole problem of information storage in the CNS, but are far from having been deciphered at this stage. In particular, the dramatic phenomena of anterograde amnesia (AA) and retrograde amnesia (RA), following, for example, head injuries, cerebral anoxia, or epileptic attacks, provide data which are likely to contribute greatly to a future theory of learning.

To appreciate the problem involved here, we cite Gooddy's analogy between the memorizing brain and a tape recorder (Gooddy, 1964). AA is then explained simply as the breakdown of the recording mechanism, just as damage to the tape recorder at time t causes the tape to be blank from t until the time the damage is repaired. The brain, fortunately, is capable of a considerable amount of self-repair, hence recording will often recommence spontaneously after a period of inactivity, depending on the severity of the damage.

A puzzling aspect, not analogous to the tape recorder (even though Gooddy attempts to extend the analogy to this point) is the fact that, upon close examination of the memory, we find that events immediately preceeding the trauma have disappeared from the record, even if these events had nothing to do with the trauma, and most certainly would have been part of the memory, had the accident not occurred. This record-destroying property of RA, as distinct from the mere loss of recording ability in AA, is selective with respect to time, but nonselective with respect to content or sense modality: the destroyed record is always the most recent, covering a certain time interval immediately preceding the trauma. Eventual recovery of the memory always starts with the earliest events, and gradually advances to, but oftentimes never quite reaches, the instant of the trauma.

It follows from the above observations, that a physical distinction must exist between recent and more distant memory. This distinction

is most likely not spatial (as on the tape), hence some character of the memory trace must change with time.

We have mentioned before the vulnerability of a dynamic memory. This fact suggests that perhaps for a period of time the information is held in the form of reverberating neuronal circuits, and only gradually becomes *fixed* as a persistent structural change in network parameters (Gerard, 1961). Thus, if we interfere with ongoing neuronal activity, all dynamic memories will be destroyed, and events which have not yet become fixed, will be lost.

GENERAL FORMULATION OF MODEL

In the present paper we attempt to find a quantitative description of the time dependence of memory processes, with particular reference to the phenomenon of RA. The description is by no means unique, but represents a possible set of assumptions. More importantly, it is hoped that the method employed here may point the way toward new experiments, and the utilization of forthcoming data.

We begin by defining a quantity $\lambda(t)$, the *learning parameter*, which represents the extent to which a given event is fixed as memory at time t. The physical nature of this engram is left open, $\lambda(t)$ could be defined in terms of scores in a learning experiment. For the sake of illustration only we shall adopt a point of view similar to Hebb's cell assembly theory (Hebb, 1963). Memory is assumed to come about through synaptic facilitation between neurons belonging to the cell assembly in question. More specifically, we could define synaptic coupling coefficients between neurons (Harth, 1966) and define λ as the increase in the average coupling coefficient for neurons belonging to a particular cell assembly.

We shall not exclude the possibility that $\lambda(t)$ involves a number of different physical changes, all contributing to the engram. These changes may involve, for example, various structural features of the synapses. Each one of these changes $\lambda_i(t)$ is assumed to have its own dependence upon the inflow of sensory information, and its own characteristic time decay.

We add the following specific assumptions:

1. The different parameters λ_i are additive in determining the effective learning parameter $\lambda(t)$.

2. The parameter λ_i shall have the dynamic range $0 \leqq \lambda_i \leqq \Lambda_i$.

3. The increase in λ_i will occur as a result of experiencing a particular event, and in proportion to the length of time the event is experienced. If the occurrence of the event is described by the function $\epsilon(t)$, where ϵ is either 0 or 1, depending on whether at that moment the event is perceived or not, then the rate of change in λ_i can be written $\alpha_i(\Lambda_i - \lambda_i)\epsilon(t)$. Here the factor $(\Lambda_i - \lambda_i)$ merely insures that λ_i does not increase beyond Λ_i, and α_i is the rate of increase of λ_i when $\epsilon = 1$ and $\lambda_i = 0$.

4. Each of the learning parameters λ_i will, in the absence of sensory input ($\epsilon = 0$), decay with a time constant $1/\beta_i$.

We summarize the above assumptions by writing

$$d\lambda_i = \alpha_i(\Lambda_i - \lambda_i)\,\epsilon(t)\,dt - \beta_i\lambda_i dt \tag{1}$$

As a particularly simple case we consider the learning due to continuous experience, expressed by $\epsilon(t) = 1$ for all t. Equation 1 will then have the solutions

$$\lambda_i = \frac{\alpha_i\Lambda_i}{\alpha_i + \beta_i}\,[1 - \exp(-(\alpha_i + \beta_i)t)]$$

Each physical change λ_i will tend toward the fraction $\dfrac{\alpha_i}{\alpha_i + \beta_i}$ of its saturation value Λ_i with a time constant $1/(\alpha_i + \beta_i)$.

On the other hand, a single experience may be thought of as a step function $\epsilon = 1$ for $0 \leqq t \leqq \Delta t$, and $\epsilon = 0$ for all other times. The memory λ for times following the experience will then be

$$\lambda(t) = \sum_i \frac{\alpha_i\Lambda_i}{\alpha_i + \beta_i}\,[1 - e^{-(\alpha_i+\beta_i)\Delta t}]\,e^{-\beta_i t}$$

Clearly with increasing t those components of memory will predominate, which have a high acquisition rate α_i and low decay rate β_i.

It is difficult to explain with this model, how single experiences of short time duration can leave lasting memory traces. Also the phenomenon of RA, particularly the subsequent recovery of memory, remain unexplained.

To remedy these shortcomings, we now modify our model as follows:

a) Following the single occurrence of an event (assumed now to be instantaneous) we assume that the neuronal activity resulting from the event, will briefly continue in what has loosely been termed *reverberations*. This extension of the primary experience will fade rapidly with time; we

take this decrease to be exponential, with a mean reverberation time $1/\varrho$, which may or may not depend on the event experienced.

b) Strengthening of the memory trace may occur also after the reverberations have died out, due to the fact that elements of the specific experience continue to occur in the steady stream of sensory inputs, which we call the *normal experience*. The rate of change of λ_i due to this effect is taken to be proportional to the total strength of the engram $\lambda(t)$. The constant of proportionality, γ, is thus a measure of the extent to which the specific event is coupled to the normal experience.

Assumptions (a) and (b) are expressed by writing the function $\epsilon(t)$ for a single event occuring at $t = 0$ as

$$\epsilon(t) = e^{-\varrho t} + \gamma \sum_i \lambda_i \tag{2}$$

Equation 2 expresses the fact that the time interval, during which the event itself occurs, is too short for any real learning to take place. This assumption has been stated formally as the *adiabatic learning hypothesis* (Caianiello, 1961). We further assume that the more or less permanent physical changes that constitute learning will occur during the reverberation period and, to a lesser extent, at all later times. Substitution of (2) into (1) gives rise to a system of coupled, non-linear differential equations.

To begin with, we shall restrict ourselves to the case of a single physical change λ with acquisition rate α and decay rate β. From (1) and (2) we then obtain the equation

$$d\lambda = \alpha(\Lambda - \lambda)\,(e^{-\varrho t} + \gamma\lambda)\,dt - \beta\lambda dt \tag{3}$$

Orders of magnitude for the parameters in (3) may now be discussed. The fixing of memory takes place in times of the order of minutes. By contrast the persistence of memory after prolonged neuronal disturbances, such as unconsciousness following brain injury, shows that there exists a component with a decay time of perhaps months or longer. If we assume that a substantial amount of learning takes place during the reverberation time following a single occurrence of the event, the learning rate α cannot be very much smaller than the reverberation rate. The parameter γ has units of $(1/\lambda)$. Restricting ourselves to cases in which the primary event is only weakly coupled with the normal experience, allows us to

set $\gamma\Lambda \ll 1$. For short times then

$$d\lambda \approx \alpha(\Lambda - \lambda)\, e^{-\varrho t} dt$$

hence

$$\lambda \approx \Lambda\left\{1 - \exp\left[-\frac{\alpha}{\varrho}(1 - e^{-\varrho t})\right]\right\} \tag{4}$$

where $\lambda = 0$ at $t = 0$. The asymptotic value of λ, call it l, for times $1/\beta \gg t \gg 1/\varrho$ will therefore be

$$l = \Lambda\left[1 - \exp\left(-\frac{\alpha}{\varrho}\right)\right] \tag{5}$$

Similarly we can write an equation for long times, $t \gg 1/\varrho$. From (3)

$$d\lambda \approx \alpha(\Lambda - \lambda)\gamma\lambda dt - \beta dt \tag{6}$$

which has the solution

$$\lambda = \frac{l(\alpha\gamma\Lambda - \beta)}{l\alpha\gamma + [\alpha\gamma(\Lambda - l) - \beta]\exp[-(\alpha\gamma\Lambda - \beta)\, t]} \tag{7}$$

Here the "initial" value l is taken to be equal to the asymptotic solution (5) to the short-term equation.

The overall behavior of the function $\lambda(t)$ is shown in Figure 1. Here the rapid rise in region A reflects the learning during the reverberation period. This is followed by the long-term solutions for different conditions.

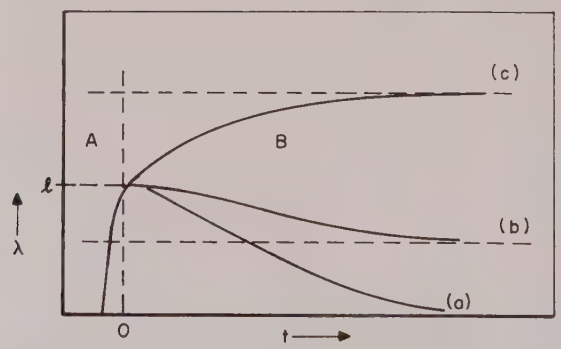


Fig. 1. Time dependence of λ. Region A represents the solution to the short-term equation, region B to the long-term equation, for (a) $(\alpha\gamma\Lambda - \beta) < 0$, (b) $0 < (\alpha\gamma\Lambda - \beta) < l$, (c) $(\alpha\gamma\Lambda - \beta) > l$.

DISCUSSION

We attempt now to give more physical meaning to some of the quantities discussed so far, and shall interpret, where possible, the results of the theory in terms of experimental findings.

The quantity α represents the rate at which the physical change responsible for the fixing of memory takes place initially, i.e. when the function $\epsilon(t)$ in (1) is unity and $\lambda = 0$.

The nature of this physical change is likely to be connected with the production of RNA in the neuron. We mentioned the strong experimental linkage between learning and RNA. It was found also that increases in protein content, as well as that of lipids and phospholipids accompany increases in neuronal RNA (Palladin and Vladimirov, 1956). But the strongest suggestion, that protein synthesis is involved in learning, comes from the observation that inhibitors of protein synthesis appear to interfere with the fixation process (Agranoff, Davis and Brink, 1965).

If, at the same time, we wish to retain the theory that synaptic facilitation constitutes the learning process, we have to explain the transport of chemicals from the cell body, where they are manufactured, to the synapses, where they act. This transport is probably mediated by the axoplasmic flow (Weiss and Hiscoe, 1948) and should cause a delay between the acquisition of memory (as evidenced for example by changes in RNA content) and its behavioral effect, as demonstrated by scores in learning experiments. We would therefore expect a high initial score due to reverberations, followed by a sharp drop, followed by a rise over a period of hours. Such learning curves were reported for normal subjects, while Korsakoff patients failed to show the rise (Talland, 1965).

The quantity $1/\varrho$, the mean reverberation time, may be taken to be a constant, or, more realistically, dependent upon the *importance* of the experience, an important experience having the longer reverberation time. It may be presumed that highly reduced neuronal activity (low temperature, anoxia, deep anaesthesia) as well as greatly heightened activity (hyperthermia, electroconvulsive shock) will terminate, or at least interfere with the reverberation process. If the reverberation time is of the order of $1/\alpha$, a single experience should be sufficient to produce a substantial memory trace, whereas repeated experiences would be required in the case of $\varrho \gg \alpha$. In an interesting experiment Hebb (1963) has shown that even in this situation a permanent, though subliminal

memory trace will be produced, and that it will have an additive effect on later learning experiences.

Finally, the quantity γ reflects the relatedness, or coupling between the experience that is remembered and the continued normal influx of information. Thus the memory of an event will be enhanced by the continued inflow of fragments of the original experience, or accidental similarities encountered in everyday experience. It is aided, furthermore, by the reinforcement of the fragments themselves into conceptual units in what Hebb calls "the conceptual development as the basis of learning."

Thus, apart from fluctuations, γ should be low at infancy for most stimuli, and increase gradually with maturation.

PUROMYCIN INDUCED AMNESIA

Memory fixation was studied in goldfish by Agranoff, Davis and Brink (1965), by administering intercranially varying doses of puromycin, an inhibitor of protein synthesis. These authors found that long-term memory (tested 3–4 days after training) was significantly reduced, if the drug was given less than one hour after the learning experience. At the same time short-term memory did not seem to be affected. If the drug was administered more than one hour after the learning experience, it did not cause any measurable impairment of long-term memory.

In more recent experiments it was shown that a decay of the retention score with a time constant of about one day follows the 40-minute training session if puromycin is injected immediately following the training session (Davis and Agranoff, 1966).

From the earlier experiment we would gather that the reverberation time in goldfish is of the order of a fraction of an hour, and that the puromycin injection will obstruct the fixing process. Thus, if the injection follows the learning experience by more than an hour, virtually all of the fixing will already have taken place. The known action of the drug as an inhibitor of protein synthesis is suggestive of the fact that fixing, unlike the reverberations, involves synthetic processes. Thus, in terms of the parameters of Eq. (3) the effect of puromycin may be expressed as a temporary reduction of the quantity α. This is shown schematically in Figure 2. Here (a) shows the abrupt drop of α at the time of the puromycin injection ($t = 0$), and the subsequent recovery. In curve (b) we show the reverberation terms for events occurring at different times before,

during, and after the period of effectiveness of the drug, assuming that ϱ remained unaffected. Finally, in (c) are drawn the corresponding curves of λ/Λ vs, time, using equation 4. We note, in particular, the reduced memory for events occuring just before administration of the drug and the absence of any memory traces for events occurring while $\alpha = 0$. A smaller dose of the drug may produce only a reduction in the parameter α, in which case λ may continue to rise slightly after $t = 0$ This will give rise to a less pronounced dependence of λ on the time between the last learning trial and the puromycin injection. This is indeed the result observed by Agranoff, Davis and Brink (1965).

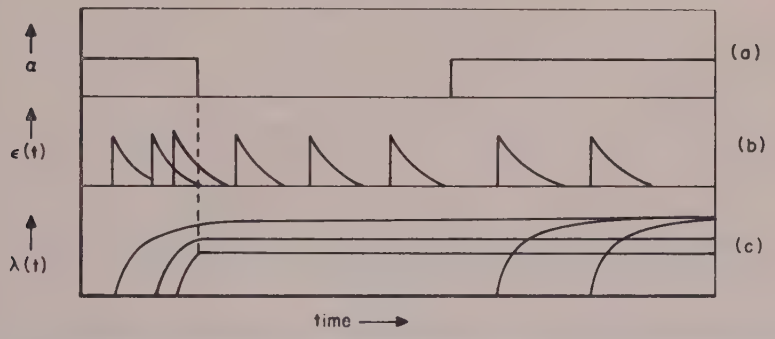


Fig. 2. Effect of impairment of fixing mechanisms. a) $\alpha(t)$; (b) reverberation term for events occuring at various times. (c) $\lambda(t)$ from Eq. 4 (short-term solution only.)

The interpretation given above is highly oversimplified. The chief objection lies in the fact that the learning experience extends over a period of 40 minutes; or about as long as the period we had deduced for the reverberation time. It would appear, therefore, that a considerable amount of fixing should have occurred already at the end of the training session, hence, the puromycin injection at that time should not have eliminated all of the long-term memory. Indeed, the later experiment shows the presence of a 24-hour component of memory (Davis and Agranoff, 1966). If this is a true memory (as distinct from a reverberation), then its fixing must be unaffected by puromycin, since it occurs even if the drug is injected prior to the 40-minute training session. The long-term memory (> 3 days) must then be attributed to a different physical process, λ_2, presumably of lower decay rate β_2. The acquisition of this memory must occur for the most part after the 40-minute training period and yet mostly within the first hour following the end of the training.

One explanation for the acquisition of this memory would be to assume that it is generated by the intermediate process λ_1 via the second term in Eq. (2), i.e. $\gamma(\lambda_1 + \lambda_2)$.

If this process, unlike the first, is puromycin sensitive, then all of the data may be accounted for, as shown in the schematic drawing of Figure 3. The upper two curves, (a) and (b), show the reverberations

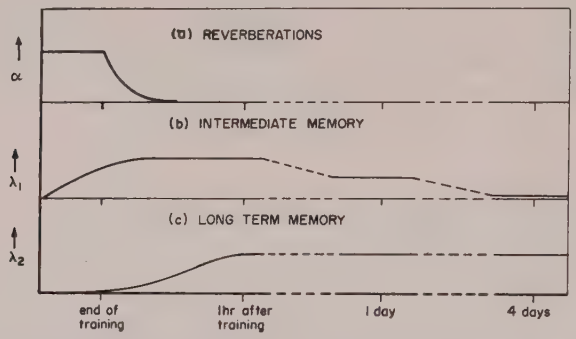


Fig. 3. Generation of long-term memory by an intermediate memory.
(a) Reverberation term due to an extended (40-minute) training period.
(b) Acquisition and decay of intermediate memory.
(c) Acquisition of long-term memory. (Note: to account for the data of Davis and Agranoff (1966), assume that α_1 is puromycin insensitive, α_2 is sensitive.)

and the intermediate, drug-insensitive memory, respectively. Curve (c) gives the long-term memory. It is seen that a puromycin injection any time during the first hour following the training period, would strongly impair the performance 4 days later.

ELECTROCONVULSIVE SHOCK

Recent experiments by Quartermain, Paolino and Miller (1965) on retrograde amnesia in rats, following electroconvulsive shock (ECS), show a dependence of acquired memory on the time interval between the (single) learning experience and the ECS, in excellent agreement with the short-term solution (Eq. 4) of our learning equation. The time constant deduced from these data is about 15 seconds. It is not clear in this case, whether the effect is due to a breakdown in the fixation process, or a cessation of reverberations, or both. The attempt made in that

experiment, to distinguish between the two alternatives, is inconclusive. Decreases have been observed in the RNA content of brain tissue in rats, following ECS (Mihailovic *et al.*, 1958) as well as in Cardiazol induced convulsions (Zakhov and Orlanskaya, 1960). This too, while seeming to favor the first interpretation, leaves open both possibilities. The argument could be settled, if we knew the effect on both short-term and long-term memory, due to ECS preceeding a learning experience.

In man therapeutic electroshock almost invariably produces short retrograde amnesias, typically of a few minutes duration (Mayer-Gross, 1943).

AMNESIAS IN MAN

Traumatic amnesias, global amnesia and the more chronic conditions associated with the Korsakoff syndrome, can provide considerable information regarding the learning process.

Severe concussions usually cause both RA and AA. Typically, the RA lasts from a fraction of a minute to several minutes, rarely for hours or days. AA may vary between a fraction of an hour and several days. The two symptoms are related in that cases of short RA generally have short AA also (Russell, 1959). With few exceptions, events occurring during AA will not be recovered later on, while the period of RA will begin to shrink upon full recovery of consciousness, the more distant memories always returning first.

One way to account for the permanent inability to form a memory of events that occurred during the period of AA, is to assume the complete cessation of the fixing process ($\alpha \to 0$) during that time, similar to the situation in puromycin induced amnesia. The strong preponderance of RA times of the order of minutes would indicate reverberation times of the same order.

An alternative way of relating the data on traumatic amnesia to our model, would be to assume that only the capability of forming reverberations is impaired as a result of injury. In either case the time dependence of λ is the same as that shown in Figure 2 for the case of $\alpha = 0$. Empiricially, the question of whether or not the reverberation time remains unchanged during AA could be settled by observing the presence or absence of short-term memory during AA.

In global amnesia the immediate causes are unknown, but the onset is as sudden and quite as dramatic as in traumatic amnesia. Again, both AA and RA are always present, an initial RA, which may cover weeks or even years, rapidly shrinks in the course of a few hours to a permanent RA of a few hours or less than an hour (Fisher, 1964). Again, the confused period following the onset until full recovery a few hours later, is one during which no permanent memory is laid down.

Among the characteristics of the chronic syndrome following Wernicke's disease, is the non-specific memory loss involving almost always both RA and AA. It has been pointed out that the deficiencies of the Korsakoff syndrome are not merely an inability to recall, since hypnosis failed in all cases to elicit the memory of any event within the amnesia periods (Talland, 1965).

Many authors have reported on the extremely rapid decay of new memory traces in Korsakoff patients. Thus, Crahey (1954) and Talland (1965) report that, while immediate recall of an event was not impaired, substantial memory loss appeared in a matter of seconds. This would indicate a substantial increase in ϱ, i.e. shortening of the reverberation time. Earlier observations (Bonhöffer, 1904) that permanent memories may be formed of events of particular significance to the patient, show that α is not zero, and may even be unaffected; it also supports the suggestion we made above, that ϱ should depend on the significance of the event.

In attempting to interpret the above observations on amnesias in man, we look for a typical time span of RA which we could associate with the reverberation time in our model. We must distinguish, first of all, between the occasional long RA's which cover weeks or even years (decades in some of the advanced cases of Korsakoff psychoses), and the true residual RA's. The first occur mostly in the confused periods of AA and may be due to an impairment of the mechanisms of recall. The last group covers – as we have seen – periods from seconds to a few hours, still a considerable span. It is interesting to note that the least severe disturbances (electroshock, mild concussions) cause the shortest RA's whereas the more severe traumas cause both long AA's and long RA's.

It is tempting to argue that this relationship between RA and AA, which was first pointed up by Russell (1959), can be explained by assuming that the memory trace λ, which exists at the start of the AA, is

further decaying at the rate β as long as α remains zero. In this picture then a long residual RA is *caused* by the long AA. The diagram in figure 4 attempts to illustrate this situation. Here the curve depicts the short-term solution (Eq. 4). Various decays are shown (straight lines) for different values of λ at $t = 0$. A threshold λ_{min} for λ is assumed (dotted line),

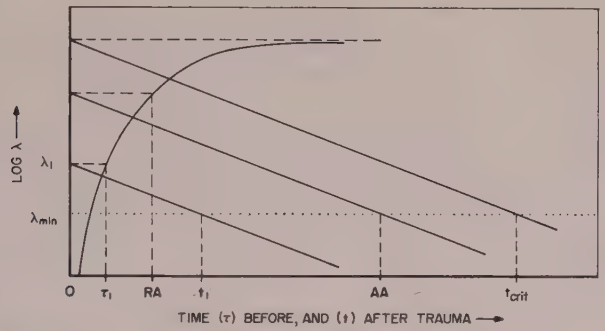


Fig. 4. Schematic of a possible interpretation of dependence of *RA* on *AA*. An event occurs at a time τ_1 before trauma. From the ordinate of the curve representing the short-term solution (Eq. 4) at τ_1 we obtain the value of λ at $t = 0$, the instant of the trauma. This value is labeled λ_1. Subsequently it will decay (solid straight line), falling below threshold of recall at time t_1. It follows that the duration of *RA* may be derived from the length of the *AA*, as shown. Further, there should be an *AA* (t_{crit} in diagram), such that all previous memory is wiped out ($\tau = \infty$).

such that an event is considered *remembered* if $\lambda > \lambda_{min}$, and *forgotten* if $\lambda < \lambda_{min}$. Consider now an event that occurred at time τ_1 preceeding a trauma and the onset of AA. From the curve we read the value of λ for $t = 0$, call it λ_1, (Fig. 4) and draw the line corresponding to the subsequent decay of the memory. To simplify the diagram, times τ before the trauma and t after the trauma are both drawn as positive numbers on the same abscissa. The time t_1 in the diagram is thus the duration of the AA which will just cause the event to be forgotten. Conversely, the diagram illustrates how, starting with a given AA, one can find the corresponding RA.

One difficulty with this very simple picture is the fact that a decay time $1/\beta$ of the order of hours or days has to be assumed to account for some of the data. If this were the only component of the memory, then complete retrograde amnesia (i.e. $\tau = \infty$) should result from AA's exceeding a critical value t_{crit}, as shown in Figure 4, which is not borne

out by the facts. Once again there is some indication that more than one fixation process is involved, in which case the more short-lived components would serve as the generators for the more permanent memory.

SUMMARY

Data on the time dependence of memory engrams following various disturbances of normal cerebral functioning, were used in conjunction with a simple dynamic model of memory.

It should hardly require emphasis, that the available data, some of which have been discussed briefly above, are still much too fragmentary and too heterogeneous to allow more than a semi-quantitative approach. What is not too much to hope for at this stage, however, is that certain very simple sets of assumptions may be tested for compatibility with the data. One such set is that permanent memory consists of a single physical component generated as a result of sensory input and decaying exponentially with a single time constant when normal brain functions are severely curtailed. We believe that strong indication for the existence of more than one fixing process—and more than one time constant—may be deduced from a comparison of the data with our model.

REFERENCES

* A listing of references citing evidence for changes in RNA content in neurons as a function of neural activity has recently been given by Pevzner (1966).

Agranoff, B. W., Davis, R. E., and Brink, J. J. *Proc. Nat. Acad. Sci.* **54**, 788 (1965).

Bonhöffer, K. *Allg. Z. Psychiat.* **61**, 744 (1904).

Caianiello, E. *J. Theor. Biology* **1**, 204, (1961).

Crahey, S. *Acta Neurol. Belga.* **57**. 570 (1954).

Davis, R. E., and Agranoff, B. W. *Proc. Nat. Acad. Sci.* **55**, 555 (1966).

Dingman, W., and Sporn, M. B. *Science* **144**, 26 (1964).

Fisher, C. M., and Adams, R. D. *Acta Neurol. Scand.* **40** Suppl. 9 (1964).

Gerard, R. W. in *Brain Mechanisms and Learning*, A. Fessard ed., pp. 21–36, Blackwell Scientific Publications, Oxford 1961.

Gooddy, W. *Brain* **87**, 75 (1964).

Harth, E. M., in *Automata Theory*, E. Caianiello ed., pp. 201–217, Academic Press 1966.

Hebb, D. O. *The Organization of Behavior*, John Wiley and Sons, 1963.

Hyden, H., and Lange, P. W. *Proc. Nat. Acad. Sci.* **53**, 946 (1965).
Katchalsky, E. *Hysteresis and Memory in Biopolymers*, Invited paper presented at Tenth Annual Meeting of the Biophysical Society, San Francisco (1965).
Mayer-Gross, W. *Lancet* **2**, 603 (1943).
Mihailović, L., Janković, B. D., Petković, M. and Isaković, K. *Experientia* **14** (1958) 144.
Palladin, A. V., and Vladimirov, G. E. Proc. Intern. Conf. on the Peaceful Uses of Atomic Energy, vol. 12, 401 United Nations, New York (1956).
Pevzner, L. Z. in *Macromolecules and Behavior*, J. Gaito, ed., pp. 43–70, Appleton Century Crofts, New York (1966).
Quartermain, D., Paolino, R. M., and Miller, N. E. *Science* **149**, 1116 (1965).
Russell, W. R. *Brain, Memory and Learning*, Oxford, Clarendon Press 1959.
Talland, G. A. *Deranged Memory*, Academic Press 1965.
Weiss, P., and Hiscoe, H. B. *J. Exper. Zool.* **107**, 315 (1948).
Zakhov, N. B., and Orlanskaya, R. L. *Vop. Med. Khim.* **6**, 249 (1960).

KENNETH KROHN

Krohn-Rhodes Research Foundation

Berkeley, California

RUDOLPH LANGER

Krohn-Rhodes Research Foundation and the University of California,

Berkeley, California

and JOHN RHODES

Krohn-Rhodes Research Foundation, the University of California, Berkeley, California

and the Institute for Advanced Studies, Princeton, New Jersey

*A Theory of Finite Physics with an Application to the Analysis of Metabolic Systems**

ABSTRACT

In Part I of this paper we present a theory of finite physics, i.e. of experiments having finite phase spaces. This theory is based on mathematical results in the theory of finite semigroups and the algebraic theory of finite state machines. We indicate how the principles of this theory can be interpreted to parallel the principles of classical physics.

Part II of the paper gives an application of this theory to the analysis of a composite model of bacterial intermediary metabolism. We conclude with a discussion of the relation of this theory to a theory of organization for arbitrary multienzyme systems.

PART I

We state two basic principles:

PRINCIPLE I

Any physical experiment gives rise to a collection of transformations on the phase space (or set of states) of the physical system. The action is given by assigning to each input a and each state q the new state q' observed.

* This research was sponsored in part by the Office of Naval Research, Medicine and Dentistry Branch, Contract Number N0014-66-C0172 and by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant Number: AF-AFOSR-848-65 and Contract Number: AF 49(638)-1550.

Thus for any experiment E we have

Q, the state (phase) space

A, the set of basic inputs

$f: A \times Q \to Q$, the action given by $f(a, q) = q'$

We may define a semigroup, which we will denote by $S(E)$ to be the semigroup generated under composition by the collection of transformations

$$\{f_a | a \in A\}$$

where $f_a: Q \to Q$ is given by

$$f_a(q) = f(a, q)$$

Denote $S(E)$ acting on Q by $(Q, S(E))$. We now state

PRINCIPLE II

A physical theory for E is merely a coordinatization of $(Q, S(E))$ in "wreath product," or "triangular" manner (see definitions below).

Note that at this point we have made no assumptions as to the finiteness of either A or Q.

Essentially "wreath product" or "triangular" coordinates provide a representation of Q in a direct product:

$$Q \subseteq X_1 \times X_2 \times X_3 \times \ldots \text{ i.e., } q \leftrightarrow (x_1, x_2, x_3, \ldots)\, x_i \in X_i$$

in such a way that if f is any transformation in $S(E)$ and $q \leftrightarrow (x_1, x_2, x_3, \ldots)$ is any element of Q, and $f(q) = q' \leftrightarrow (x_1', x_2', x_3', \ldots)$, then x_j' depends only on $x_1, x_2, \ldots, x_j$ for $j = 1, 2, 3, \ldots$ Thus we have a decomposition of f into $(f_1, f_2, f_3, \ldots)$ where

$$f_k: X_1 \times X_2 \times \ldots \times X_k \to X_k$$

and

$$f((x_1, x_2, x_3, \ldots)) = (f_1(x_1), f_2(x_1, x_2), f_3(x_1, x_2, x_3), \ldots)$$

This wreath product coordinatization thus yields a "triangularization" of the action observed in experiment E. The connection between this representation of the action and several notions from classical physics will be brought out in examples at the end of this section.

We now give a precise definition of the wreath product for semigroups. See Krohn and Rhodes.[1]

DEFINITION

Let the pair (X_k, S_k) denote the semigroup S_k acting faithfully on the right of a set X_k. (The pair $(Q, S(E))$ is just such a pair).

Then (X, S) is the wreath product of (X_1, S_1), (X_2, S_2), ..., (X_k, S_k), ... written:

$$(X, S) = (X_1, S_1)\, w\, (X_2, S_2)\, w \ldots w\, (X_k, S_k)\, w \ldots$$

if and only if:

$$X = X_1 \times X_2 \times \ldots \times X_k \times \ldots \tag{1}$$

$$S = S_1\, w\, S_2\, w \ldots w\, S_k\, w \ldots \tag{2}$$

The meaning of (2) is that S is the set of all functions $f: X \to X$ such that

(a) $f((x_1, x_2 \ldots, x_k, \ldots)) = (f_1(x_1), f_2(x_1, x_2), \ldots, f_k(x_1, x_2, \ldots, x_k), \ldots)$

(b) $f_1 \in S_1$; and for k greater than 1;
$(a_1, a_2, \ldots, a_{k-1})$ a fixed point of $X_1 \times X_2 \times \ldots \times X_{k-1}$, and $g: X_k \to X_k$ defined by $g(x) = f_k(a_1, a_2, \ldots, a_{k-1}, x)$ implies $g \in S_k$, i.e., any function on X_k obtained by holding fixed in the first $k-1$ coordinates of some k component action of a function in S must be in S_k.

We now define a relation which formalizes the notion that a semigroup action denoted by the pair (X, S) is really just a "special case", or restriction, of the action (Y, T) or, equivalently, whenever one has the action (Y, T) one also has (X, T) acting somewhere within it.

DEFINITION

$$(X, S) \mid (Y, T)$$

read "(X, S) *divides* (Y, T)" iff there exists
(a) a subsemigroup $T_0 \subset T$ and a homomorphism

$$\tau: T_0 \to\!\!\!\!\to S \text{ onto} \quad \text{and}$$

(b) a T_0-invariant subset $Y_0 \subset Y$ and a function

$$\theta: Y_0 \to\!\!\!\!\to X \text{ onto}$$

such that for any $t \in T_0 \subset T$

$$\theta(t(y)) = \tau(t)\,(\theta(y)) \quad \text{for all} \quad y \in Y_0$$

i.e. $$\theta \cdot t|_{Y_0} = \tau(t) \cdot \theta$$

Let $C = \{(X_1, S_1), (X_2, S_2), \ldots, (X_k, S_k), \ldots\}$ then we define a *coordinatization of* $(Q, S(E))$ *from the class* C to be any solution of $(Q, S(E))|(X,S)$ where (X, S) is a *wreath product* of members of C.

In the event that Q is *finite*, a theorem from the algebraic theory of finite state machines[1] allows us to assert that especially simple coordinatizations of $(Q, S(E))$ exist. To be more specific, there are always coordinatizations of $(Q, S(E))$ from the class C where all the S_i occurring in pairs of C are either

(a) Non-commutative finite *groups* which have no non-trivial normal subgroups (Simple non-abelian groups or "SNAG's")
(b) Cyclic *groups* of prime order ("p-counters" or "Z_p's") or
(c) One of three *trivial semigroups* the largest of which has three elements ("Units" or "U_1", "U_2" and "U_3").

We postpone further discussion of the significance of this result for finite phase spaces until Part II.

We now wish to show some parallels to Principle II in notions of classical physics:

1. Conservation Laws Give A First Wreath Product Coordinate

A description of an experiment that takes place within a conservative system can often be given by stating that with each state q of the system is associated a value $x_1(q)$ such that under any input a this value either remains constant, or changes in a way dependant only on $x_1(q)$ (and no further aspect of the state) and a. Thus if input a drives q to q', we have

$$x_1(q) = x_1(f_a(q)) = f_{a_1}(x_1(q))$$

This conserved quantity $x_1(q)$ frequently corresponds to the energy or momentum associated with a state of the phase space. Any phase space that has such a conserved quantity can be coordinatized using two coordinates the first of which represents the quantity conserved. To do this we set up a correspondence

$$q \leftrightarrow (x_1(q),\ x_2(q))$$

in which $x_2(q)$ is chosen so as to make the correspondence one to one. Then for any $f_a \in (Q, S(E))$

$$f_a(x_1(q),\ x_2(q)) = (f_{a_1}(x_1(q)),\ f_{a_2}(x_1(q),\ x_2(q)))$$

exhibiting the desired wreath product action.

2. Symmetry Laws Give A Last Group Coordinate

Intuitively, the symmetries of the experiment described by $(Q, S(E))$ will be defined by those permutations of Q that are interchangeable with any transformation of Q. That is, the symmetries will be characterized by the *centralizer*, C, of $S(E)$ which is defined by:

$$c \in C \leftrightarrow c(f(q)) = f(c(q)) \quad f \in S(E),\ q \in Q$$

Assume $S(E)$ is transitive, i.e. there is an element of $S(E)$ that will drive any state of Q to any other state of Q. Then it is trivial to prove that C is a regular permutation group. That is, if $c \in C$ and $c(q) = q$ for some $q \in Q$, then $c(q) = q$ for all $q \in Q$. It follows that the equation $c(q_1) = q_2$, with $q_1, q_2 \in Q$, has at most one solution $c \in C$.

Let the orbits of Q under the action of elements of C be $\tau_1, \tau_2, \ldots, \tau_n$. Intuitively these orbits are the essentially asymmetric classes of states of Q, i.e. each τ_i is made up of all those states related by symmetry to one another and therefore all obtainable from some specified state, say p_i, by some unique symmetry transformation $c \in C$.

Since $\tau_i = C_{p_i} = \{c(p_i) | c \in C\}$, it follows that for $f \in S(E)$ we have

$$f(\tau_i) = f(C_{p_i}) = C_{f(p_i)}$$

which is contained in the orbit containing $f(p_i)$. Thus $f \in S(E)$ maps orbits into orbits.

Now let us assign to each $q \in Q$ the pair $(\tau(q), c(q))$ where $\tau : Q \to N$, $(N = \{1, 2, \ldots, n\})$ specifies the orbit of q, i.e.

$$q \in \tau_{\tau(q)}\ \forall\, q \in Q$$

and $c : Q \to C$ gives the unique symmetry transformation necessary to obtain q from $p_{\tau(q)}$, i.e.

$$c(q)\,(p_{\tau(q)}) = q \quad \forall q \in Q$$

We now have

$$(Q, S(E)) \subseteq (N \times C, N \text{ w } C)$$

since we may write any $f \in S(E)$ as (f_1, f_2) where:

$f_1 : N \to N$ and $f_1(k)$ gives the orbit to which f sends any state in τ_k. That is

$$f(\tau_k) \subseteq \tau_{f_1(k)}$$

Further, $f_2 : N \times C \to C$, and $f_2(k, c)$ gives the symmetry transformation required to reach $f(c(p_k))$ from $p_{f_1(k)}$. This function may be obtained from a function $\bar{f} : N \to C$ defined by

$$f(p_k) = \bar{f}(k)\,(p_{f_1(k)})$$

so that

$$f_2(k, c) = c \cdot \bar{f}(k)$$

This works, since $c(p_k) \leftrightarrow (k, c)$ and

$$\begin{aligned} f(c(p_k)) &= c(f(p_k)) \\ &= c(\bar{f}(k)\,(p_{f_1(k)})) \\ &= (c \cdot \bar{f}(k))\,(p_{f_1(k)}) \\ &= f_2(k, c)\,(p_{f_1(k)}) \leftrightarrow (f_1(k), f_2(k, c)) \end{aligned}$$

which establishes a wreath product action.

Now to every $f \in S(E)$ there can be associated an $n \times n$ matrix (n the number of orbits of Q under C) having as entries either 0 or elements of C. The matrix denoted by $\{f_C\}$ is defined as follows:

$$\{f_C\}_{ij} = \begin{cases} \bar{f}(i) & \text{if } \ j = f_1(i) \\ 0 & \text{otherwise} \end{cases}$$

That is, all the entries in the ith row are 0 except the $f_1(i)$th, and that is $\bar{f}(i) \in C$.

Note that this matrix $\{f_C\}$ is row-monomial, i.e., it has only one non-zero entry occurring in each row. Given that M is a matrix representation of the group C as $m \times m$ matrices, a representation of $S(E)$, the *induced representation*, is obtained by substituting $m \times m$ blocks of 0's for the zero entries of $\{f_C\}$ and $M(\bar{f}(i))$ for the non-zero entries.

3. Reversible Processes Give A Special Wreath Product Decomposition (Lagrange Coordinates)

To say that all processes in an experiment E are reversible means that for any input $a \in A$

$$q' = f(a, q) \Rightarrow \exists a' \in A \quad f(a', q') = q$$

i.e. whatever one input does, some other input can undo. If we define a "do nothing" input $e \in A$ by

$$f(e, q) = q \quad \forall q \in Q$$

then f_e will act as an identity for $S(E)$. Now we may state the reversibility property in terms of $S(E)$:

$$f_a \in S(E) \Rightarrow \exists f_a, \in S(E)\, 3 f_a f_a = f_e$$

i.e. every element of $S(E)$ has an inverse. It follows at once that:

The processes of E are reversible $\Leftrightarrow$ $S(E)$ is a group. This situation is, of course, very frequently found in classical physics.

In the event that the semigroup $S(E)$ is in fact a group there is always one way to enter wreath product coordinates from a class C in which all the S_i are simple groups. The coordinates obtained in this way are called "Lagrange coordinates" because of their relation to the subgroup structure of the group. We now sketch this procedure.

In the discussion that follows, let us denote this $S(E)$ which is a group simply by S and suppose H is a non-trivial *normal* subgroup of S. If no such H exhists then S is already simple and we are done. Let us denote the factor group G/H by B and let the natural map of each element of S onto its coset be denoted by $n: S \to B$, finally, select an arbitrary set of representatives for each coset and denote them by $\bar{}^{*}$, i.e., $b^* = f \in S$ where $n(f) = b \in B$, subject to the stipulation that $1_B^* = 1_S (= f_e)$. We now claim that

$$(Q, S) | (B \times H, \bar{B} \,\mathrm{w}\, \bar{H}) \qquad (*)$$

where $\bar{}$ denotes action by multiplication on the left, e.g., $\bar{t}(u) = t \cdot u$.

Indeed, if we fix some $q_\theta \in Q$ and define a map θ

$$\theta : B \times H \to Q \quad \text{by} \quad \theta(\mathrm{b}, h) = b^* \cdot h(q_\theta)$$

then θ is onto, for given any $q \in Q$, by transitivity of S, there is some $f \in S$ such that $q = f(q_\theta)$. Now if $n(f) = b$, for some $h \in H$, $f = b^* h$ so that

$$q = f(q_\theta) = b^* h(q_\theta) = \theta(b, h)$$

We remark that θ partitions $B \times H$, and hence S, by identifying all transformations which agree at the base state q_θ, and since q_θ is arbitrary it could be chosen to have desirable special properties relative to the particular application.

We recall that by the definition of wreath product, any element of $\bar{B}\, w\, \bar{H}$ may be written as (f_1, f_2) where

$$f_1 : B \to B \quad f_2 : B \times H \to H$$

We will now define a subset T_0 of $\bar{B}$ w $\bar{H}$ which is in one to one correspondence with S: for every element $f \in S$ pick an element $(f_1, f_2) \in \bar{B}$ w $\bar{H}$ where we define f_1 and f_2 by

$$f_1(b) = n(f) \cdot b$$

and

$$f_2(\mathrm{b}, h) = \left[((n(f) \cdot b)^*)^{-1} \cdot f \cdot b^*\right] \cdot h$$

Thus defined, (f_1, f_2) is indeed an element of $\bar{B}$ w $\bar{H}$, and the correspondence $\tau : T_0 \to S$ given by

$$\tau((f_1, f_2)) = f$$

can be shown to be an isomorphism. Thus, in particular, T_0 is a subsemigroup and τ an onto homomorphism which, along with the map $\theta : B \times H \to Q$ defined above, establish the division (*) according to the definition given earlier in this paper.

A few remarks on the nature of the correspondence $f \leftrightarrow (f_1, f_2)$ are in order. First we note that f_1 codes the coset action of f; it has the form of simple multiplication in B (i.e. $f_1 \in \bar{B}$) because of the normality of H. Also, if we denote the square bracketed operations in the definition of f_2 by $\bar{f}$ we have

$$\bar{f} : B \to \bar{H} \quad \text{and} \quad f_2(b, h) = \bar{f}(b)\,(h) = \bar{f}(b) \cdot h$$

and now if $f = b_0^* h_0$ then

$$\begin{aligned} \bar{f}(b) &= ((n(f) \cdot b)^*)^{-1} \cdot f \cdot b^* \\ &= ((b_0 \cdot b)^*)^{-1} \cdot f \cdot b^* = (b^*)^{-1} \cdot (b_0^*)^{-1} \cdot (b_0^* h_0) \cdot b^* \\ &= (b^*)^{-1}\, h_0 b^* \end{aligned}$$

Thus $\bar{f}$, which yields the element of $\bar{H}$ necessary to obtain the second coordinate under the action of f from the original second coordinate, is simply conjugation of the element h_0 (which distinguishes f itself in its coset $b_0^* H$) by the original first coordinate.

We have thus exhibited the construction which established a wreath product with first coordinate in the factor group. In the event that H is not simple we can obtain another H' normal in H and carry out this process (which can be computerized) again to obtain a three term wreath product. Indeed, if G has a subnormal series of maximum length L then this process gives rise to a decomposition into L coordinates each of which is an element of a simple group. It is this *algebraic* criterion that imposes the wreath product order on any group, i.e. any $S(E)$ for which E involves only reversible actions.

PART II

As we have seen, there are strong parallels between certain important properties of systems as studied in physics and certain formal, algebraic properties of abstract transformation semigroups. Such concepts as conservation and symmetry are seen to correspond to aspects of a special basic decomposition, the wreath-product decomposition, which seems to exist for many phenomena successfully analyzed by mathematical physics, and which is *known* to exist for any representation of an experiment on a *finite* phase space. At the present state of sophistication in classical physics, to characterize a system in a wreath product manner is really just to specify it in an obvious and "intuitive" way in terms of its "significant" parameters. This physical intuition was a long time in the making. There are many areas of experimental knowledge—including some in physics itself—where this intuition for building good theory simply does not now exist. It would therefore seem reasonable to see if our notion of physical theory, formally defined in Principle II, can provide some useful insight into such areas. In order to guarantee that a physical theory can be defined, we have chosen to investigate systems whose state spaces are discrete; however it seems plausible that physical theories under our definition exist for many significant continuous systems.

Biology in general, and cellular biology in particular, is an area of great interest and experimental activity which would seem to provide many examples of discrete state spaces. We have chosen a metabolic process to be considered as the physical experiment in Principle I. We wish to investigate a physical theory for this experiment using Principle II. There is no unique or traditional way to assign states and inputs to describe a metabolic process as there is for most phenomena studied in

physics. The problem is to use some of the traditional parameters in terms of which the experimental, as well as the theoretical, content of the field is expressed, and from them to obtain a state-input assignment. This problem of choosing states and inputs for domains never investigated from the state-space of view is delicate but amounts, in the final analysis, to the problem of choosing a model, and that problem arises whenever one applies any mathematical discipline to an area outside mathematics. We will therefore treat our own state assignment procedure in the case of metabolic systems in some detail.

We begin by stating our primary interest to be the investigation of the organization of normal cellular metabolism with an emphasis on the biochemical aspects. Thus in our model we will not attempt to describe such phenomena as membrane transport or diffusion, although such physical aspects are important, perhaps crucial, to the maintenance of many metabolic pathways which will be studied. Another aspect of our model is that it is not kinetic, i.e., it will not be concerned with rates or, strictly speaking, with equilibrium. The model will be built on the basis of biochemical reactions that are known to occur, but whose rates need not be known, especially when they occur as part of a complex system of reactions. Simplifications are usually necessary, sometimes even desirable, in any analysis; however distortions (and falsifications) are to be avoided as much as possible. Accordingly, the model has been conceived as being imbeddable in a real, albeit in vitro, situation.

Metabolism is viewed as a collection of biochemical reactions, and a metabolic pathway is a connected series of these. Since in living things there is very nearly a one to one correspondence between biochemical reactions and the specific enzymes that catalyze them, we will consider the terms metabolism, metabolic process, and multienzyme system to be synonymous as we have done tacitly heretofore. We characterize a multienzyme system M as $M(E, S, I)$ where:

E is a set of *enzymes* (high molecular weight catalytic proteins) denoted E_1, E_2, etc.

S is the set of *substrates* or metabolites (molecules usually of low molecular weight relative to the enzymes) produced by reactions catalyzed by elements of E and denoted by $s_1, s_2, \ldots$

I is a set of *inorganic ions* or "minerals" (e. g. inorganic phosphate, ammonium, metallic activator cations, etc.) required for reactions to be catalyzed by elements of E.

Now we wish to define a "metabolic state-space" Q_M corresponding to $M = M(E, S, I)$:

$$Q_M = \{q_i = E \cup I \cup \{s_i\} \mid s_i \in S\} \quad M = M(E, S, I)$$

Heuristically, we can regard each state $q_i \in Q_M$ as a suspension of all the enzymes of E in a solution with the inorganic ions of I to which has been added some fixed amount, say one mole (or one molecule), of substrate s_i. Elements of $E \cup I$ should be assumed to be in such high concentrations that they limit no reaction catalyzed by the system. Thus we can assume that any reaction catalyzed by elements of E for which the substrates are present goes to completion. We also assume that any state is an *open system* with respect to all substances produced by, or introduced into it except for elements of S. This is plausible since members of S may be thought of as being bound to specific members of E.

The state space which we use is derived from Q_M by identifying with q_n all subsets $P_{q_n} = \{q_i, q_j, \ldots, q_n\}$ of Q_M for which reaction sequences

$$s_i \rightarrow s_j \rightarrow \cdots \rightarrow s_n$$

exist that require only the elements of E or I (e.g. that require no cofactors). Intuitively this makes sense, since one mole of any element of P_{q_n} in the hypothetical solution yields state q_n (since time is not an explicit part of our model and all reactions go to completion). We will denote Q_M thus partitioned as Q_m and denote its elements by numbers. In general Q_m will have fewer states than either substrates or enzymes but not many fewer since most biochemical reactions do require some organic cofactor.

We can now study the activities of a multienzyme system $M(E, S, I)$ as experiments on its state space Q_m according to Principle I. To do this we need to define a set of basic inputs. In our present model this is A_m where:

A_m is the set of all cofactors (usually coenzymes or prosthetic groups) required for reactions to be catalyzed by elements of E, and not included in I

The stipulation that a true prosthetic group which is tightly bound to its enzyme, such as *FAD*, could be introduced by itself into a system, or that its reduced counterpart, $FADH_2$, would then diffuse out in a finite period of time is, admittedly, quite artificial. There are technical ways around this; e.g. let the entire flavoprotein, i.e. the *FAD* + apoenzyme complex, be the input. This is not necessarily more realistic in the event that the apoenzyme itself is part of an enzyme complex with other enzymes

of the system. The safest route is perhaps to consider models in which the system is not entirely open. This viewpoint is not at all repugnant to our theory and has been considered by us, but it induces complications not germane to the basic outline of our approach which we are attempting to present here.

To illustrate our definitions let us obtain the state space of the multi-enzyme system corresponding to the hypothetical metabolic map in Figure 1(a). This state space together with the actions of the inputs upon it is given in Figure 1(b). To obtain the states of Q_m from Q_M we made the following identification:

$$
\begin{aligned}
Q_M \quad & Q_m \\
\{q_1, q_2\} & \leftrightarrow 1 \\
\{q_3, q_4\} & \leftrightarrow 2 \\
q_5 & \leftrightarrow 3 \\
\{q_6, q_7\} & \leftrightarrow 4 \\
q_8 & \leftrightarrow 5 \\
q_9 & \leftrightarrow 6 \\
q_{10} & \leftrightarrow 7
\end{aligned}
$$

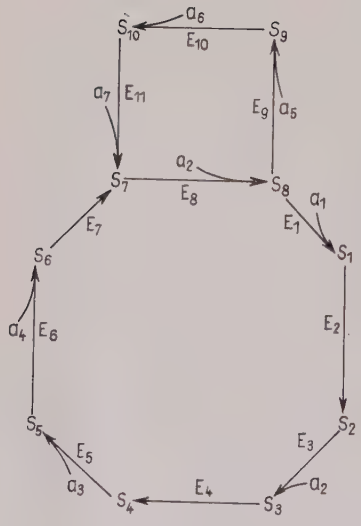


Figure 1(a).

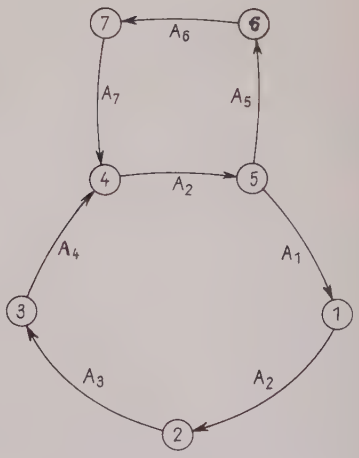


Figure 1(b).

The state space of Figure 1(b) is fully typical of experiments viewed in the light of Principle I but is not large enough to be of great interest organizationally. A model of some complexity is given in Figure 2. This model is in fact a representation of a composite multienzyme-system of about eighty enzymes which occur in various microorganisms (most would be found in E-coli) as enzymes of the intermediary metabolism, the hub of all metabolic activity. The reactions and pathways involved are almost all well known and may be found in standard references.[2,3,4,5,6] This data, which has been so painstakingly gathered

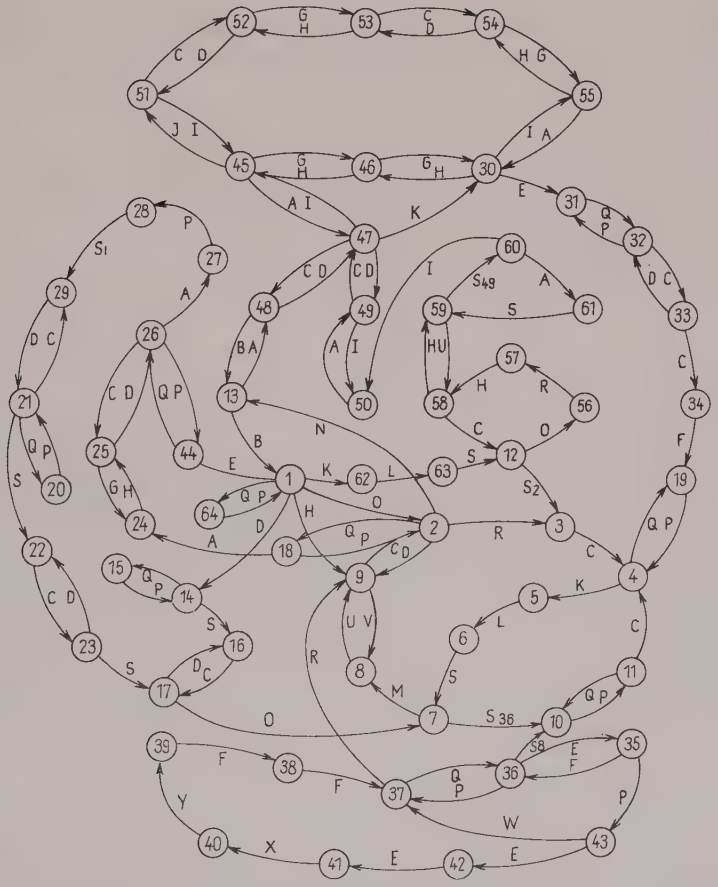


Fig. 2. A partitioned state-input model of bacterial intermediary metabolism.

over several decades, compiled together into a large system (as biochemists have been doing for years, see Umbreit[7,8]) should provide a firm basis for analysis using Principle II via Principle I. To indicate what we hope to learn by such an analysis, we resume our discussion of coordinatizations of finite phase spaces.

What we pointed out in the first part of this paper was essentially that an experiment carried out on a finite phase space could be analyzed as a product of simpler actions in concert. Furthermore, the way these actions were combined to yield the observed results of the experiment could also be specified. Its general character was "wreath product" and that is really only the way the component actions mesh, but what is specific to the experiment, and hence of greatest analytical import, is the identity and nature of these component actions. Thus, given that

$$(Q_m, S(M)) \mid (X_1, S_1) \text{ w } (X_2, S_2) \text{ w } \ldots \text{ w } (X_n, S_n) \qquad (*)$$

that is, given a coordinatization of the experimental (observed) action in a multienzyme system, no matter how complex, one has a decomposition of the global action in the transformation semigroups S_1 to S_n acting respectively on coordinate "aspects", X_1 to X_n, of the state space Q_m. We have seen that in physics these "aspects" may correspond to such concepts as the energy or the momentum or even the symmetries associated with a state of the system. The isolation of such concepts for any system is a concomitant of a wreath-product coordinatization for that system. Now, since for finite systems there are *purely algebraic* methods to obtain coordinatizations for whatever experiments may exist, there is thus the possibility of discovering ingredients for a theory explaining these experiments even where no theory of any kind existed before. This is the reason that we refer to a coordinatization as a "physical theory". To sum up, a "physical theory" (*) for a system implies two things:

(a) A classification of any state of the system in terms of a set of "aspects":

$$q \leftrightarrow (x_1, x_2, \ldots, x_n)$$

(b) A description of any observed action of the system in terms of restricted types of actions on each of the "aspects" of the system:

$$f(q) \leftrightarrow (f_1(x_1), f_2(x_1, x_2), \ldots, f(x_k, x_2, \ldots, x_n))$$

Returning to our multienzyme system M, we note that to every metabolic pathway there can be associated an experiment on Q_m. This experiment consists of successive application of the inputs required to drive the state corresponding to the initial substrate of pathway through the intermediate states and finally to the state corresponding to the end product of the pathway.* The important thing here is that this sequence of inputs corresponds in turn to a product of elements of the semigroup $S(M)$ and hence to a single element of $S(M)$. Thus every metabolic pathway corresponds to an element (or a set of elements) in a semigroup. The import of this element of the semigroup is broader than the metabolic path from which it arose, for it may also move states of the multienzyme system that are not on the path; in fact, it relates the individual metabolic path or paths it encodes to the whole multienzyme system. Since this is the case, we propose to study the disposition of semigroup elements among the various groups acting as coordinate spaces of M, i.e., occurring as factors in the wreath product decomposition of $S(M)$. Indeed the identification of even a few of these groups should provide insight and possibly answers to a number of questions that bear on complex aspects of integrated metabolic organization. Such questions as:

Why under certain conditions is metabolism shifted from one major pathway to another?

What are probable alternative pathways within a given system and why are some possible ones not found?

are questions that we feel a physical theory for metabolism will greatly elucidate and possibly settle.

In conclusion, we mention the directions in which our further work will tend. Within the bourne of the present model we wish to carry the analysis of the semigroup associated with the state space (illustrated in Fig. 2) to the extent of discovering some of the simple non-abelian groups (SNAG's) which occur in its wreath product decomposition. These SNAG's, we feel, will have particular biological significance; indeed we believe that they may characterize the aspects of metabolism which are evolutionarily stable. We might mention here that we are developing computer programs to isolate SNAG's from any process that can be described in a format such as that of Figure 2. Furthermore,

* In general there will be more than one sequence of inputs that will do this, hence there is really a class of experiments corresponding to a given metabolic path.

mechanization of large portions of the formal algebraic manipulations associated with the notion of a theory embodied in Principle II is under way. This is being done as an example of man-machine interaction on the pilot version of the SDS-940 time sharing system which has been developed at Berkeley in the course of an ARPA sponsored project. The methods of this phase of the research will be published along with the results as soon as they are obtained. We would also like to study other models of metabolism, as, for example models that take explicitly into account the fact that inputs (especially coenzymes) are themselves metabolites. We are particularly interested in variations on our state-input assigment schemes that will make the resulting model especially sensitive in the area of cellular regulation and control, for it was from such investigations outside the traditional realm of biology that the theorems arose which underlie our present methods.

REFERENCES

1. Krohn, K., and Rhodes, J., "Algebraic theory of machines. I. Prime decomposition theorem for finite semigroups and machines." *Transactions of the American Mathematical Society*, Vol. **116**, Issue 4, p. 450 (April 1965).
2. Greenberg, D. M., "Amino acid metabolism." *Annual Review of Biochemistry*. **33**, 633 (1964).
3. Greenberg, D. M., (ed.), "Metabolic pathways, Vol. I," *Academic Press*, New York (1960).
4. Holzer, H., "Carbohydrate Metabolism," *Annual Review of Biochemistry*, **28**, 171 (1959).
5. Karlson, P., "Introduction to modern biochemistry." *Academic Press*, New York (1956).
6. Meister, Anton, "Biochemistry of the amino acids," especially Part IV, "Intermediary metabolism of the amino acids," pp. 256–393, *Academic Press*, New York (1957).
7. Umbreit, W. W., "Metabolic maps, Vol. I," *Burgess Publishing Company*, Minneapolis (1952).
8. Umbreit, W. W., "Metabolic maps, Vol. II," *Burgess Publishing Company*, Minneapolis (1960).

PETER B. BRIGHT

Committee on Mathematical Biology, University of Chicago,

Chicago, Illinois*

Relations Between System and Component Behavior

In designing a model of the neuron, the engineer wants those properties which will give the results desired but not those properties which will only add unnecessary cost. The most obvious property is that of time dependency. The system will be able to detect temporal patterns only if the output function of the component has a time delay or time dependency on the input. In the biological organism, individual neurons may detect temporal on-off or off-on patterns. Many comparisons, such as A is larger, louder, brighter, higher, etc. than B, could be expressed by on-off, off-on elements when attention is shifted from A to B. The basic learning cue is time dependent since conditioning of an input occurs to outputs which follow or concur but not to those which precede. Since the McCulloch-Pitt's model operates on a time delay and Rashevsky's model (see Eq. 10) accounts for the on-off and off-on property by two differential equations relating the rate of change of output to the input, it seems to be more important that the output be time dependent than the particular manner in which it is.

One of the more troublesome properties for theoretical analysis is that of threshold. To see what will be gained by including this property rather than omitting it as has been proposed by some, it will be convenient to employ the operation "$\dot{-}$" introduced by Kleene defined as

$$a \dot{-} b = \begin{cases} a - b & \text{if} \quad b < a \\ o & \text{if} \quad a \leqq b \end{cases} \tag{1}$$

This operation is referred to as proper subtraction by Davis and as diminish by Murphy but also concisely describes the threshold operation.

* Present address: Center for Theoretical Biology, State University of New York, Buffalo, New York.

The McCulloch-Pitt's model is based on a somewhat different operator $f_h(x)$. However, an identity

$$f_h(x) = 1 \div (h \div x) = \begin{cases} 1 \text{ if } h \leqq x \\ 0 \text{ if } x < h \end{cases} \quad \text{for } h, x \text{ integers} \tag{2}$$

given by Murphy and the obvious identity

$$a \div b = f_1(a - b) + f_2(a - b) + \ldots \tag{3}$$

show that the two operations must yield the same set of functions over any finite range of integers when they are introduced to a group structure involving $+$ and $-$.

The relations

$$|a - b| = (a \div b) + (b \div a) \tag{4}$$

and

$$(a \div b) = \tfrac{1}{2}(a - b + |a - b|) \tag{5}$$

given by Kleene and by Seeber (see paper by Murphy) show that the threshold and absolute value operations also yield the same set of functions when either is added to the operations of $+$ and $-$. The absolute value operation is introduced into the discussion of geometry in order to discuss the concept of distance which can not be expressed by linear operations.

Murphy also showed that the min and max or cap and cup operations could be expressed by $\div$ using the identities

$$x \cup y = x + (y \div x); \quad x \cap y = x \div (x \div y) \tag{6}$$

These identities coupled with

$$x \div y = (x - y) \cup 0 \tag{7}$$

show that both the threshold and the min and max (cap and cup) operations similarly yield the same set of functions. Thus the decision on whether to include the threshold among the properties of the component is related to the decision on whether or not the system of components should be able to express the concept of distance or to find the min and max.

The capacity of a system to demonstrate classical Pavlovian conditioning is also related to the properties of the component. The conditioning reflects the correlation of activity in two different neurons. If either neuron fires without the other, no conditioning takes place.

Among the numerous hypotheses regarding memory recording, only those incorporating Hebb's suggestion that a reinforcement takes place only when both the pre- and postsynaptic neurons are active are capable of demonstrating classical Pavlovian conditioning with a network of three neurons. Other hypotheses such as Hyden's RNA hypothesis do not depend on concurrent activity in two neurons and hence this aspect would have to be incorporated in a nonchanging supporting network that would probably have to be genetically determined. It is interesting to note that this gives a means for distinguishing between these two classes of hypotheses. If the anatomist could find an isolated Y shaped network of three neurons in which only one of two neurons can initially elicit but both can potentially elicit a response in the third neuron, then a crucial test could be made to distinguish these two classes of reinforcement hypotheses. Such a network might more easily be formed in tissue culture. If the second neuron could be conditioned to elicit a response in the third if and only if both pre- and postsynaptic neurons were firing, then the correlative process can take place in a synapse as suggested by Hebb. Failure to bring about conditioning would suggest that the greater number of neurons in the supporting network would be required to detect and record the correlation of activity.

By considering the class of all possible networks which can be generated by a component and by calling two networks equivalent if they have, or essentially have the same input-output relation, one can build a lattice hierarchy among components. If the class of all possible networks C_A which can be formed from component A contains a subset which is equivalent to the class of all possible networks C_B which can be formed from component B, then one may say that $B \leqq A$. If in addition there exists a network of A components for which there is no equivalent network of B components then, $B < A$. The basic technique for demonstrating the inequality is to show that there exists a network of A components which will give the same input-output relation as any particular B component. An exact comparison may be obtained only if the two components have the same input and output range. A more useful scheme may be constructed by calling two sets of networks essentially equivalent if any input-output relation of one set can be matched to within any degree of accuracy desired.

To demonstrate let us consider the McCulloch-Pitt's model, the linear threshold model and the Rashevsky two factor model. The output $E_k(t)$

of the kth McCulloch-Pitt's component at time t is

$$E_k(t) = f_h\left(\sum_i c_{ik}E_i(t-1)\right) \tag{8}$$

where the c_{ik}'s are constant integral weighting factors, h is the threshold number of inputs and f_h was defined in equation 2. Incorporating a time delay into the linear threshold model it is described as

$$E_k(t) = \sum_i c_{ik}E_i(t-1) \doteq h \tag{9}$$

For the Rashevsky neuroelement there are two intermediates e and j which are related to the inputs by the equations

$$de_{ik}(t)/dt = A_{ik}E_i(t) - a_{ik}e_{ik}(t)$$

and

$$dj_{ik}(t)/dt = B_{ik}E_i(t) - b_{ik}j_{ik}(t) \tag{10}$$

Rashevsky discussed three functional forms by which the output might be related to the $\sum_i [e_{ik}(t) - j_{ik}(t)]$. One was the linear threshold function which was looked on as an approximation to which the other two forms reduced for small inputs. The other two forms saturated at high inputs so that the output remains bounded despite how high the input might be.

For a constant input one has asymptotically

$$e_{ik} - j_{ik} = (A_{ik}/a_{ik} - B_{ik}/b_{ik})\, E_i(t) \tag{11}$$

and so by setting A_{ik} and B_{ik} sufficiently large while maintaining the ratios in equation 11, the Rashevsky model becomes a linear threshold device. Hence the linear threshold model $\leqq$ Rashevsky two factor model.

If one restricts to integers the range of the inputs, the weighting factors, and hence the outputs, then one may compare the linear threshold model with the McCulloch-Pitt's model. If initial activity is permitted such that a loop may exist and give a constant output, then one may replace a McCulloch-Pitt's component by a network of linear threshold components arranged to express equation 2 as in Figure 1. Thus the McCulloch-Pitt's component $\leqq$ linear threshold component.

If the range of inputs and outputs is bounded then most any output function can be represented as closely as desired by a sum of sufficiently small units. Any such function could then be represented by a network

of McCulloch-Pitt's components. Thus, in particular, Rashevsky's component $\leqq$ McCulloch-Pitt's component and the three are equivalent.

Since the McCulloch-Pitt's, linear threshold and Rashevsky's components are essentially equivalent, the selection of the one to use should depend on other considerations such as cost and convenience. The great extent of the analysis of the Rashevsky component by himself, Householder and Landahl and others, and the seeming ease of construction suggest that it should be particularly attractive.

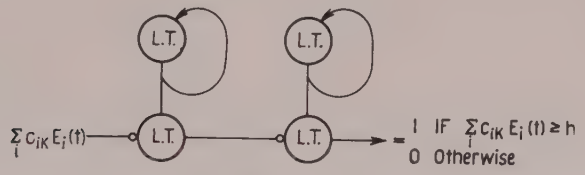


Fig. 1. A Network of Linear Threshold Elements LT which could Replace a McCulloch-Pitt's Element.

In contrast to the equality of the above components the linear component cannot contain the linear threshold component, since linear operations cannot express a nonlinear function. On the other hand, one may use linear threshold devices to give a linear output, as in Figure 2, by giving the same input pattern to two components but with excitation

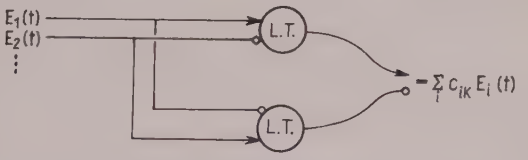


Fig. 2. A Network of Linear Threshold Elements LT which could Replace a Completely Linear Output.

and inhibition interchanged in one of them. Connecting the two components to a third, one as an excitation and the other as inhibition, there results the same affect as a linear function. Thus the linear component $<$ linear threshold component.

REFERENCES

Davis, M. (1958). *Computability and Unsolvability.* New York, McGraw Hill.

Householder, H. S., and Landahl, H. D. (1945). *Mathematical Biophysics of the Central Nervous System.* Bloomington, The Principia Press.

Hyden, H. (1960). "The Neuron" in *The Cell.* Ed. Jean Brachet and Alfred E. Mirsky. pp. 215–323. New York and London, Academic Press.

Kleene, S. C. (1952). *Introduction to Metamathematics.* New York, Van Nostrand.

Murphy, R. W. (1957). "A Positive Integer Arithmetic for Data Processing." IBM Journal of Research and Development. **1,** pp. 158–170.

Rashevsky, N. (1960). *Mathematical Biophysics; Physico-Mathematical Foundations of Biology.* New York, Dover Publications.

JAMES R. GOUGE, JR.

*Bird Engineering-Research Associates, Inc.**

Vienna, Virginia

Reliability Prediction for Networks of Probability State Variable Devices

INTRODUCTION

Recently, bionics systems using networks of probability state variable devices have been applied in areas of the control field which previously have been the exclusive domain of deterministic systems. These applications have spurred an interest in the reliability characteristics of these networks and the self-organizing systems in which they have been utilized. No small part of this interest has centered on the development of a reliability prediction method which would provide an analytic means of quantitatively assessing network and system reliability.

This paper presents a reliability prediction method which is an extension of classical reliability prediction techniques into the area of probabilistic networks. This method was developed by the author under an Air Force sponsored investigation under subcontract to Adaptronics, Inc., McLean, Virginia. Although it was developed specifically for the Adaptronics MK III Self-Organizing Controller (SOC), it should find general application in the field of incremental probability state variable controllers.

AN ELEMENTARY SELF-ORGANIZING CONTROL SYSTEM

A self-organizing control system employing the probability state variable (PSV) principle is shown in Figure 1 in its most elementary form. The controller portion of this system includes two elements—a performance assessment (PA) module, and a PSV module. In essence, the PA module evaluates system performance on the basis of the error signal (e) and informs the PSV module through the binary reward/punish-

* Now, Adaptronics, Inc., McLean, Va.

ment ($\overline{rp}$) signal of the results of this evaluation. The $\overline{rp}$ signal may be thought of as a "good/bad" comment by the PA module on the last incremental change (Δu) made by the PSV module in its output signal (u). The PSV module performs a self-evaluation by comparing the sense of its most recent Δu with $\overline{rp}$. As a result of this self-evaluation, it either incrementally increases or decreases (as appropriate) the probability that the next Δu will have the same sense as the last one. Through this iterative process of system performance evaluation and self-evaluation, the SOC controls the plant in a manner which is ideally suited to the plant dynamic characteristics.

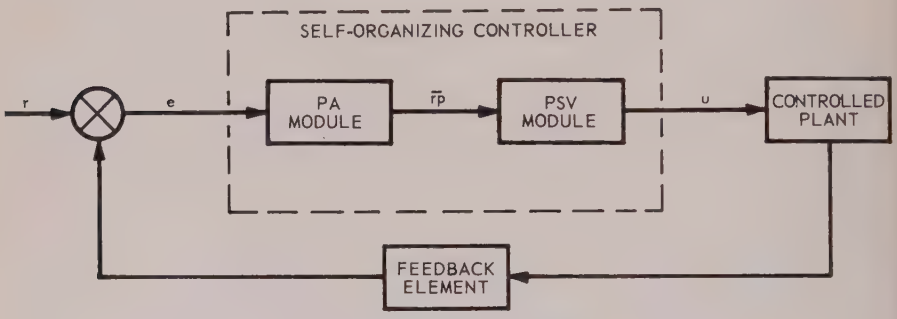


Fig. 1. Diagram of an Elementary Self-Organizing Control System.

In the MK III SOC, the output of the PSV module is an incremental analog signal which at a given moment may be at any one of 15 discrete levels. These levels are zero, and seven equal valued voltage steps to either side of zero—i.e., seven positive and seven negative levels. The incrementing rate of u is controlled by an internal clock and the magnitude of each increment is one voltage step (except at the two extremes where it may be zero). Although u is not truly a continuous analog signal, the integrating effect of a real plant on this incremental signal will for all practical purposes reduce it to a continuum.

RELIABILITY PREDICTION FOR THE ELEMENTARY PSV SOC

A knowledge of classical reliability analysis and prediction techniques is not essential to an understanding of the concepts presented in this paper and therefore will not be presented in any detail here. There are

a number of texts related to these subjects, and the reader is referred to them in the event of any personal interest.

Carrying out a reliability prediction on a given system involves the following procedural steps:

(a) Identifying, through the application of routine circuit analysis procedures, all possible failure modes of each system subassembly or module.

(b) Determining, through the application of classical reliability prediction procedures, the probability of occurrence of each of these failure modes during the operating period of interest.

(c) Determining, through the application of routine analytical procedures, the effect at the system level of the occurrence of each subassembly failure mode identified in (a).

(d) Determining, from the failure mode and effects analysis of (c) and the predicted probabilities of (b), the probability of occurrence of each system failure mode for the operating period of interest.

An analysis of the MK III PA module revealed the existence of only a single major failure mode—that of an $\overline{rp}$ output signal improperly related to its input. The occurrence of this *PA* module failure mode would result in a catastrophic (as opposed to a degradation) system failure.

The predicted reliability of the *PA* module for a two-hour airborne mission* was determined to be $R_{PA} = 0.999870$. Conversely, it may be stated that the predicted probability of failure of the *PA* module (under the same conditions) is $Q_{PA} = 130 \times 10^{-6}$—i.e., a total of 130 *PA* module failures would be expected in the course of one million two-hour airborne missions.

An analysis of the MK III PSV module revealed the existence of three major failure modes, defined as:

(a) *Hardover output*—a fixed or time-varying output signal restricted by the nature of the failure to within three incremental steps of either output limit.

(b) *Fixed output*—essentially a fixed (limited to a one-step variation) output signal at an unpredictable level other than those of (a) above.

* Since the previously referenced Air Force study related to the use of the MK III SOC as an adaptive flight controller, all numerical results presented in this paper relate to a two-hour airborne mission. The part failure rates employed in the analyses were those of MIL-HDBK-217 for operation at 55 °C, modified by the multiplier of 6.5 as specified for airborne systems in MIL-STD-756.

(c) *Erratic output*—a time-varying output signal which for one-fourth of the time or more is not properly related to the system response, or which is not probabilistic in nature. The output signal increments may be larger than the single step of a properly functioning module.

The predicted reliability of the MK III PSV module for a two-hour airborne mission was determined to be $R_{PSV} = 0.999657$. Conversely, it may be stated that the predicted probability of failure of the PSV module (under the same conditions) is $Q_{PSV} = 343 \times 10^{-6}$.

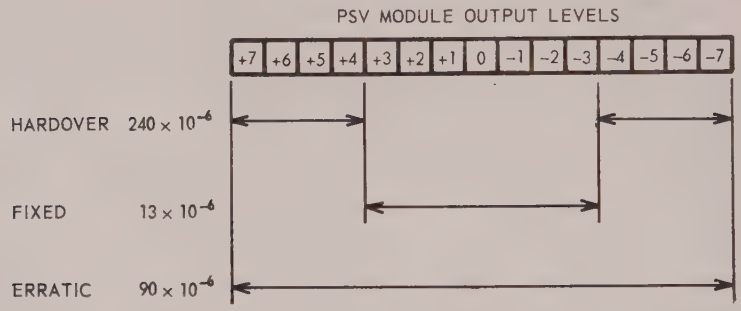
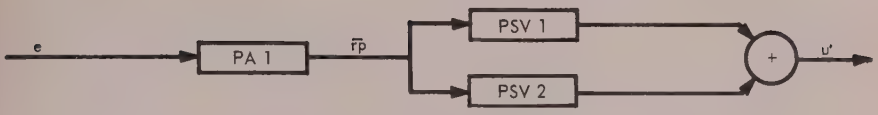


Fig. 2. Relationship Between PSV Module Output Levels and Major PSV Failure Modes; and the Probabilities of Occurrence for a Two-Hour Airborne Mission

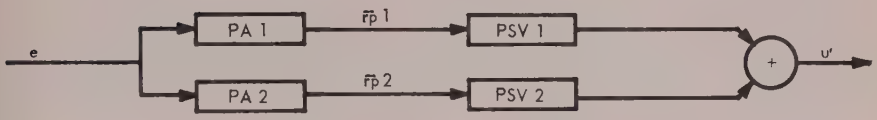
Based on the foregoing analyses, the predicted reliability for a two-hour airborne mission of a controller composed of a single PA module and a single PSV module is:

$$R_{SOC} = R_{PA}R_{PSV} = 0.999870 \times 0.999657 = 0.999527$$

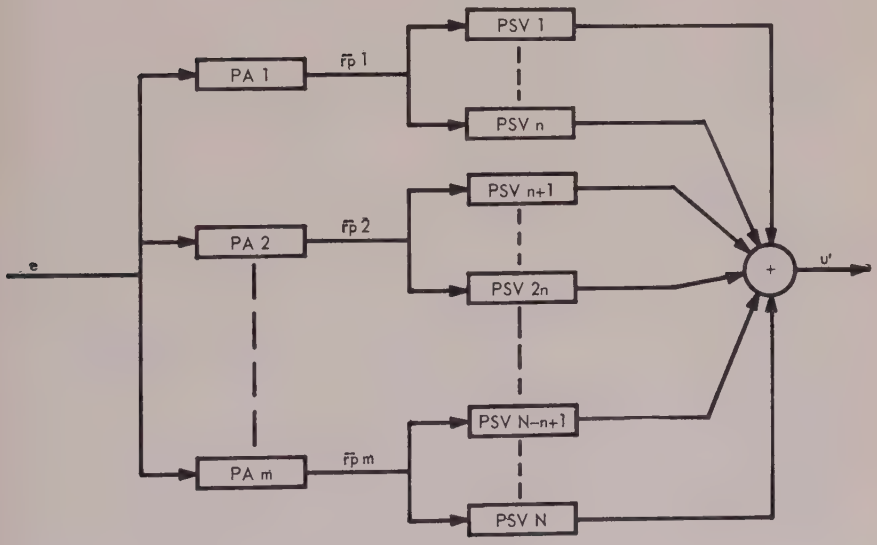
On the basis of this prediction, 473 failures would be expected in each million two-hour flights. For many controller applications, such a failure probability would be acceptable, but for flight controller applications it falls far short of the goal proposed by some for a single-axis controller of the equivalent mean-time-between-failures (MTBF) of 50,000 hours based on a two-hour flight. This figure translates to a single-flight reliability of 0.999960 or, conversely, a failure probability of 40×10^{-6}. (This translation assumes a preflight test verification that all elements of the controller are fully operational prior to the initiation of each mission.)



(a.) One PA Module, Two PSV Modules



(b.) Two PA Modules, Two PSV Modules



(c.) m PA Modules, N PSV Modules

Fig. 3. Examples of PA/PSV Networks Considered for Reliability Analysis

This disparity between the proposed reliability goal for a single-axis flight controller and the predicted reliability for a single-PA, single-PSV controller led to an investigation of the reliability of controllers composed of networks of PA and PSV modules. The first step was to determine the network configurations which it appeared practical to implement with the existing modules, these then being considered the configurations worthy of immediate investigation. For various reasons, it was decided to limit the effort to networks where the number of PSV modules is an integral multiple of the number of PA modules, with each PA module supplying the $\overline{rp}$ signal to its proportionate share of PSV modules. The outputs of all PSV modules were in turn summed to provide the controller output signal. Examples of PA/PSV networks considered in the investigation are shown in Figure 3.

RELIABILITY PREDICTIONS FOR PARALLEL PSV MODULES

Early in the investigation of reliability prediction techniques for *PA/PSV* networks, it became apparent that there was a need for developing a means for predicting the reliability of parallel PSV modules (exclusive of PA modules), and that the three failure modes previously presented for PSV modules were too restrictive to allow a comprehensive analysis of parallel PSV's. To facilitate the analysis, the three PSV

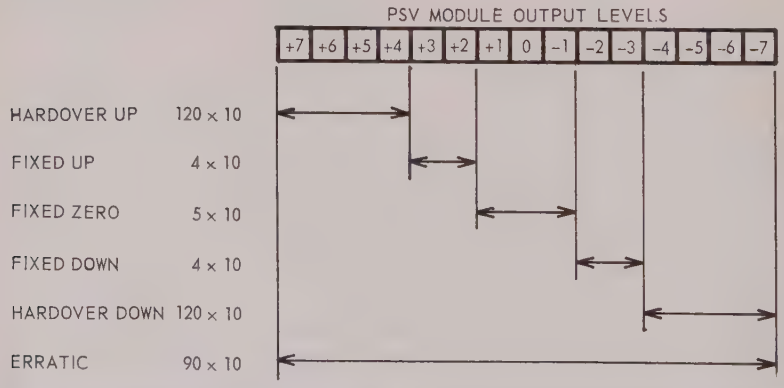


Fig. 4. Relationship Between PSV Module Output Levels and PSV Failure Modes; and the Probabilities of Occurrence for a Two-Hour Airborne Mission

failure modes were further subdivided into six failure modes, as defined below and shown in Figure 4:

(a) *Hardover up output*—a fixed or time-varying output signal restricted by the nature of the failure to within three incremental steps of the positive output limit.

(b) *Hardover down output*—a fixed or time-varying output signal restricted by the nature of the failure to within three incremental steps of the negative output limit.

(c) *Fixed zero output*—essentially a fixed output signal restricted by the nature of the failure to the zero and first positive and negative output steps.

(d) *Erratic output*—a time-varying output signal which for one-fourth of the time or more is not properly related to the system response, or which is not probabilistic in nature. The output signal increments may be larger than the single step of a properly functioning module.

(e) *Fixed up output*—essentially a fixed output signal restricted by the nature of the failure to the second and third positive output steps.

(f) *Fixed down output*—essentially a fixed output signal restricted by the nature of the failure to the second and third negative output steps.

The failure probabilities presented in Figure 4 resulted from an extension of the intra-PSV module failure mode and effects analysis previously performed. The up and down senses were arbitrarily assigned to positive and negative PSV outputs, respectively, for ease of discussion.

In addition to expanding the number of module failure modes to be considered, it became necessary to establish certain "ground rules" consistent with the functional characteristics of the PSV module, the nature of its failure modes, and the simple summing process by which the combined output signal (u') from parallel connected PSV's is obtained. Essentially, these rules involve the assignment of weighting factors to each of the six failure modes on the basis of their influence on u' and establishing a definition (utilizing these weighting factors) for the proportion of normal full-scale output which can be realized from a given number (n) of PSV modules connected in parallel. Normal full-scale output is defined as the absolute value of that value of u' which would obtain if all PSV's were functioning normally and the output of each were at the same extreme output level, either positive or negative.

With n PSV modules connected in parallel with their outputs combined through simple summation, the weighting factors assigned to the six module failure modes are:

(a) Hardover up output $+1/n = W_a$
(b) Hardover down output $-1/n = W_b$
(c) Fixed zero output $0 = W_c$
(d) Erratic output $0 = W_d$
(e) Fixed up output $+1/2n = W_e$
(f) Fixed down output $-1/2n = W_f$

These weighting factors were derived using the information of Figure 4. A hardover output failure (either up or down) is seen to represent approximately full-scale output for a single PSV module—thus its assigned weighting factor of $\pm 1/n$ (direction dependent) indicating that the full weight of the output of the failed module is felt in u', which is considered to have a normal full-scale value of unity. A fixed up or down output failure is seen from Figure 4 to represent approximately one-half of full-scale output for a single PSV module—thus their assigned weighting factors of $\pm 1/2n$, respectively. A fixed zero output failure is seen to represent approximately zero output for a single PSV module—thus its assigned weighting factor of zero. In the case of an erratic output failure, the output of a failed PSV will typically be a randomly time-varying signal with an average value (for even short periods of time) of zero. Thus an erratic output failure was considered in the same class as a fixed zero output failure and assigned a weighting factor of zero.

For purposes of this analysis, the proportion (k) of the normal full-scale output which can be realized from a number (n) of parallel connected PSV modules was defined as the minimum value of the absolute value of:

$$\begin{aligned}
&[\text{Number of Hardover Up Failures } (a)]\,[W_a] \\
&+ [\text{Number of Hardover Down Failures } (b)]\,[W_b] \\
&+ [\text{Number of Fixed Zero Failures } (c)]\,[W_c] \\
&+ [\text{Number of Erratic Failures } (d)]\,[W_d] \\
&+ [\text{Number of Fixed Up Failures } (e)]\,[W_e] \\
&+ [\text{Number of Fixed Down Failures } (f)]\,[W_f] \\
&\pm \left[1 - \frac{\text{Number of Failed PSV's } (q) = (a + b + c + d + e + f)}{\text{Number of Parallel PSV's } (n)}\right]
\end{aligned}$$

or, using the above symbols and introducing the actual weighting factors (W_c and W_d being defined as zero):

$$k = \min \left| \frac{a}{n} - \frac{b}{n} + \frac{e}{2n} - \frac{f}{2n} \pm \left(1 - \frac{q}{n}\right) \right| \tag{1}$$

Implied in this definition are the following assumptions which are consistent with capabilities of parallel connected PSV modules and the nature of their failure modes:

(a) A module failed in the hardover up state can be "balanced out" by a module failed in the hardover down state, by two modules failed in the fixed down state, or by a single nonfailed module. (A similar assumption applies to "balancing out" a hardover down failure.)

(b) A module failed in the hardover up state can be "balanced out" by a module failed in the fixed down state plus one-half the capacity of a nonfailed module. (A similar assumption applies to "balancing out" a hardover down failure.)

(c) A module failed in the fixed zero state requires no "balancing out," but simply results in reducing the realizable output of the parallel connected PSV's by $1/n$th.

(d) A module failed in the erratic output state requires no "balancing out," but simply results in reducing the realizable output of the parallel connected PSV's by $1/n$th.

(e) A module failed in the fixed up state can be "balanced out" by a module failed in the fixed down state, or by one-half the capacity of a nonfailed module. (A similar assumption applies to "balancing out" a fixed down failure.)

As an example of the use of the foregoing definition of k, consider six PSV's connected in parallel. If all modules are functioning properly, then:

$$k = \min \left| \frac{0}{n} - \frac{0}{n} + \frac{0}{n} - \frac{0}{n} \pm \left(1 - \frac{0}{n}\right) \right| = 1$$

Thus, full-scale output will be realized.

However, if one hardover up failure ($a = 1$), one fixed zero failure ($c = 1$), and one fixed down failure ($f = 1$) were experienced, then the proportion of normal full-scale output which would be realized

would be:

$$k = \min \left| \frac{a}{n} - \frac{f}{2n} \pm \left(1 - \frac{q}{n}\right) \right|$$

$$= \min \left| \frac{1}{6} - \frac{1}{(2)(6)} \pm \left(1 - \frac{3}{6}\right) \right|$$

$$= \min \left| \frac{1}{6} - \frac{1}{12} \pm \left(1 - \frac{1}{2}\right) \right|$$

$$= \min \left| \frac{1}{12} \pm \left(\frac{1}{2}\right) \right| = \frac{5}{12}$$

Thus, by the definition established for k, the group of six PSV's would be considered capable of fulfilling its requirements if 5/12 or less of full-scale output were required for satisfactory system operation.

The probability of failure of groups of parallel connected PSV modules was calculated, using the data of Figure 4 and the foregoing definitions and "ground rules." These calculations were carried out for an airborne mission of two hours duration and required proportions of normal full-scale output of 1, 2/3, 1/2, 1/3, and 1/100. The results are tabulated in Table 1. The columns for $k = 1$ and $k = 1/100$ are shown to indicate limiting values, and it should not be inferred that a value of $k = 1$ may be required for a typical system or that $k = 1/100$ will yield satisfactory results for any system.

Table 1. Probability of Failure ($\times 10^6$) of Parallel Connected PSV Modules for a Two-Hour Airborne Mission

Number of PSV's	Proportion of Normal Full-Scale Output Required				
	1	2/3	1/2	1/3	1/100
1	343	343	343	343	343
2	686	686	496	496	480
3	1029	744	720	0.24	0.23
4	1371	992	0.47	0.46	0.17
5	1714	1200	0.76	0.29	<0.001
6	2056	1.20	0.30	<0.001	<0.001
7	2399	1.67	0.60	<0.001	<0.001
8	2741	2.16	0.001	<0.001	<0.001
9	3083	1.11	0.001	<0.001	<0.001

The failure probabilities of Table 1 indicate that for required proportions of full-scale output of one-half or less, the number of PSV modules in parallel should be four or more in order to maximize the reliability of the network. It is also of interest to note that a single PSV will be more reliable than two parallel modules in all cases, and will be more reliable than three parallel modules in the event system characteristics dictate that one-half or more of normal full-scale output must be realized from the parallel modules.

Connecting PSV modules in parallel does not yield the reliability improvement that would be predicted from an assumption of simple operational redundancy. This is due to the fact that PSV failure modes are such that not only is the capability of the failed module lost, but its output will (except for the fixed zero and erratic output failure modes) actually have a degrading effect on the combined output signal, requiring some portion of the cabability of the remaining properly functioning modules to compensate for or "balance out" this degrading signal.

RELIABILITY PREDICTIONS FOR PARALLEL PA/PSV NETWORKS

Analysis showed that a PA module failure will result in a hardover failure of its associated PSV's (those to which it supplies the $\overline{rp}$ signal) of an unpredictable sense. The unpredictable sense of the hardover output is a result of the probabilistic nature of the operation of the PSV module and the fact that for many types of PA failures the polarity of the $\overline{rp}$ output signal of the failed module will vary with system conditions, hence with time. PA module failures, therefore, are not considered capable of being "balanced out" by other PA failures, but only by another fully operational parallel PA/PSV network arm. This finding leads to a definition for the proportion (k') of the normal full-scale output which can be realized from a system of parallel connected PA/PSV networks of:

$$k' = \frac{k(m - q_m) - q_m}{m} \tag{2}$$

where: k is found from equation (1) with n equal to the number of PSV's associated with nonfailed PA's
m equals the number of PA's in the system
q_m is the number of failed PA's in the system

With this definition, it is possible to arrive at calculated values for k' less than one; a negative value of k' has no significance other than it represents a useless system.

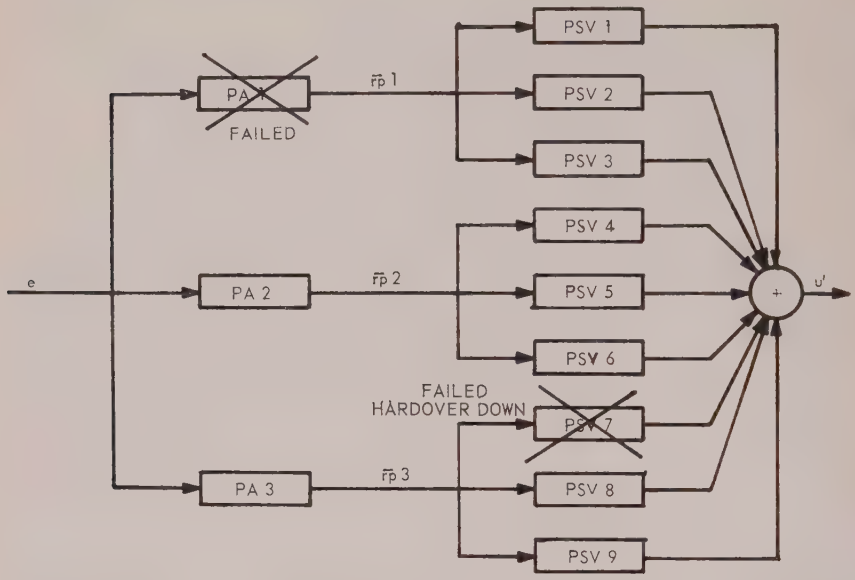


Fig. 5. 3 PA/9 PSV Network Configuration with One Failed PA Module and One PSV Associated with a Non-Failed PA Failed in the Hardover Down Mode

As an example of the use of the foregoing definition of k', consider a system of parallel PA/PSV networks with a total of three PA modules ($m = 3$) and nine PSV modules which experiences one PA module failure as shown in Figure 5. Then, in this event:

$$k' = \frac{\min \left| -\frac{b}{n} \pm \left(1 - \frac{q}{n}\right) \right| (m - q_m) - q_m}{m}$$

$$= \frac{\min \left| -\frac{1}{6} \pm \left(1 - \frac{1}{6}\right) \right| (3 - 1) - 1}{3} = \frac{\frac{4}{6}(2) - 1}{3} = \frac{1}{9}$$

Thus, by the definition established for k', the system of three PA's and nine PSV's having experienced the failures noted above would be considerd

capable of fulfilling its mission requirements if 1/9 or less of normal full-scale output were required for satisfactory operation.

The probability of failure of parallel connected networks of PA and PSV modules was calculated using the data of Table 1, the above definition for k', and the previously presented value of $R_{PA} = 0.999870$. These calculations were carried out for an airborne mission of two hours duration and required proportions of normal full-scale output of 1, 2/3, 1/2, 1/3, and 1/100. The results are tabulated in Table 2.

The data of Table 2 show that reliability of the PA/PSV systems considered is limited by their inability to accommodate PA module failures

Table 2: Probability of Failure ($\times 10^6$) of Parallel Connected PA/PSV Networks for a Two-Hour Airborne Mission

Number of		Proportion of Normal Full-Scale Output Required				
PA's	PSV's	1	2/3	1/2	1/3	1/100
1	1	473	473	473	473	473
1	2	816	816	626	626	610
1	3	1160	874	850	130	130
1	4	1500	1120	131	130	130
1	$R = 1$	130	130	130	130	130
2	2	946	946	756	756	480
2	4	1630	1250	260	260	260
2	6	2320	261	261	260	260
2	$R = 1$	260	260	260	260	260
3	3	1420	1130	1110	0.56	0.55
3	6	2450	391	390	0.59	0.44
3	9	3470	391	390	0.85	0.05
3	$R = 1$	390	390	390	0.05	0.06
4	4	1890	1510	1.11	1.10	0.65
4	8	3260	522	1.17	0.87	0.10
4	$R = 1$	520	520	0.10	0.10	0.10
5	5	2360	1850	1.82	1.10	0.17
5	$R = 1$	650	650	0.17	0.17	<0.001
6	6	2830	2.79	1.86	0.97	0.001
6	$R = 1$	780	0.25	0.25	<0.001	<0.001

in any case when only one or two PA's are used, and in the case of one-half or more output proportion requirements even when three PA's are used. Again, as in the case of parallel PSV modules, the procedure of connecting PA/PSV networks in parallel does not yield the reliability improvement that would be predicted from an assumption of simple redundancy due to the degrading effect on u' of module or network failures.

The data of Table 2 are presented as a guide for selecting the most appropriate system configuration of PA/PSV modules for an airborne mission of two hours duration when the required value of k' is known and other inputs have become available which will allow carrying out the required weight/space/power/cost/reliability tradeoff studies. At present, the required value of k' for satisfactory operation of an SOC in a particular application must be determined empirically. Preliminary results of tests carried out under computer simulation on a 6-PA/6-PSV flight controller designed for pitch axis control of a high-performance aircraft indicate that the required value of k' will fall somewhere between 1/3 and 2/3 (see Table 2) which will result in a predicted failure probability more than two orders of magnitude better than that suggested as a design goal.

The validity of the PA/PSV network reliability models presented has been verified to a limited (though promising) extent using the Adaptronics MK II (forerunner to the MK III) and MK III SOC's, and more comprehensive efforts are contemplated in this area in the near future.

Two important characteristics of the PSV type system which are not taken into account in the analyses and thus are not reflected in the data of Table 1 or Table 2 are the ability of the PSV system to compensate for degradation type failures of portions of the system other than the SOC, and the recovery time required for the adaptive system to compensate for module failures within itself. Both of these characteristics will play an important role in defining the operational utility of an adaptive control system in a particular application.

SUBJECTS FOR FUTURE INVESTIGATION

Additional empirical studies are required to further validate the reliability models presented in this paper, although the work performed in this area to date has shown promising results. Another subject of interest

to potential users of SOC's in whatever field is the development—through either theoretical or empirical means—of a method of predicting required values of k' for a given controlled plant. A number of means have been suggested for improving the reliability of adaptive controllers, such as the MK III, which utilize PA/PSV networks. One of the most promising of these is the possibility of employing voting redundancy at the $\overline{rp}$ signal level. Since the principal limiting factor on reliability is currently the PA module (see Table 2), this approach to SOC reliability improvement is particularly attractive.

APPENDIX: EXAMPLES OF CALCULATIONS OF PREDICTED FAILURE PROBABILITIES FOR PARALLEL *PSV* MODULES AND PARALLEL *PA/PSV* NETWORKS

SAMPLE CALCULATIONS FOR PARALLEL *PSV* MODULES

Examples of calculations are given below for determining the predicted failure probabilities of parallel PSV modules for a two-hour airborne mission.

1. Four Parallel *PSV* Modules With Required $k = 1$

A requirement of $k = 1$ dictates that no module failures can be tolerated. Hence, the probability of failure for four parallel modules is found by subtracting from one the probability that all PSV's will survive the mission without failure. With $Q_{4PSV,\,k=1}$ representing the probability of failure of 4 parallel PSV modules with a required $k = 1$:

$$\begin{aligned} Q_{4PSV,\,k=1} &= 1 - R_{PSV}^4 \\ &= 1 - (0.999657)^4 \\ &= 1 - 0.998629 \\ &= 1371 \times 10^{-6} \end{aligned}$$

2. Two Parallel *PSV* Modules With Required $k = 1/2$

If all combinations of failed and nonfailed modules which are possible with two PSV's are substituted in equation (1) of the paper it will be found that only three of these combinations will yield $k \geqq 1/2$, as shown below (with $\bar{q}$ representing the number of nonfailed PSV's).

a	b	c	d	e	f	$\bar{q}$	k	
0	0	0	0	0	0	2	1	R_{PSV}^2
1	0	0	0	0	0	1	0	
0	1	0	0	0	0	1	0	
0	0	1	0	0	0	1	1/2	$R_{PSV}p(c)$
0	0	0	1	0	0	1	1/2	$R_{PSV}p(d)$
0	0	0	0	1	0	1	1/4	
0	0	0	0	0	1	1	1/4	
1	1	0	0	0	0	0	0	
1	0	1	0	0	0	0	0	
1	0	0	1	0	0	0	0	
1	0	0	0	1	0	0	0	
1	0	0	0	0	1	0	0	
0	1	1	0	0	0	0	0	
0	1	0	1	0	0	0	0	
0	1	0	0	1	0	0	0	
0	1	0	0	0	1	0	0	
0	0	1	1	0	0	0	0	
0	0	1	0	1	0	0	0	
0	0	1	0	0	1	0	0	
0	0	0	1	1	0	0	0	
0	0	0	1	0	1	0	0	
0	0	0	0	1	1	0	0	
2	0	0	0	0	0	0	0	
0	2	0	0	0	0	0	0	
0	0	2	0	0	0	0	0	
0	0	0	2	0	0	0	0	
0	0	0	0	2	0	0	0	
0	0	0	0	0	2	0	0	

Hence:

$$\begin{aligned} Q_{2PSV,k=1/2} &= 1 - R_{2PSV,k=1/2} \\ &= 1 - [R_{PSV}^2 + 2R_{PSV}(p(c) + p(d))]^* \\ &= 1 - [(0.999657)^2 + 0.999657\,(5 + 90) \times 10^{-6}] \\ &= 1 - [0.9993141 + 0.0001899] \\ &= 1 - 0.999409 \\ &= 496 \times 10^{-6} \end{aligned}$$

SAMPLE CALCULATIONS FOR PARALLEL PA/PSV NETWORKS

Examples of calculations are given below for determining the predicted failure probabilities of parallel PA/PSV networks for a two-hour airborne mission.

1. One PA Module and Two PSV Modules With Required k′ = 1

A requirement of $k' = 1$ dictates that no module failures can be tolerated. Hence, the probability of failure for a one-PA and two-PSV network is found by subtracting from one the probability that all modules will survive the mission without failure. Therefore:

$$\begin{aligned} Q_{1PA/2PSV,k'=1} &= 1 - R_{PA}R_{PSV}^2 \\ &= 1 - (0.999870)\,(0.999657)^2 \\ &= 1 - 0.999184 \\ &= 816 \times 10^{-6} \end{aligned}$$

2. Three PA Modules and Three PSV Modules With Required k′ = 1/3

Use is made here of data from Table 1, specifically $Q_{3PSV,k=1/3}$. It should also be recognized that should a PA module fail, no failures can

* The term $R_{PSV}(p(c) + p(d))$ is multiplied by two because with two PSV's there are two possible combinations of one good PSV and one failed PSV, i.e., the first may fail and the second remain nonfailed or the first may remain nonfailed and the second may fail.

be tolerated in the PSV's associated with the nonfailed PA's. The probability of failure of a 3PA/3PSV network is then found to be:

$$
\begin{aligned}
Q_{3PA/3PSV,\,k'=1/3} &= 1 - [R_{PA}^3 R_{3PSV,\,k=1/3} + 3R_{PA}^2 Q_{PA} R_{PSV}^2] \\
&= 1 - [R_{PA}^3 1 - (Q_{3PSV,\,k=1/3}) + 3R_{PA}^2 Q_{PA} R_{PSV}^2] \\
&= 1 - [(0.999870)^3\,(0.24 \times 10^{-6}) \\
&\quad + 3(0.999870)^2\,(130 \times 10^{-6})\,(0.999657)^2] \\
&= 1 - [0.99960981 + 0.00038963] \\
&= 1 - 0.99999944 \\
&= 0.63 \times 10^{-6}
\end{aligned}
$$

V. V. GRIFFITH

Goodyear Aerospace Corporation

Akron, Ohio

An Iterated Element Theory of Neuron Networks

INTRODUCTION

Most present models of the nervous system assume that the neuron is a threshold logic element in which synaptic weights or thresholds are adjusted by higher centers.[1] Available physiological data do not completely support these assumptions. It is well known that pulse rates in the nervous system reflect stimulus intensities and that a neuron's response to stimuli is altered by the immediate past history of stimuli.[2] These characteristics are not displayed by threshold logic devices. No physiological mechanism has been discovered which shows definitely that higher centers can and do adjust synapses or thresholds in lower centers.

The following work continues previous work by the writer,[3] and starts from two premises: (a) The neuron handles quantized analog information (pulse rate modulated) and can be approximately modeled by continuous functions, (b) Each neuron is self-adjusting, and uses a rule of adjustment that causes the overall system (i.e., nervous system plus environment) performance to extremize a performance functional, rather than to conform to a pre-established system model. (A discussion of physiological data supporting these premises is given in reference 3.) Previous work extremized a static criterion; the present model extremizes an integral.

Following paragraphs use results of optimal control theory to arrive at an iterative control element. It is shown that this element has some characteristics resembling those displayed by neurons.

BASIC ANALYSIS

It is assumed that we wish to control a multi-input, multi-output process (an environment) and do not have complete knowledge of the process, although we may know some process parameters or their algebraic signs

or possible ranges of values. The controller is to operate in a hostile environment; portions of the controller may fail or be damaged. For reliable operation, we wish the controller to be made up of multiple adaptive elements which operate autonomously and with a minimum of direction and control from other units, so that if elements are damaged, other elements can "take over."

Let the outputs from the "environment" to be controlled be y_i, $i = 1, 2, \ldots, n$, and let the control responses be m_j, $j = 1, 2, \ldots, m$. We assume for the moment that the "environment" is linear, and described by a vector differential equation of the form

$$\dot{Y} = AY + BM + D \tag{1}$$

where Y, M, D, are respectively an n-vector, m-vector and an n-vector. A is an n by n, B an n by m matrix, and D is a disturbance vector. We wish to minimize system energy "cost" given by

$$C = \frac{1}{2}\int_0^T (Y^t Y + \mathbf{M}^t\mathbf{M})\, dt \tag{2}$$

where the superscript t represents the transpose of a matrix. T is some arbitrary time limit of integration. This cost criterion seems intuitively to compare with observed creature behavior, but was selected primarily because it is mathematically tractable.

It can be shown by use of variational methods[4] that a minimum of equation (2) subject to the constraint of equation (1) is obtained if an auxiliary set of variables p_i, $i = 1, 2, \ldots, n$, satisfies the equation

$$\dot{P} = -A^t P - Y \tag{3}$$

with the additional constraint that

$$M = -B^t P. \tag{4}$$

It can further be shown that if T is much larger than the longest time constant of the system, one can with negligible error set

$$P = KY \tag{5}$$

where K is a symmetric, n by n matrix of constants which is non-negative definite. Equations (3) and (5) only hold during times when no disturbance is present.

Routine algebraic manipulation of the above yields

$$Y - AY + BB^t P = D \tag{6}$$

$$KY + A^t P + Y = 0 \tag{7}$$

$$P = KY \tag{8}$$

These three equations can be used to define an adaptive system.

SYSTEM DESIGN

We have as a basic relation that $P = KY$, $M = -B^t P$. Such a basic, non-adaptive system is shown in Figure 1. Proper choice of the parameters k_{ij} and b_{jk} will produce optimum response. Note that the intermediate variables, p_i, are the weighted sums of the y_i, and that the outputs, m_i, are in turn weighted sums of the intermediate variables. One can, with little stretch of the imagination, see a similarity between a sensor-internuncial-effector reflex arc and the system of Figure 1. However, the system has no adaptive ability, which can be introduced as follows.

Let the k_{ij} be initially unknown. Instead, parameters $\varkappa_{ij}$, which are the best estimates available, are inserted into the network. For simplicity, assume temporarily that A is zero. We have that

$$P = \mathbf{K}Y; \quad \mathbf{K} = [\varkappa_{ij}]. \tag{9}$$

One obtains from equation (7) for $A = 0$ that

$$\mathbf{K}Y + Y = \varepsilon_1, \tag{10}$$

where ε_1 is an error signal.

The error signal can be used to revise the values of the $\varkappa_{ij}$. One such method is the "method of steepest descent" in which one sets

$$\dot{\varkappa}_{ij} = -\frac{\partial \varepsilon_1}{\partial \varkappa_{ij}} \operatorname{sgn} [\varepsilon_1]_i = -y_j \operatorname{sgn} [\varepsilon_1]_i \tag{11}$$

Figure 2 illustrates the method. Note that each element uses only information found in its own inputs to adjust its input weights.

It was assumed above that B was known. Assume now that, like K, B is only approximately known, and that estimated β_{ij} are originally

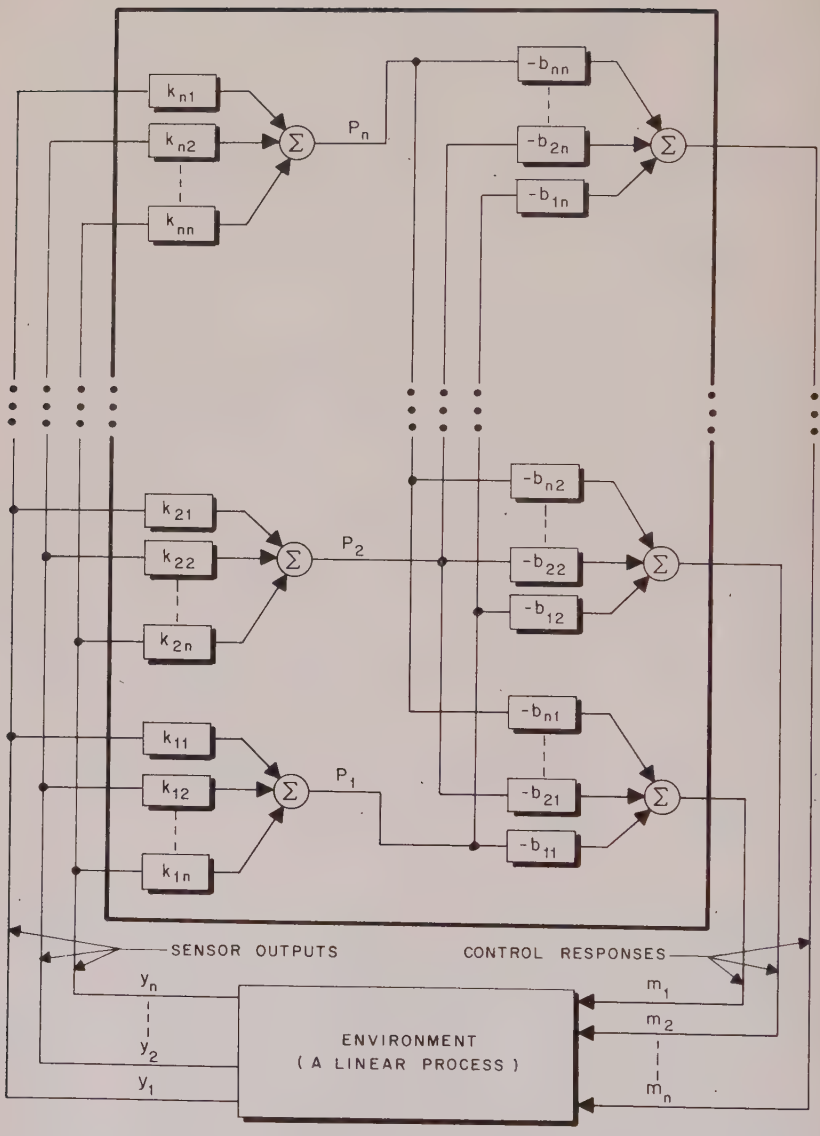


Fig. 1. Basic System.

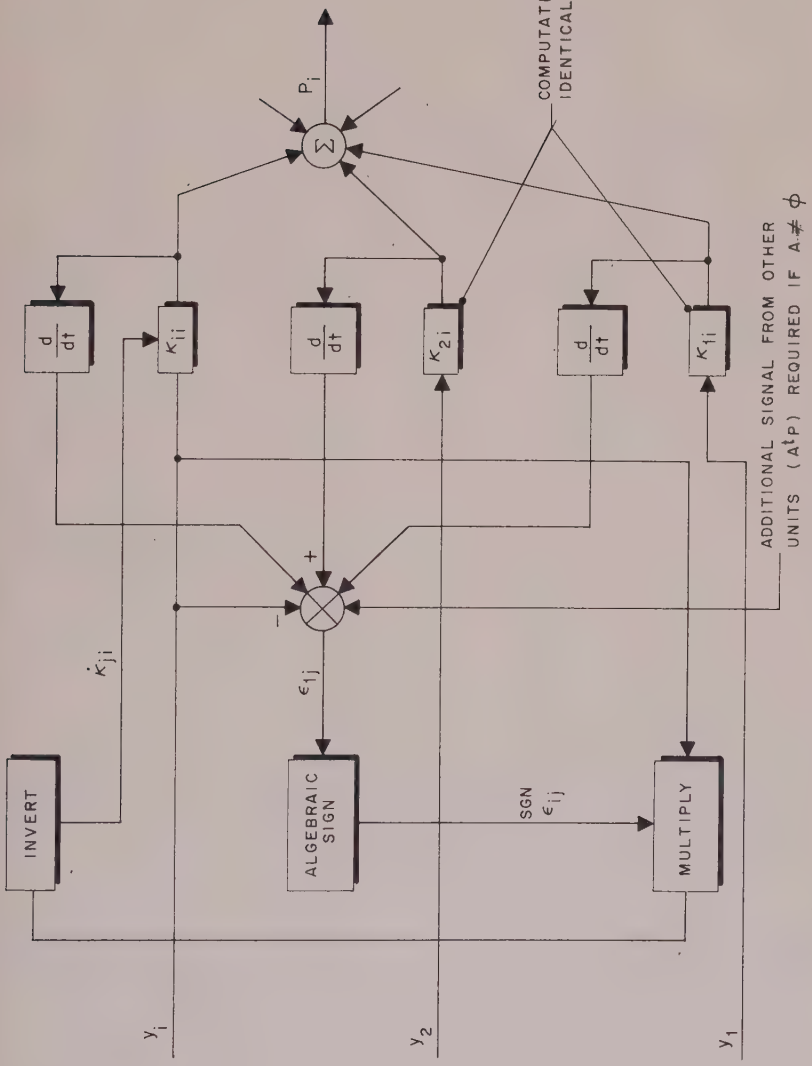


Fig. 2. κ-Computation.

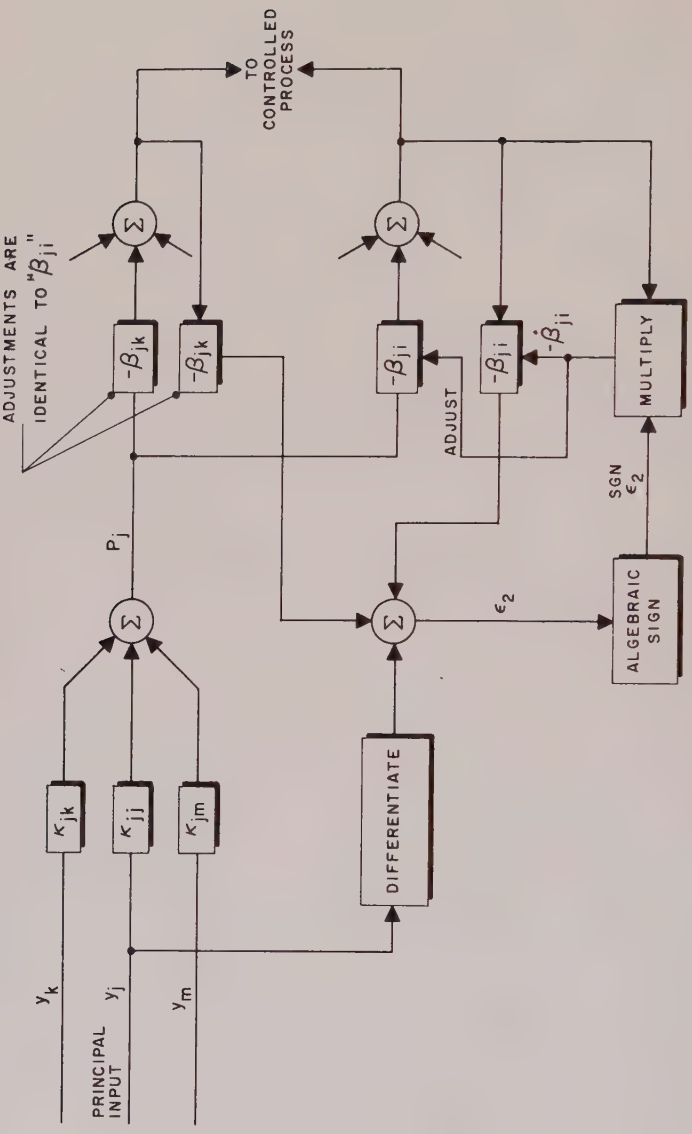


Fig. 3. β-Computation.

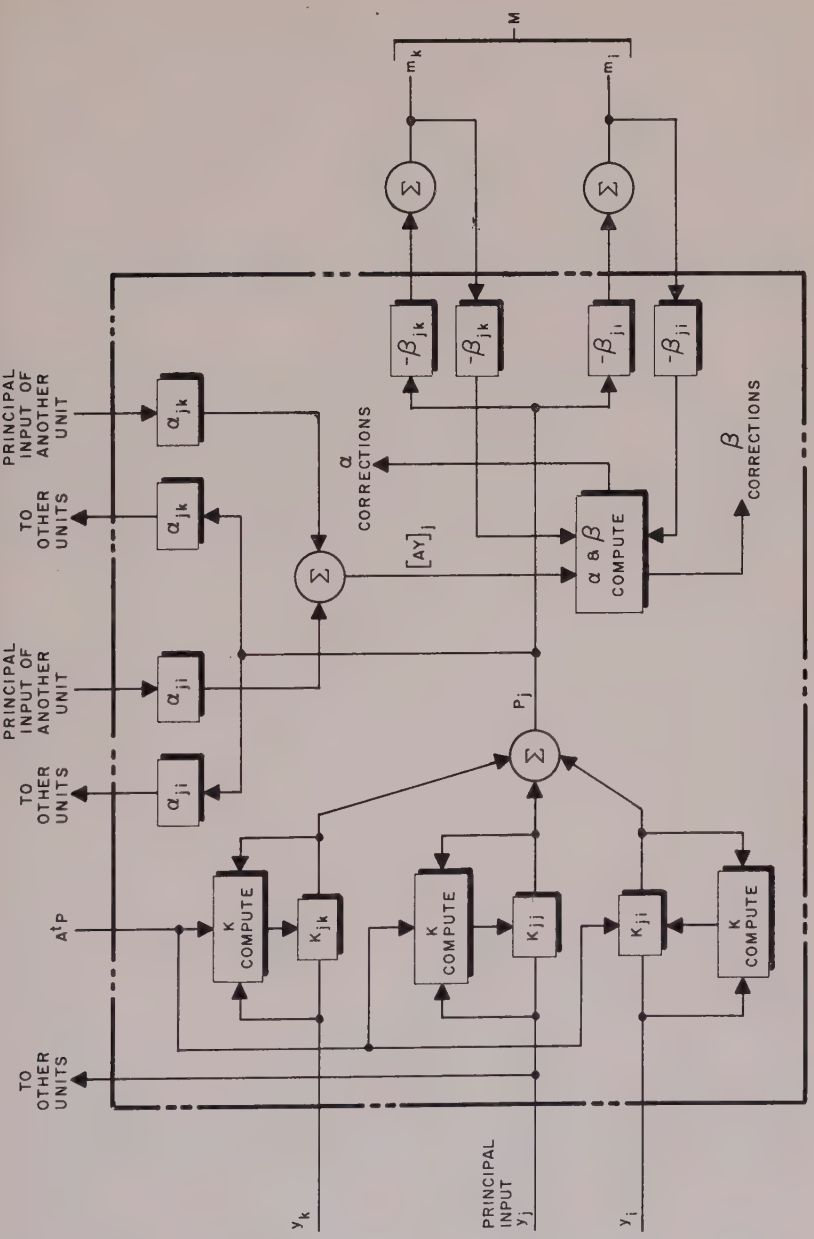


Fig. 4. A Complete Unit.

inserted. The equation

$$Y - BM = D \tag{12}$$

becomes

$$Y - \mathbf{B}M = \varepsilon_2;\ \mathbf{B} = [\beta_{ij}] \tag{13}$$

when the disturbance is zero.

The values β_{ij} can be adjusted by a method similar to that used for adjusting K. Figure 3 shows the method. The output signals m_i are multiplied by the same coefficients β_{ij} that have been computed as weights on the output of the element. The β_{ij} are altered by equations of the form

$$\dot{\beta}_{ij} = -m_j \operatorname{sgn} [\varepsilon_2]_i. \tag{14}$$

Again, each element uses for computation only signals which are directly accessible to it.

From the description of the simplified system it is now possible to see how an element to handle a more general case would be constructed. A block diagram is shown on Figure 4. The parameters α_{ij} are adjusted in a method analogous to that already presented for the β_{ij}, using the equations (6), (7) and (8) in their complete form ($A \neq 0$).

The above has described only one method for adjusting the parameters. Others, perhaps perturbations or correlation computations, might be suitable for some sort of iterative computation, depending on the exact system to be controlled and the statistics of the disturbances expected. A previous version used a correlation computation.

REDUNDANCY

Suppose now that two elements are inserted in parallel someplace in the controller. Each of the elements receives all the appropriate inputs and each sends outputs to all the places served by the other. Elementary calculation shows that asymptotic system performance will be unchanged if the one principal input to both of the parallel elements is multiplied by a factor $\sqrt{2}/2$. All of the weights on the inputs to the parallel pair and the outputs and output weights of the parallel pair will be related to the input weight, the outputs, and output weights of a single element by the same factor, $\sqrt{2}/2$. For $n > 2$ elements in parallel, parameters change by a factor $\sqrt{n}/n$. The system performance can be insensitive to insertion or removal of redundant elements by having each sensor "count," by

any convenient scheme, the number of elements for which it is the principal input. An alternative method is to use enough units in parallel to render the change from $\sqrt{n}/n$ to $\sqrt{(n-1)}/n-1$ relatively insignificant.

In the above, it was noted that the equations for computing the various parameters hold only when the disturbance is zero. Obviously, this condition is not obtained in realistic situations. A number of schemes for cancelling the disturbance or for stopping computation of parameters while disturbances were present have been evaluated. One which has so far been effective is as follows.

Assume that disturbances are relatively rare, so that the system has adequate settling time between disturbances. The system will then ordinarily be near a null when a disturbance occurs. As a result, during the initial period of the disturbance both the disturbed signal y_i and its derivative $\dot{y}_i$ will have the same algebraic sign. We can avoid disturbing the previously computed values of the α_{ij} and β_{ij} during this period by requiring always that y_i and $\dot{y}_j$ have opposite signs before computation of the α_{ij} and β_{ij} will take place.

EXPERIMENTAL DATA

Figures 5 and 6 present some results of simulation experiments.* These results, chosen from simulation of a second-order system as an environment, show the changes in system behavior with time as parameters were optimized (Fig. 5), and typical results of a test in which all parameters but one were fixed and the other varied by hand around the system-computed value to test whether the value was indeed optimum (Fig. 6). Other experiments, which simulated partial failures of multiple elements and /or changes in the environment, have also agreed with the theoretical predictions.

CONCLUSIONS

The foregoing shows that it is possible to design a system made up of iterative elements which are in a sense "self-healing" and which optimize system performance under certain constraints. The elements certainly

* Simulations were conducted by Mr. G. H. Bolen, whose assistance is gratefully acknowledged.

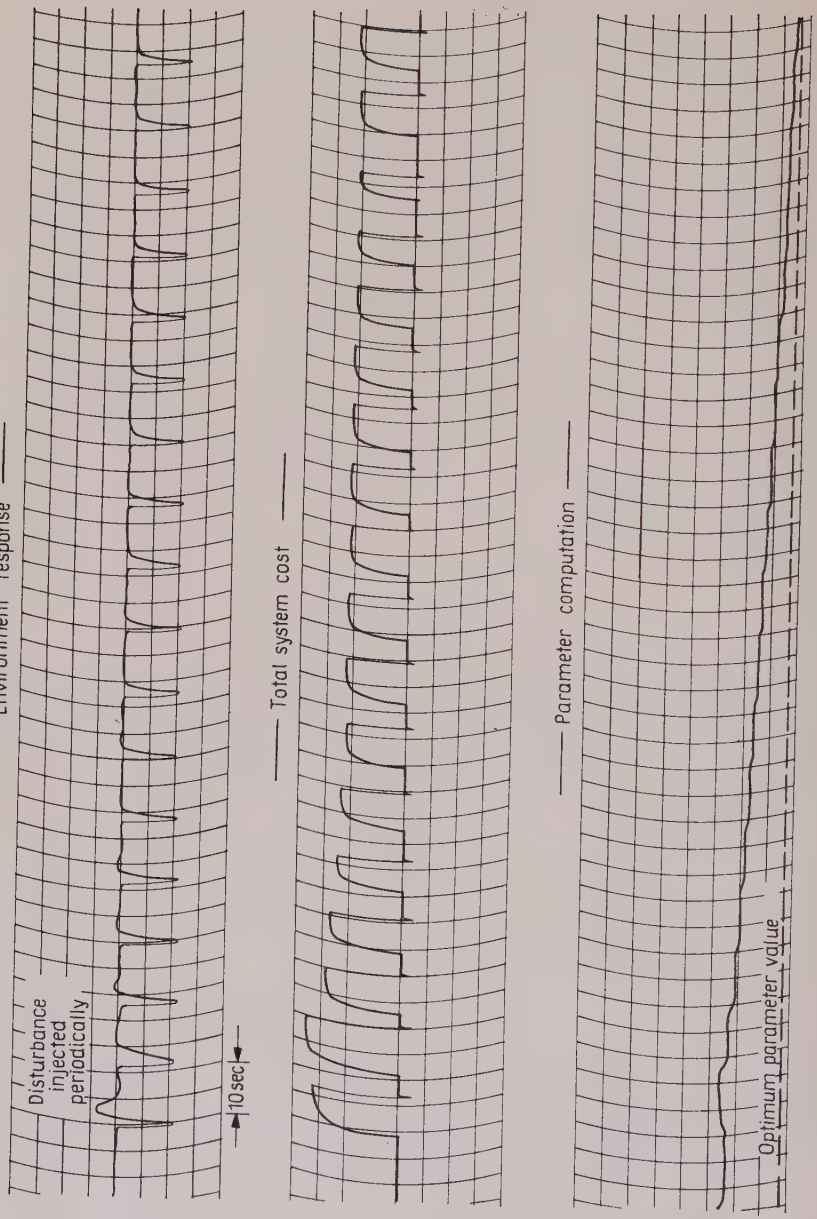


Fig. 5. Typical Time History.

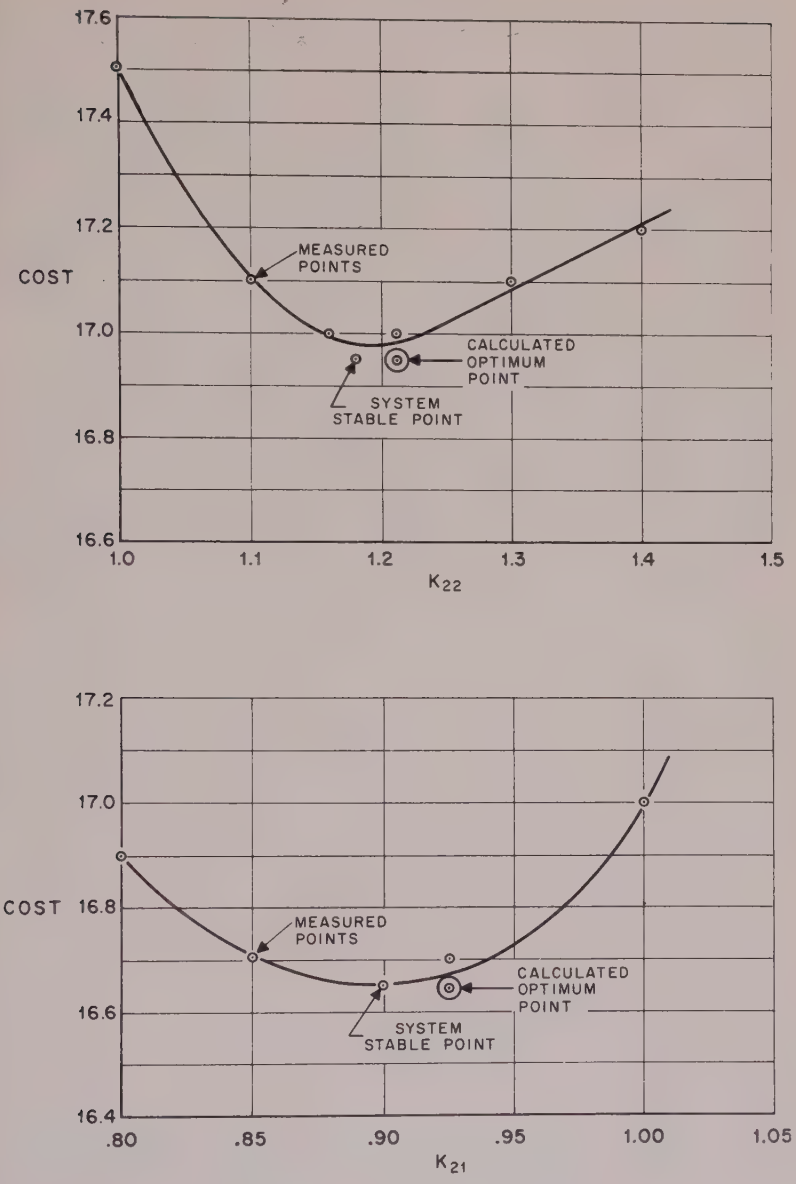


Fig. 6. "Cost" vs Parameter Curves

do not resemble the usual model of the neuron in the literature. It is the writer's belief, however, that the system presented displays behavior that in some respects models that of living creatures much more closely than more conventional threshold models. No *a priori* knowledge of "correct" system response is required, as is represented by the "teacher" in the usual neural net schemes. Extension of the ideas presented above to nonlinearities is not straightforward, although the threshold behavior of the neuron provides a clue. The requirement that inputs be differentiated offers a cogent explanation for the cortical currents observed by Kohler, *et al.*[5] The two-way communication required along neural pathways may be an explanation for longitudinal currents observed by Becker. [6]The requirement for detecting the fact that y_i and $\dot{y}_i$ have the opposite signs is simplified to requiring that y be negative if y_i has only a single polarity and signals of the opposite polarity are handled by another sensor y_i, as is seen in the nervous system.

Partitioning of an A-matrix which is sparse or having a B-matrix with a row of zeroes can lead to multi-level configurations like those studied by Mesarovic[7] and Pearson.[8] The reader can confirm readily that multiple control tasks with few interaction terms cause elements to group into functional units, with "decussations" and "cross-couplings," resembling somewhat the cytoarchitectonics of multiple level reflex arcs.

Some features of the model do not fit known physiological data, e. g., no clear explanation of slow glial potentials is provided, and facilitation is not explained. Nevertheless, it is the writer's belief that extensions of the ideas presented here promise explanations of nervous system phenomena and creature behavior not explained by conventional models. The concepts, coupled with present advances in microelectronics, also promise control systems of extremely high reliability and unusual adaptive ability.

ACKNOWLEDGEMENT

The research reported in this paper was sponsored in part by the Aerospace Medical Research Laboratories. Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Contract No. AF 33 (615)-1891. Further reproduction is authorized to satisfy needs of the U. S. Government.

REFERENCES

1. Hawkins, J. K. "Self-Organizing Systems—A Review and Commentary," Proc. IRE **49**, 1, pp. 31–48, January 1961.
2. Eccles, J. C. "The Physiology of Nerve Cells," Johns Hopkins Press, Baltimore, Md., 1957.
3. Griffith, V. V. "A Model of the Plastic Neuron," IEEE Trans. on Mil. Electr., 7 Bionics Issue (eds. L. M. Butsch and H. L. Oestreicher), pp. 243–253 (1963).
4. Kalman, R. E. "Contributions to the Theory of Optimal Control," Bol. de la Sociedad Matematica Mexicana, pp. 102–119, 1960.
5. Kohler, W., Held, R., and O'Connel, D. N. "An Investigation of Cortical Currents," Proc. Am. Phil. Soc., **96**, 3 (1952).
6. Becker, R. O. "Some Observations Indicating the Possibility of Longitudinal Charge Flow in the Peripheral Nerves," 2nd Annual Bionics Symposium, Ithaca, N. Y., Plenum Press, New York, N.Y., 1961.
7. Mesarovic, M. D. "Multilevel Heuristic and Problem Solving," 1963 Bionics Symposium, Report No. ASD-TDR-63-946, Wright-Patterson Air Force Base, Ohio.
8. Pearson, J. D. "Multilevel Programming" and other papers in "Papers on Multilevel Control Systems," Report No. SRC 70-A-65-25, Case Institute of Technology, Cleveland, Ohio.

J. SKLANSKY*†

University of California

Irvine, California

Time-Varying Threshold Learning

INTRODUCTION

Threshold learning processes ("TLPs") are a class of models of simple adaptation appearing in pattern recognition, pulse-coded communication, and psychophysics. For workers in automatic control, TLPs represent a new direction of application of the feedback concept.

Earlier studies of TLPs have been concerned with fixed-increment or "simple incremental" feedback. Examples of such TLPs are Rosenblatt's "perceptron"[1] and Widrow's "adeline"[2]. The present paper is concerned with TLPs having varying-increment feedback. The statistical performance of these models is referred to here as "time-varying threshold learning" to distinguish them from "fixed-increment" or "simple" threshold learning. An interesting approximate relation between time-varying and simple threshold learning has been discovered, and is derived in this paper.

WHAT IS A THRESHOLD LEARNING PROCESS?

The threshold learning process, or TLP for short, consists of an information source, a noisy channel, a threshold detector, and a feedback policy. A block diagram showing the interconnection of these elements is given in Figure 1. The source transmits a binary signal u, in the form of a random sequence of 0's and 1's, through the channel. The threshold detector senses an analog signal u at the output of the channel, and,

* Formerly with The National Cash Register Co., Dayton, Ohio

† The research for this paper was done at RCA Laboratories, Princeton, N.J., and at The National Cash Register Company, Dayton, Ohio. Part of this research was sponsored by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under contract No. AF 33 (615)-1764 with RCA Laboratories.

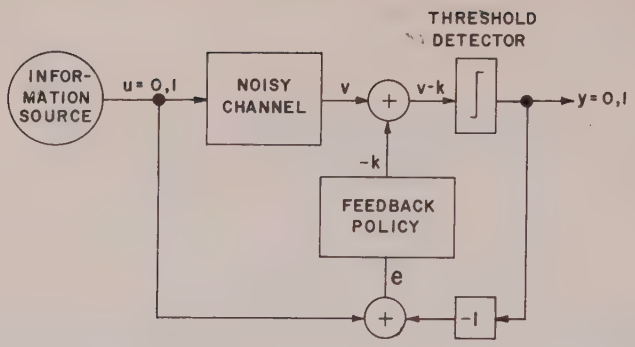


Fig. 1. The Threshold Learning Process (TLP).

comparing this quantity to a "threshold" k, guesses the value of the transmitted signal.

Learning takes place by means of a "feedback policy" which compares the guess y with the correct answer, u. After a training period of prescribed length, the TLP moves into a "working" phase, in which the observer receives no reinforcement signals. The effect of the noisy channel on the source is described by a pair of "constituent" probability densities $f_0(v)$ and $f_1(v)$. Examples of these densities are shown in Figure 2.

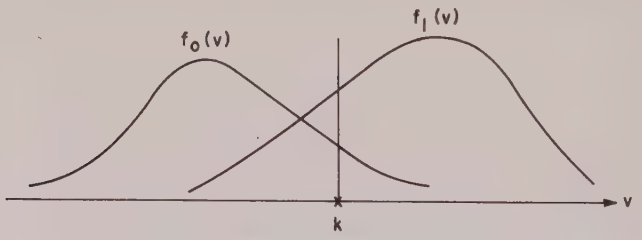


Fig. 2. Constituent Densities of v.

The differential quantity $f_i(v)\,dv$ is defined as the joint probability of transmitting i and observing a signal occupying the interval $(v, v + dv)$. In other words, $f_0(v)$ is the distribution of the v's caused by the 0's, while $f_1(v)$ is the distribution of the v's caused by the 1's. The area under $f_0(v)$ is ϱ, the probability of transmitting a 0; the area under $f_1(v)$ is $1 - \varrho$, the probability of transmitting a 1. The noise in the channel contributes to the obtuseness or variance of the constituent densities; without noise $f_0(v)$ and $f_1(v)$ would be nonoverlapping narrow spikes.

The "feedback policy" determines the new threshold after each reinforcement. The "fixed increment" feedback policy works as follows. The threshold has K possible values. The threshold is moved up or down by one increment in response to a false alarm or false rest, respectively. A false alarm is a guess of 1 when the source sent a 0; a false rest is a guess of 0 when the source sent a 1. In Figure 1, false alarm and false rest correspond to $e = -1$ and $e = 1$, respectively, where $e \equiv u - y$. The threshold remains fixed if no error is incurred ($e = 0$) or if a boundary threshold prevents a desired adjustment. We refer to TLPs having this feedback policy as "simple incremental TLPs" or "fixed-increment" TLPs.

The "learning wave" of a TLP is defined as an ensemble average of the performances of a large number of identical TLPs as a function of the number of trials, n. If the guess y and the signal u are equal, the performance index is given a value "1"; if $y \neq u$, the performance index is given the value "0". The ensemble average of these performance indices as a function of the number of trials, n, is denoted by $z(n)$, and called the "learning wave" of the TLP. The quantity $z(n)$ may be interpreted as the probability of a successful guess at trial n.

In earlier work on threshold learning, fixed-increment TLPs were found to have learning waves restricted to these shapes: monotonic increasing, monotonic decreasing, and single-peaked.[4] Although a proof of this restriction was obtained only for the case where the feedback increment is infinitesimally small, empirical evidence supports the conjecture that the restriction holds with only slight deviations even when the feedback increments are large. Among the three shapes, the monotonic increasing shape is usually associated with the longest learning time. ("Learning time" is defined as the time spanned by the transient component of the learning wave in reaching within ten per cent of the transient component's initial value.) The monotonic increasing shape has been observed much more frequently than the other two.

The motion of the threshold in a fixed-increment TLP can be modeled by a Markov chain.[6] Consequently the learning waves of fixed-increment TLPs can be found with the help of the well developed theory of Markov chains. In a time-varying TLP, however, the Markov chain's transition probabilities as well as the number of states are functions of time. Consequently the utility of Markov chain techniques for time-varying TLPs is not at all obvious. In this paper a certain amount of utility of Markov chain techniques has been retained for time-varying TLPs.

WHY IS TIME-VARYING FEEDBACK USEFUL?

One of the disadvantages of fixed-increment feedback is that the variance of the threshold is never zero, even after an infinite length of training. To be specific, the asymptotic variance of the threshold of a fixed-increment TLP is directly proportional to the size of the increment when the increment is sufficiently small. To achieve an asymptotic variance of zero when the input statistics are stationary, the size of the threshold increment must be brought toward zero asymptotically. One way to achieve this is by the Robbins-Monro method of stochastic approximation.[3] In this method the size of the threshold increment is inversely proportional to the number of training samples or "trials." The Robbins-Monro method yields an asymptotic threshold for which the false-alarm and false-rest probabilities are equal. In practice this threshold is often close to the "optimal" threshold, where the error probability is a minimum. The Robbins-Monro technique is an example of a time-varying feedback policy, and achieves the desired asymptotic compression of the threshold's variance.

In this paper we show how the learning times of time-varying TLPs are related to the learning times of fixed-increment TLPs. This contributes an insight into the design of time-varying feedback policies for threshold learning processes.

THE LEARNING TIMES OF TIME-VARYING *TLPS*

Since the learning wave $z(n)$ of a simple incremental TLP is usually monotonic and has a nonzero asymptote, $z(n)$ has approximately the following x-transform:[4]

$$z(x) = \frac{\zeta}{1 - x} + \frac{z_0 - \zeta}{1 - \dfrac{x}{\xi}} \tag{1}$$

In the sequel we shall assume that this equation is exact, even though it is actually an approximation.

Equation (1) contains three parameters: z_0, ξ, and ζ/x. The quantity

z_0, the initial value of z_n, may be evaluated by the formula

$$z_0 = \boldsymbol{r}(0)\,\boldsymbol{q}$$

where $\boldsymbol{r}(0)$ is the initial row vector of threshold probabilities, and $\boldsymbol{q}$ is the column vector of conditional success probabilities. The quantity ξ is the smallest nonunity root of the equation $|\boldsymbol{I} - \boldsymbol{P}\xi| = 0$, where $\boldsymbol{P}$ is the transition probability matrix of the TLP, $\boldsymbol{I}$ is the unit matrix, and $|\boldsymbol{I} - \boldsymbol{P}\xi|$ is the determinant of $\boldsymbol{I} - \boldsymbol{P}\xi$. The asymptotic success probability, $\zeta - z(\infty)$, may be computed by the signal flow techniques described in the earlier report.[4]

In a time-varying TLP, the learning wave has the following "frozen" form:[5]

$$z(x, n) = \frac{\xi}{1 - x} + \frac{z_n - \zeta_n}{1 - \dfrac{x}{\xi_n}} \tag{2}$$

where $z(x, n)$ is the x-transform of the learning wave of a TLP whose threshold increment is "frozen" at its value at time n. The quantity z_n is the success probability at time n, and ζ_n is the asymptotic success probability of the TLP whose threshold increment is frozen at its value at time n. ("Time" is equivalent to "trial number.")

Equation (2) represents the following difference equation:

$$z_{n+1} = \zeta_n + (z_n - \zeta_n)\,\xi_n^{-1}$$

Or

$$z_n - \xi_{n-1}^{-1} z_{n-1} = \zeta_{n-1}(1 - \xi_{n-1}^{-1}).$$

Let

$$a_{n-1} = \zeta_{n-1}(1 - \xi_{n-1}^{-1}).$$

Then

$$z_n - \xi_{n-1}^{-1} z_{n-1} = a_{n-1}. \tag{2a}$$

This equation has the following solution:

$$z_n = a_{n-1} + \frac{a_{n-2}}{\xi_{n-1}} + \frac{a_{n-3}}{\xi_{n-1}\xi_{n-2}} + \cdots + \frac{a_1}{\prod_{2}^{n-1} \xi_i} + \frac{a_0}{\prod_{1}^{n-1} \xi_i} + \frac{z_0}{\prod_{0}^{n-1} \xi_i} \tag{2b}$$

where

$$a_i = \zeta_i(1 - \xi_i^{-1}).$$

Substituting Eq. 2a into Eq. 2b yields

$$z_n = \zeta_{n-1} + \frac{\zeta_{n-2} - \zeta_{n-1}}{\xi_{n-1}} + \frac{\zeta_{n-3} - \zeta_{n-2}}{\xi_{n-1}\xi_{n-2}} + \cdots + \frac{\zeta_0 - \zeta_1}{\prod_1^{n-1} \xi_i} + \frac{z_0 - \zeta_0}{\prod_0^{n-1} \zeta_i} \tag{3}$$

A ROUGH ESTIMATE

A relation between the learning time of a time-varying TLP and the learning times of the constituent frozen (fixed-increment) TLPs, under the assumption that the ζ_n's of the frozen TLPs are equal, will now be derived. The empirical basis for this assumption is shown in Figure 3.

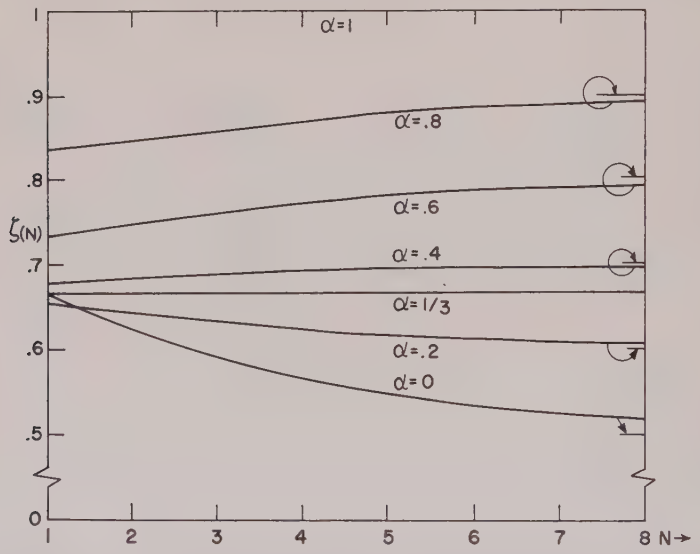


Fig. 3. Asymptotic Success Probability of Simple TLPs as a Function of the Size of the Threshold Increment, 1/N.

In this figure the asymptotic success probabilities $\zeta(N)$ for a large number of fixed-increment TLPs are plotted. The quantity N is the reciprocal of the normalized increment size. The parameter α is a quantity inversely related to the obtuseness of the constituent densities (Fig. 2). The variation of ζ with the size of the increment, $1/N$, is observed to be small. Denote

the learning time of the time-varying TLP by λ, and the learning times of the constituent frozen TLPs by L_i. Then find the value of m which minimizes

$$\left| \frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_m} - 1 \right| \tag{4}$$

This value of m is approximately equal to λ.

To derive Eq. 4, recall that we are assuming that

$$\zeta_i - \zeta_{i-1} = 0 \text{ for all } i.$$

Hence, by Eq. 3,

$$z_n = \zeta + \frac{z_0 - \zeta}{\prod_0^{n-1} \xi_i} \tag{5}$$

where $\zeta = \zeta_i$ for all i.

Note that, by Eq. 5, $\lim_{n \to \infty} z_n = \zeta$, as it should. By Eq. 5

$$\frac{\zeta - z_n}{\zeta - z_0} = \frac{1}{\prod_0^{n-1} \xi_i} \tag{6}$$

The learning time of z_n is defined as that value of n such that

$$\frac{\zeta - z_n}{\zeta - z_0} = \frac{1}{c} \tag{7}$$

where c is an arbitrary constant chosen by the designer. A typical value of c is $c = 10$.

Hence, by Eq. 6, $\xi_i^{L_i} = c$, where L_i is the learning time of a fixed-increment TLP. Note that $L_i = (\log_c \xi_i)^{-1}$.

Hence: if $\prod_0^{m-1} \xi_i = c$, then m is the learning time λ of the time-varying TLP.

Recapitulating:

$$\prod_0^{\bar{\lambda}-1} \xi_i \simeq c \tag{8}$$

where $\bar{\lambda}$ is the nearest-integer approximation of λ. Take $\log_c$ of both members of Eq. 8. This yields

$$\frac{1}{L_0} + \frac{1}{L_i} + \cdots + \frac{1}{L_{\bar{\lambda}-1}} \simeq 1 \tag{9}$$

Now renumber all of the subscripts of the ξ_i's as follows:

$$i = i + 1.$$

Then

$$\prod_{1}^{\bar{\lambda}} \xi_j \simeq c \tag{10}$$

Hence

$$\frac{1}{L_1} + \frac{1}{L_2} + \cdots + \frac{1}{L_{\bar{\lambda}}} \simeq 1 \tag{11}$$

which completes the derivation.

Note in the above analysis that the way in which the threshold increments may be varied is not specified. Hence the Robbins-Monro feedback is just one of a large class of feedback policies to which Eq. 11 is applicable.

An improved estimate of λ that takes into account variations of ξ with respect to n is available. A discussion of this improved estimate is omitted here because of space limitations.

AN EMPRIRICAL STUDY

Preliminary computations of the learning waves of a Robbins-Monro TLP and its constituent fixed-increment TLPs have been executed on a computer. For the case $c = 3$, the Robbins-Monro learning time was observed to be 6, while Equation 11 yielded an estimate of 4. Improved performance of Equation 11 is expected for larger values of c, since the fixed-increment learning waves become more truly exponential as the number of trials increases.

Our computer study of time-varying TLPs is being continued, and will be reported in a later paper.

SUMMARY AND CONCLUSIONS

The learning time of any time-varying TLP may be estimated in terms of the constituent "frozen" TLPs provided the learning wave is monotonic or approximately monotonic. A TLP with a Robbins-Monro feedback is an example of a time-varying TLP.

An estimate of the relation between the learning time of a time-varying TLP and the learning times of the constituent frozen TLPs is given by Eq. 11. In that equation, all of the asymptotic success probabilities of the frozen TLPs are assumed to be equal. A more accurate formula is also available. In the more accurate formula small variations of ζ_n with n can be taken into account.

These techniques are contributions toward understanding the dynamics of training threshold devices for pattern recognition and for modeling adaptive threshold phenomena in psychophysics.

In many cases simple Markov chains are inadequate models of the motion of the thresholds, particularly when the variance of the thresholds is reduced asymptotically with increased training. This paper considers feedback policies that vary with time independently of the responses of the TLP, and which can achieve the desired asymptotic compression of the threshold's variance. Future work should consider feedback policies that depend on the responses.

ACKNOWLEDGEMENT

The author is grateful to C. E. Thorstensen of the National Cash Register Co., Dayton, Ohio for programming and executing on a computer the preliminary computations of the learning waves of time-varying TLPs.

REFERENCES

1. Rosenblatt, F. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, D. C., 1962.
2. Widrow, B. "Pattern Recognition and Adaptive Control," *IEEE Trans. on Applications and Industry*, **83**, No. 74, Sept. 1964, pp. 269–277.
3. Robbins, H., and Monro, S. "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, **22**, 400–407 (1951).
4. Kaplan, K. R., and Sklansky, J. "Analysis of Markov Chain Models of Adaptive Processes," Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio, Report AMRL-TR-65-3, January 1965.
5. Zadeh, L. A. "Frequency Analysis of Variable Networks," *Proc. IRE*, **38**, No. 3, March 1950, pp. 291–299.
6. Sklansky, J. "Adaptation Theory—a Tutorial Introduction to Current Research," *Medical Electronics and Biological Engineering*, RCA Publication PE-238, RCA Laboratories, Princeton, New Jersey, 1965.

SATOSI WATANABE*

Dunham Laboratory, Yale University

New Haven, Connecticut

A Method of Self-Featuring Information Compression in Pattern Recognition

ABSTRACT

A succesful method of pattern recognition depends on two kinds of selection of "good" variables. The "post-selection" chooses those variables which serve the purpose of inductive generalization on the basis of given samples of classes. The "pre-selection" aims at eliminating less useful variables without using the knowledge about particular classes and their members. The present-day deadlock in the field of pattern recognition seems to suggest that more emphasis should now be placed on the study of pre-selection. There should be many different approaches to tackle this problem, and SELFIC (Self-Featuring Information Compression) should be considered as one of them. It automatically chooses those variables which serve effectively the purpose of discriminating individual objects among themselves in a given collection of objects which require classification. The strong point of SELFIC is that it is a universal method independent of the type of classification problem one may have at hand. The relationship of this method to the Karhunen-Loève expansion and to factor analysis will be discussed. An application of the mechanical method to speech recognition will be adduced.

LIMITATIONS OF THE METHOD OF SELFIC

In view of the Theorem of the Ugly Duckling[1] and its natural extension to the case of continuous variables, it is obvious that any task of classifying objects with known properties into classes presupposes, and depends essentially on, an uneven distribution of weight or importance over the possible predicates or variables which can participate in describing these properties. Generally speaking, there are two kinds of tasks of classification: In "clustering problems," we are given an "ensemble" of objects with known properties and we are asked to group them into classes in such a way that two members of the same class resemble each other much more strongly than two members of two

* Now at University of Hawaii, Honolulu, Hawaii.

different classes, whereby the degree of resemblance must be determined on the basis of *more important* predicates or variables. In "recognitive problems," we are first shown a certain number of sample members, or "paradigms," of given classes, and we are asked thereafter to place newcomers in these classes inductively generalizing the class characteristics suggested by the paradigms. In either case, "more important variables" are those which are intimately related to the "features" of the classes, which are yet to be formed or are already exemplified by the paradigms.

In clustering problems, the weighting of variables must be determined first, and the classes must be thereafter created in accordance with this weighting. For this reason, this type of weighting is called "pre-weighting" or "pre-selection."[2] The term selection is used because unimportant variables can often be ignored. In "recognitive problems," we should first go through a "pre-selection" to eliminate obviously unimportant variables, and then apply a second-stage weighting process in order to bring out in bold relief the class features as suggested by the paradigms. This second stage weighting is called "post-weighting" or "post-selection."[2]

The ground for pre-selection can be either "external" or "internal." In the former case, an externally given prescription or our vast experience in broader problems will be the guide. If, in the character recognition, only those properties are said to be important which are invariant for translational, rotational, dilatational, and some topological transformations, then this is an example of pre-selection of "external" nature, in the sense that it has nothing to do with the particular examples of written or printed letters at hand. In a pre-selection of "internal" nature, the weighting or selection must be derived from the ensemble of objects at hand in such a way that those variables which serve efficiently to discriminate different members of the collection get larger importance and those variables which give more or less equal observed values for different members of the collections get less importance. The Euclidian distance in the space of these variables is assumed to have an intrinsic meaning. The self-featuring information compression (or SELFIC) is precisely a method of such an internal pre-weighting or pre-selection. It is "self-featuring" in the sense that it discovers the main features of the given ensemble in accordance with its own statistic properties. It is independent of "external" criteria or of "paradigms." It is a method of information compression, because it serves to eliminate information

which is not pertinent to our task of class formation. Sometimes, information compression in the sense of elimination of redundancy is also effected by SELFIC. A post-selection must follow this SELFIC to perform a real recognitive task.[2] The possibility, however, cannot be entirely excluded that a variable which may become important in the post-selection is already eliminated by a careless pre-selection.

The raw data about the properties of the objects are, of course, based on an observation of a certain set of qualities and quantities, and implies tacitly already a certain selection of variables, or rather, a pre-selection of variables. But, a discussion of this process is outside the scope of the present paper.

SELFIC COORDINATE SYSTEM

We consider an ensemble of objects in which an object-type α ($\alpha = 1, 2, \ldots, \nu$) is represented with relative frequency $w^{(\alpha)}$, where $w^{(\alpha)} \geqq 0$, $\sum_{\alpha=1}^{\nu} w^{(\alpha)} = 1$. An object of type α is assumed to be represented by a square-integrable real function $f^{(\alpha)}(\xi)$, defined in a real domain $a \leqq \xi \leqq b$, and normalized to unity in this domain. The case where the objects are represented by n-component vectors will be discussed later, on page 702. By the help of an orthonormal set of functions, $\{\psi_i(\xi)\}$, we expand $f^{(\alpha)}(\xi)$ as

$$f^{(\alpha)}(\xi) = \sum_{i=1}^{\infty} x_i^{(\alpha)} \psi_i(\xi). \tag{1}$$

The quantity $(x_i^{(\alpha)})^2$ can be considered as a measure of importance of the base function, or coordinate, $\psi_i(\xi)$ in expressing $f^{(\alpha)}(\xi)$. Hence, its average in the ensemble

$$\varrho_i = \sum_{\alpha=1}^{\nu} w^{(\alpha)} (x_i^{(\alpha)})^2$$

is the measure of importance of $\psi_i(\xi)$ in mathematically representing the ensemble. We have evidently

$$\varrho_i \geqq 0, \quad \sum_{i=1}^{\infty} \varrho_i = 1 \tag{3}$$

From the point of view of information compression which aims at extracting important variables, it is desirable that the $\{\psi_i(\xi)\}$ is such

that the ϱ's are concentrated on a few coordinates instead of being widely spread over many. To formulate mathematically this idea, it will be convenient to introduce the entropy function

$$S(\{\psi_i\}) = -\sum_{i=1}^{\infty} \varrho_i \log \varrho_i \tag{4}$$

Then, an optimal coordinate system $\{\varphi_i(\xi)\}$ will be characterized by

$$S(\{\varphi_i\}) = \min_{\{\psi_i\}} S(\{\psi_i\}) \tag{5}$$

If we use an optimal coordinate system $\{\varphi_i(\xi)\}$ in Eq. (1), then we can consider those x_i's of Eq. (1) corresponding to larger ϱ_i's as "more important" variables in describing the statistically salient features of the ensemble. We shall call such $\{\varphi_i\}$ a SELFIC coordinate system. The SELFIC itself, as the above derivation shows, does not select properties which are invariant for some geometrical transformations. However, if the important properties are invariant for translations in the argument variable (ξ of $f^{(\alpha)}(\xi)$), then one should first get its Fourier transform (or power spectrum) and then apply SELFIC to this transform. See our example on page 703. If an important property can be extracted only by a non-linear transformation, such a transformation must be applied before using SELFIC, because this is essentially a linear transformation.

KARHUNEN-LOÈVE EXPANSION

A Karhunen-Loève coordinate system $\{\varphi_i(\xi)\}$ is defined in the ensemble, $\{f^{(\alpha)}(\xi), w^{(\alpha)}\}$, as the set of eigen-functions of the symmetric matrix defined by

$$(\xi|G|\xi') = \sum_{a=1}^{\nu} w^{(a)} f^{(a)}(\xi) f^{(a)}(\xi'), \tag{6}$$

i.e.,

$$\int_a^b (\xi|G|\xi')\, \varphi_i(\xi')\, d\xi' = \lambda_i \varphi_i(\xi). \tag{7}$$

The labeling of the eigen-functions $\varphi_i(\xi)$ and corresponding eigenvalues λ_i is assumed to be determined so as to satisfy

$$\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \ldots \tag{8}$$

If $\{f^{(\alpha)}(\xi)\}$ contains only μ linearly independent functions, Eq. (7) can define only μ eigen-functions. In order to obtain a complete set $\{\varphi_i\}$, we

have to add functions which are mutually orthogonal and orthogonal to the original μ eigen-functions. These additional functions can be considered as eigen-functions corresponding to the infinitely degenerate zero eigen-value of G. In practice, however, we do not need these functions.

Theorem **1**. A Karhunen-Loève coordinate system defined by Eqs. (6) and (7) is a SELFIC coordinate system defined by Eq. (5), and vice versa.

Proof. A brief proof that a K.-L. coordinate is a SELFIC coordinate system is given in the Appendix of the present paper. For the converse, see Reference 2 or Reference 3.

The Karhunen-Loève expansion is usually known[4] as the error-minimizing coordinate system in the following sense. The average error committed by taking only a finite number of terms in the expansion (1) is given by

$$E(\{\psi_i\}, m) = \sum w^{(a)} \int_a^b \left[f^{(a)}(\xi) - \sum_{i=1}^{m} x_i^{(a)} \psi_i(\xi) \right]^2 d\xi \tag{9}$$

Theorem **2**. A Karhunen-Loève coordinate $\{\varphi_i(\xi)\}$ is such that for every integer m

$$E(\{\varphi_i\}, m) = \operatorname*{Min}_{\{\psi_i\}} E(\{\psi_i\}, m) \tag{10}$$

Proof. See Reference 2 or Reference 3. Theorem 2 is a well-known fact, while Theorem 1 seems to have been unnoticed until now.

It should be mentioned that the first eigen-function, $\varphi_1(\xi)$, corresponding to the largest eigen-value tends to represent the average of the $f^{(\alpha)}(\xi)$ if this is not zero, hence, it is advisable to subtract the average of the $f^{(\alpha)}(\xi)$ from each $f^{(\alpha)}(\xi)$ from the beginning so that the first eigen-function already serves the purpose of discriminating different $f^{(\alpha)}(\xi)$. In recognition problems (as distinct from clustering problems), additional possibilities appear if we apply SELFIC separately to each class of paradigms. For the two-dimensional K.-L. expansion, see Reference 11.

REMARKS ABOUT FACTOR ANALYSIS

In the case where each object of the ensemble is represented by an n-component vector, $f^{(\alpha)}(\xi)$, $\xi = 1, 2, \ldots, n$, all we need to change in the foregoing discussion is to substitute a summation over $\xi = 1, 2, \ldots, n$

for an integration with respect to ξ from a to b. Similarly, the summation $\sum_{i=1}^{\infty}$ in Eq. (1) and subsequent equations has to be replaced by $\sum_{i=1}^{n}$. The number μ of linearly independent vectors in $\{f^{(\alpha)}(\xi)\}$ can never be larger than n. Under these circumstances, the eigen-functions $\varphi_i(\xi)$, of Eq. (7), become n eigen-vectors of an n by n symmetric matrix $(\xi\,|G|\,\xi')$, $\xi, \xi' = 1, 2, \ldots, n$. Of these n eigen-vectors only μ are actually meaningful.

In factor analysis, we consider n real random variables, x_i, $i = 1, 2, \ldots, n$. In the population under consideration, we assume that we obtain a set of observed values $(x_1^{(\alpha)}, x_2^{(\alpha)}, \ldots, x_n^{(\alpha)})$ with relative frequencies $w^{(\alpha)}$, $\alpha = 1, 2, \ldots, \nu$. We agree that the origin and the scale of each random variable are so chosen that

$$\left.\begin{aligned} \langle x_i^{(\alpha)2}\rangle_\alpha &= \sum_{\alpha=1}^{\nu} w^{(\alpha)} x_i^{(\alpha)2} = 0 \\ \langle x_i^{(\alpha)2}\rangle_\alpha &= \sum_{\alpha=1}^{\nu} w^{(\alpha)} x_i^{(\alpha)} = 1 \end{aligned}\right\} \tag{11}$$

The central idea of factor analysis is to consider the existing correlation among the x's as stemming from a relatively small number m of common variables on which the x's are linearly dependent. Thus, we write

$$x_i = \sum_{j=1}^{m} a_{ij} y_j + b_i z_i, \quad i = 1, 2, \ldots, n \tag{12}$$

where the y's are the hidden common variables (the so-called "factors") and the z's are the so-called specific variables introduced as a necessary compromise because the variation of x_i cannot be entirely attributed to the common factors. The conditions of the type (11) are assumed for the y's and z's. We can point out already at this starting point the basic dilemma of the method of factor analysis, namely, in an effort to attribute the multifariousness of the apparent data (x_i) to fewer hidden "causes" we are forced to introduce more new variables than we had at the beginning ($m + n$ versus n). The requirement that all $m + n$ variables be statistically independent (i.e., orthogonal) cannot eliminate the enormous degree of arbitrariness introduced by this assumption (12).

The obvious analogy* between the factor analysis and the method of

* The first line of Eq. (11) is satisfied in the SELFIC method, if the average function is subtracted from each function. The normalization in SELFIC is such that the sum over i of the quantity in the second line of (11) is equated to unity. This difference, of course, does not change the essence of the method.

SELFIC suggests the following solution to this dilemma. Ignore all z's from (4.2) and make $m = n$. Consider the correlation matrix

$$r_{ij} = \langle x_i^{(\alpha)} x_j^{(\alpha)} \rangle_\alpha = \sum_{k=1}^{n} a_{ik} a_{jk}, \tag{13}$$

which corresponds to the G of the SELFIC method. The degree to which the variable y_j influences the variable x_i can be measured by a_{ij}^2, and the degree of the overall influence of the "factor" y_j on the x's can be expressed by $\sum_{i=1}^{n} a_{ij}^2$. The eigen-vectors of r_{ij}, arranged in the descending order of eigen-values as in Eq. (8), will produce the y's in the descending order of this overall influence. The degree to which the influence of a y_j is spread out over different x's can be measured by

$$\left.\begin{aligned} \sigma_j &= -\sum_{i=1}^{n} \tau_{ij} \log \tau_{ij} \\ \text{with} \quad \tau_{ij} &= a_{ij}^2 / \sum_{i=1}^{n} a_{ij}^2 \end{aligned}\right\} \tag{14}$$

The distinction between a common factor y_j and a specific variable z_i is thus reduced to a matter of degree. Those y_j's with large σ_j are more like a common factor and those y_j's with small σ_j are more like specific variable. In this way, we can avoid the enormous redundancy inherent in the method of factor analysis. See References 2 and 3 for more details. It may be noted that the idea of diagonalization of the correlation matrix was suggested as early as in 1901, by Karl Pearson,[5] and it took a detour of more than half a century and the invention of electronic computers for the factor-analysts to come back to this simple starting point.[6,7,8]

APPLICATION OF THE METHOD OF SELFIC TO SPEECH RECOGNITION

Figure 1 represents results obtained by applying the SELFIC method to the deviation from the average of the power-spectra (obtained by 36 bandpass filters) of 12 different vowels (*i*, *I*, *ε*, *a*, *e*, *o*, *ʌ*, *u*, *U*, *æ*, *ɔ*, *ə*.) spoken by 19 different persons (male and female).* Each record is a 36-component vector. The center-frequencies of the filters range from

* The experimental data were provided by S. Bakis of IBM Research Laboratory.

100 to 10,000 cycles per second and their bandwiths from 50 to 1,200 cycles per second. Figure 1 represents 12 × 19 points in the space defined by "longitude" $\tan^{-1}(c_2/c_1)$ and "latitude" $\tan^{-1}\left(c_3/\sqrt{(c_1^2 + c_2^2)}\right)$, where c_1, c_2, and c_3 are the first, second and third coefficients in the expansion (2.1) using the SELFIC coordinate system.* The subscript α in this case identifies each of the 12 × 19 points.

Figure 1 shows also a rough determination of 9 disjoint regions, purposely tolerating intrusion of some "alien" elements in each territory

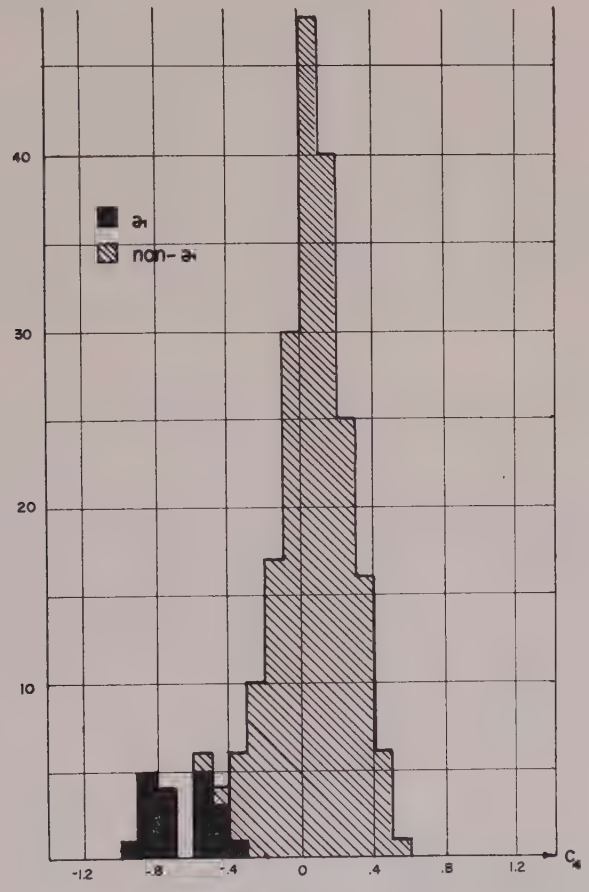


Fig. 2

* The programming on IBM 7094 was carried out by Joseph Harry of IBM Research Laboratory.

in order to avoid excessive gerrimandering. These zones have been drawn by a human hand to facilitate inspection. The *o* and the ɔ occupy a common region, and this may lie in the nature of things. The distinction between *I* and *e* is so small that a confusion of these two may be forgiven at this stage of the game. The most troublesome one is ə (*er*, *ir*, or *ur* in the usual spelling) which intrudes different regions. However, the histogram of Figure 2, which represents the distribution along the fourth coefficient c_4, shows that the ə forms almost a separate island in the c_4 space. It is to be expected from the beginning that considerable overlapping of regions is inevitable in speech recognition. We should probably be surprised that the extent of overlapping is not more than seen on Figure 1. Determination of the class-affiliation of a newly arriving signal can be done by the customary decision function method or the Bayesian decision method. After the author decided to try the present method to a speech recognition problem, it came to his attention that Kramer, Mathews[9] and Bakis[10] had earlier undertaken somewhat similar attempts, but for an unknown reason, their attempts were not as successful as the present one. Their theoretical motivations were quite different from our point of view.

ACKNOWLEDGEMENT

The results reported here have been obtained by work which was partly supported by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Contract AF 33 (657)-11347 and by the Decision Science Laboratory under Contracts AF 19 (628)-4303 and AF 19 (628)-4804. Further reproduction is authorized to satisfy needs of the U. S. Government.

APPENDIX

The ϱ's are the diagonal elements of G in some coordinate systems $\{\psi_i\}$, while the λ's are the eigen-values of G, i.e., they are the ϱ's in the particular coordinate system $\{\varphi_j\}$ which consists of the eigen-vectors of G. Hence, the ϱ's and the λ's are connected by

$$\varrho_i = \sum_j A_{ij}\lambda_j, \tag{A.1}$$

where

$$A_{ij} = T_{ij}(T^T)_{ji}, \tag{A.2}$$

and

$$T_{ij} = \int_a^b \psi_i(\xi) \cdot \varphi_j(\xi)\, d\xi. \tag{A.3}$$

The condition that T represents an ortho-normal transformation, $T^{-1} = T^T$, guarantees the double stochasticity of A_{ij}:

$$\sum_i A_{ij} = \sum_j A_{ij} = 1. \tag{A.4}$$

The assertion of Theorem 1, which can be written as

$$-\sum_i \varrho_i \log \varrho_i \geqq -\sum_i \lambda_i \log \lambda_i \tag{A.5}$$

becomes, in view of (A. 4), simply the well-known H-Theorem.

REFERENCES

1. Watanabe, S. An article in S. Dockx and P. Bernays (editors), *Information and Prediction in Science*, New York, Academic Press, 1965.
2. Watanabe, S. *Knowing and Guessing*, New York, John Wiley and Sons, (in preparation).
3. Watanabe, S. "Karhunen-Loève Expansion and Factor Analysis—Theoretical Remarks and Applications." in the *Proceedings of the Fourth Prague Conference on Information Theory, etc, 1965*. Czechoslovac Academy of Science, 1966.
4. Loève, M. *Probability Theory*, 3rd Edition, Princeton, Van Nostrand, 1963.
5. Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space," Philos. Mag., 1901, p. 559.
6. Harman, H. H. *Modern Factor Analysis*, University of Chicago Press, 1960.
7. Thurstone, L. L. *Multiple Factor Analysis*, University of Chicago Press, 1947.
8. Hotelling, H. "Analysis of Complex Statistical Variables into Principal Components," Exp. Psy. **24** (1933) 417 qnd 498.
9. Kramer, H. P., and Mathews, M. V. "A Linear Coding for Transmitting a Set of Correlated Signals," IRE Trans. IT-2, 1956, p. 41.
10. Bakis, S. "Some Observations Concerning the Number of Filters Needed for Speech Analysis," Research Memorandum PR 13, AR 31, February 17, 1960, (IBM internal publication).
11. Wong. E. "Vector Stochastic Processes in Problems of Communication Theory," Thesis, Princeton University, May, 1959.

IVAN H. SUBLETTE

RCA Laboratories

Princeton, N. J.

Recognition of Class Membership by Means of Weak, Statistically Dependent Features

INTRODUCTION

Consider the problem of designing a machine M that will recognize objects in some universe U as members of classes. Each object in U belongs to one of the two classes A and B, and the probability that an object belongs to a particular class is determined by the values of features that it possesses. The set of features $\{f_i\}$ $(i = 1, 2, \ldots, n)$ has already been chosen, and $f_1 = 0$ or 1 for all i. The classes are defined for M only by a small set of examples, which are objects whose class membership and feature values are given. M is then required to recognize the class of any object not included in the set of examples.

The machine M is visualized as a digital computer containing in its memory the class membership and the feature values for all of the examples. It is assumed that the memory capacity of M is adequate for this purpose, and there will be no attempt to reduce the required storage capacity by storing only "typical" examples, as in Sebestyen's adaptive sample set construction technique.[1] The basic limitation is assumed to be the smallness of the number of examples, not the memory capacity or the computing speed of M.

Let x be an unknown object whose class is to be determined. For each feature f_i we define a corresponding "value set" F_i consisting of those members u of U for which $f_i(u) = f_i(x)$. A "product set" is defined as the intersection of a finite number of distinct value sets (a single value set is also considered to be a product set). Let $P(Y_1|Y_2)$ denote the probability that an object contained in the set Y_2 is also contained in the set Y_1, and let J denote the product set formed by intersecting all n of the value sets. It is known that the recognition error rate is minimized by the following rule: decide $x \in A$ if $P(A|J) \geqq P(B|J)$, and decide $x \in B$ if $P(A|J) < P(B|J)$.

The main difficulty with this approach arises when the probabilities $P(A|J)$ and $P(B|J)$ must be estimated with a small number of examples. It would be naive to attempt, say, a maximum likelihood estimate of $P(A|J)$ based on the examples contained in J, since it is unlikely that a set defined by the intersection of n value sets will contain any examples. One way out of this difficulty is to assume that the features are statistically independent; in this case the problem of estimating the ratio $P(A|J)/P(B|J)$ reduces to the problem of estimating the ratio $P(A|F_i)/P(B|F_i)$ for every i. The probabilities $P(A|F_i)$ and $P(B|F_i)$ can be estimated rather easily, since each value set is likely to contain many examples. Unfortunately, the assumption of independence is unrealistic.

The new recognition procedure to be developed here has resulted from attempts to find ways of estimating probabilities with a small number of examples. This procedure is called a "gerrymandering" procedure because it makes decisions without necessarily looking at all of the features of x.

THE BASIS OF GERRYMANDERING

Let V and W be a product set and a value set, respectively, and let Z be a variable ranging over the classes A and B; if $Z = A$, then $\bar{Z} = B$, and vice versa. Under the hypothesis that x belongs to class Z, the information imparted by the observation that x is contained in W is $-\log P(W|Z)$. When it is also known that x is contained in the set V, the information imparted by the same observation is $-\log P(W|V \cap Z)$. The fundamental hypothesis we shall make about the universe U and the features is that the information obtained in the second case is no greater than the information obtained in the first case *when the class of x is guessed correctly*, i.e., we shall assume that $K(V, W, X) \geqq 1$, where X is the class of x and

$$K(V, W, Z) = \frac{P(W|V \cap Z)}{P(W|Z)} = \frac{P(V \cap W|Z)}{P(V|Z)\,P(W|Z)}.$$

The ratio $K(V, W, Z)$ is commonly called the "association factor." When the events V and W are independent of the hypothesis Z, $K(V, W, Z) = 1$.

For reasons that will become clear, we shall forbid the use of certain value sets as constituent value sets of product sets. Let Λ denote the set

of all value sets that may be used in the formation of product sets. The set Λ, being a subset of the set of all value sets, is a function of x. It will be assumed that the definition of the function Λ is known to the recognition machine M. To avoid repetitious mention of Λ and of the conditions under which $K(V, W, Z)$ is defined, the following definition is introduced. A product set $D = V \cap W$ will be called a "class Z product set" if and only if: (1) all constituent value sets of D belong to Λ; (2) $P(V|Z) > 0$; (3) $P(W|Z) > 0$. We are now prepared to introduce the fundamental hypothesis about the universe U.

The Discrimination Hypothesis

There exists a $\gamma < 1$ *such* that:

1. *For all* x *and all class* X *product sets* $D = V \cap W$, $K(V, W, X) \geqq 1$.
2. *For all* x *there exists a class* $\overline{X}$ *product set* $D = V \cap W$ *such that* $K(V, W, \overline{X}) \leqq \gamma$.

A class Z product set D such that $K(V, W, Z) \leqq \gamma$ for some class Z will be called a "class Z terminal set." The gerrymandering procedure, in its ideal form, consists of deciding that x belongs to class Z if and only if it is possible to construct a class $\overline{Z}$ terminal set. According to the Discrimination Hypothesis, this procedure will yield an error probability of zero. Since the condition $K(V, W, Z) \leqq \gamma$ contradicts the independence condition $K(V, W, Z) = 1$, the features must be statistically dependent to some degree.

A simple example of a universe satisfying the Discrimination Hypothesis is shown in Figure 1, where the classes A and B consist of three vertical and three horizontal bars, respectively. This universe will be called "Universe I". The features consist of nine matrix cells, each of which may be black or white. To satisy the Discrimination Hypothesis, the set Λ will be defined as the set of all value sets corresponding to the black cells of x. In the following calculations, it will be assumed that all objects in this universe are equally probable.

The result $K(V, W, X) = 3$ is obtained for all class X product sets. For instance, let x be the leftmost vertical bar (a member of class A) and let $V = F_1$ and $W = F_4$. Then

$$K(V, W, A) = \frac{P(F_1 \cap F_4|A)}{P(F_1|A)\,P(F_4|A)} = \frac{(1/3)}{(1/3)\,(1/3)} = 3.$$

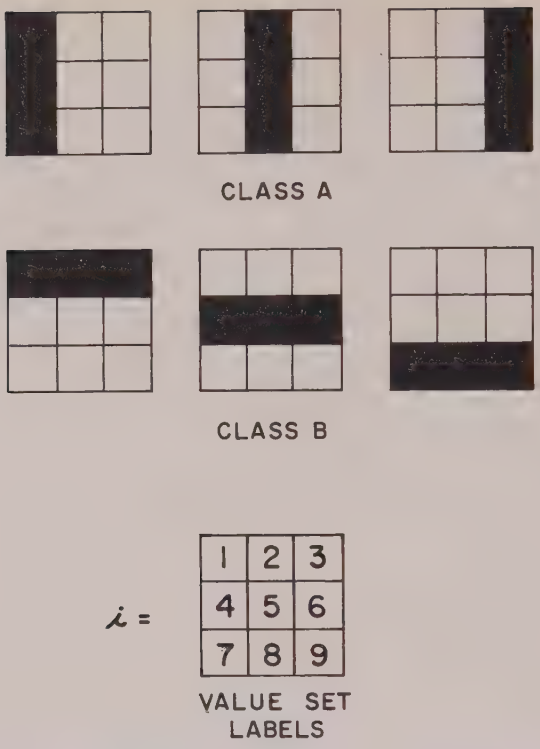


Fig. 1. Universe I.

Finally, it is easy to verify that $K(V, W, \overline{X}) = 0$ for any class $\overline{X}$ product set D. For instance, if x is the leftmost vertical bar and $V = F_1$ and $W = F_4$, then

$$K(V, W, B) = \frac{P(F_1 \cap F_4|B)}{P(F_1|B)\, P(F_4|B)} = \frac{0}{(1/3)\,(1/3)} = 0.$$

The Discrimination Hypothesis would not have been satisfied by Universe I if A had been defined to be equal to the set of all value sets. Suppose, for example, that x is the leftmost vertical bar and that $V = F_2$ and $W = F_3$ (both F_2 and F_3 correspond to the white cells of x). Then

$$K(V, W, A) = \frac{P(F_2 \cap F_3|A)}{P(F_2|A)\, P(F_3|A)} = \frac{(1/3)}{(2/3)\,(2/3)} = 3/4,$$

which contradicts part 1 of the Discrimination Hypothesis.

Let the complexity of Universe I be increased by replacing each object in it with six distinct objects formed by changing one of the six white cells into a black cell. The resulting 36-object universe will be called "Universe II." Some typical objects in this universe are shown in Figure 2. The black cells that represent changed white cells will be called "noisy" cells, while the original, unchanged cells will be called "ideal" cells. The same adjectives will be used for the corresponding value sets.

CLASS A

CLASS B

Fig. 2. Typical members of Universe II

Suppose that A is equal to the set of value sets corresponding to the black cells of x and that all objects in Universe II are equally probable. Then the smallest value of $K(V, W, X)$ for Universe II is obtained when $V = Y_1 \cap Y_2$ and $W = Y_3$, where Y_1 and Y_2 are ideal value sets and Y_3 is a noisy value set. For this case,

$$K(V, W, X) = \frac{P(Y_1 \cap Y_2 \cap Y_3|X)}{P(Y_1 \cap Y_2|X)\, P(Y_3|X)} = \frac{(1/18)}{(6/18)\,(8/18)} = 3/8,$$

which contradicts part 1 of the Discrimination Hypothesis.

Suppose now that all constituent value sets of $D = V \cap W$ are required to be ideal. Then the smallest value of $K(V, W, X)$ is obtained when $V = Y_1$ and $W = Y_2$, where Y_1 and Y_2 are ideal value sets.

In this case,

$$K(V, W, X) = \frac{P(Y_1 \cap Y_2|X)}{P(Y_1|X)\,P(Y_2|X)} = \frac{(6/18)}{(8/18)\,(8/18)} = 27/16,$$

which does satisfy part 1 of the Discrimination Hypothesis.

For any universe, let N_A be the minimum number of value sets that must be removed from Λ to eliminate all class A terminal sets, and let N_B be the corresponding number for class B. If the number N_Z is interpreted as the number of noisy value sets under the hypothesis that x belongs to class Z, it is reasonable to decide in favor of the hypothesis corresponding to the smallest number of noisy value sets. Therefore, the following rule for deciding the class of x is proposed:

The Feature Count Rule

Determine the minimum number N_A of value sets that must be removed from Λ to eliminate all class A terminal sets and the corresponding number N_B for class B. Decide $x \in A$ if $N_A \leqq N_B$ and decide $x \in B$ if $N_A > N_B$.

Suppose that x is a vertical bar (a member of class A) in Universe II. It is not difficult to show that all class A terminal sets can be eliminated by dropping the single noisy value set, while it is necessary to drop two value sets to eliminate all class B terminal sets. The latter two value sets must correspond to two ideal value sets that are not horizontally aligned with the noisy black cell. Therefore, $N_A = 1$ and $N_B = 2$, and the Feature Count Rule gives the correct decision that x belongs to class A.

TESTING THE HYPOTHESIS K(V, W, Z) $\leqq \gamma$

Let $N(Y)$ denote the number of examples in an arbitrary set Y. The test of the hypothesis $K(V, W, Z) \leqq \gamma$ must be based on the observation of the quantities $N(D \cap Z)$, $N(V \cap Z)$, $N(W \cap Z)$, and $N(U \cap Z)$. For convenience, we shall adopt the definitions

$$m_1 = N(V \cap Z) \qquad m_2 = N(W \cap Z) \qquad b = N(D \cap Z)$$

$$e = N(U \cap Z) \qquad q = P(V \cup W|Z) \qquad \lambda = K(V, W, Z).$$

Let y be the random variable of which b is the observed value, and let $P(y \leqq b|q, \lambda, m_1, m_2, e)$ be the probability that D contains b or fewer class Z examples, given q, m_1, m_2, λ, and e. Let $a = \text{Max}(0, m_1 + m_2 - e)$ and $c = \text{Min}(m_1, m_2)$. Then

$$P(y \leqq b|q, \lambda, m_1, m_2, e) = \frac{\sum_{s=a}^{s=b} R(s)}{\sum_{s=a}^{s=c} R(s)}, \tag{1}$$

where

$$R(s) = \binom{m_1}{s}\binom{e - m_1 + s}{s}\binom{e - m_1}{m_2 - s}\left[\frac{\lambda(1-q)}{1-\lambda q}\right]^{(s-a)} \tag{2}$$

Now consider the problem of testing the hypothesis $\lambda \leqq \gamma$. For known and fixed q, m_1, m_2, and e, and for any reasonable definition of the risk function, all Bayes solutions[2] of this problem are of the form "accept the hypothesis $\lambda \leqq \gamma$ if and only if $b \leqq d$," where d is some constant. To obtain a specific procedure for testing this hypothesis, we shall pretend that the composite hypothesis $\lambda \leqq \gamma$ is being tested against the simple null hypothesis $\lambda = 1$. Therefore, the following test is proposed:

The Terminal Set Acceptance Test

Accept D as a class Z terminal set (accept the hypothesis $K(V, W, Z) \leqq \gamma$) if and only if $P(y \leqq b|q, 1, m_1, m_2, e) \leqq \varepsilon$, where ε is the significance level.

Let $m_0 = \text{Min}(m_1, m_2)$ and $m_3 = \text{Max}(m_1, m_2)$. With the aid of equations (1) and (2), the Terminal Set Acceptance Test can be translated into a "counterdependence table," which shows for each possible pair (m_0, m_3) the largest value of b for which the hypothesis $\lambda \leqq \gamma$ should be accepted. Figure 3 shows two counterdependence tables, one for $\varepsilon = 1/2$ and one for $\varepsilon = 1/4$; in both tables, $e = 10$ (in practice, it would be necessary to use smaller values of ε and larger values of e). It is only necessary to tabulate the largest acceptable value of b, since if the hypothesis $\lambda \leqq \gamma$ is acceptable for a given triplet (m_1, m_2, b), it is acceptable for the same pair (m_1, m_2) and any smaller b. A counterdependence table can be constructed without knowing the value of q, since $P(y \leqq b|q, 1, m_1, m_2, e)$ is independent of q (see equation (2)).

It can be seen from Figure 3 (and more generally from equations (1) and (2)) that a small significance level in the acceptance of a product set $D = V \cap W$ as a class Z terminal set requires large values of $N(V \cap Z)$ and $N(W \cap Z)$ and a small value of $N(D \cap Z)$. See Figure 4 for an illustration of the conditions required for the acceptance of a product set

	m_3									
	1	2	3	4	5	6	7	8	9	10
1	–	–	–	–	0	0	0	0	0	–
2	–	0	0	0	0	0	1	1	–	
3	0	0	1	1	1	1	1	2	–	
4		1	1	1	2	2	3	–		
m_0 5			2	2	3	3	4	–		
6				3	3	4	4	–		
7					4	5	5	–		
8						5	6	–		
9							8	–		
10								–		

$\varepsilon = \frac{1}{2}$

	m_3									
	1	2	3	4	5	6	7	8	9	10
1	–	–	–	–	–	–	–	–	0	0 –
2		–	–	–	0	0	0	0	1	–
3			–	0	0	0	1	1	1	–
4				0	0	1	1	2	2	–
m_0 5					1	1	2	3	3	–
6						2	3	3	4	–
7							3	4	5	–
8								5	6	–
9									8	–
10										–

$\varepsilon = \frac{1}{4}$

Entries in these tables are the largest values of b for which the hypothesis $\lambda \leqq \gamma$ should be accepted.
"–" means the hypothesis cannot be accepted for any value of b.

Fig. 3. Two counterdependence tables for $e = 10$

$V \cap W$ as a terminal set. Now a large value of $N(W \cap Z)$ is likely only if $P(W|Z)$ is large. Therefore, it is desirable that $P(F_i|\bar{X})$ be large for each feature f_i. Features having this property will be said to be "weak." Even with weak features, a large value of $N(V \cap Z)$ is not likely when the constituent value sets of V are chosen at random, since a V set may be the intersection of many value sets. An important part of the algorithm presented in the next section is a procedure for constructing V sets that contain a substantial number of class Z examples.

A PROPOSED GERRYMANDERING PROCEDURE

The purpose of the algorithm given below is to attempt the formation of terminal sets for both classes. Attempts at finding class A and class B terminal sets are alternated until a bound on the computing effort is reached. Attempts to obtain a large value of $N(V \cap Z)$, where $V = G_1 \cap G_2 \cap \ldots \cap G_k$, are made by picking the value set G_j that

maximizes $N(G_1 \cap G_2 \cap \ldots \cap G_j \cap Z)$ for $2 \leqq j \leqq k$. This is only a heuristic procedure, since it is not generally true that these local maximizations of $N(G_1 \cap G_2 \cap \ldots \cap G_j \cap Z)$ will maximize $N(V \cap Z)$ for a given choice of G_1.

The Terminal Set Discovery Algorithm

Let h be the maximum number of constituent value sets of a V set, and let t be a positive integer specifying the upper bound on the computing effort. Set the Class Indicator (CI) equal to B and set $l = 1$. Execute the following procedure.

1. *Set $j = 1$. If $CI = B$ set $CI = A$ and vice versa. Let Z be the new value of CI. Choose the value set G_1 at random from the set Λ.*

2. *Let $V = G_1 \cap G_2 \cap \ldots \cap G_j$. Find all value sets W in Λ such that the product set $D = V \cap W$ passes the Terminal Set Acceptance Test. If at least one such W set is found, proceed to the next step; otherwise jump to step* **4**.

3. *Record $D = V \cap W$ as a class Z terminal set for every W set found in step* **2**.

4. *If $j = h$, jump to step* **5**; *otherwise proceed as follows. From the set or all value sets contained in Λ and not already used as constituent value sets of V, choose the set G_{j+1} that maximizes $N(V \cap G_{j+1} \cap Z)$. Replace j by $j + 1$ and return to step* **2**.

5. *Replace l by $l + 1$. If now $l > t$, stop; otherwise return to step* **1**.

The following two-step procedure is proposed for recognizing the class of an unknown object x:

1. Apply the Terminal Set Discovery Algorithm using the largest values of h and t allowed by the available computing resources;

2. Apply the Feature Count Rule.

CONCLUDING REMARKS

When the features are statistically independent the error rate is minimized by the following rule: Decide $x \in A$ if $Q(x) \geqq 1$, and decide $x \in B$ if $Q(x) < 1$, where

$$Q(x) = \frac{P(A|U)}{P(B|U)} \prod_{i=1}^{n} \frac{P(F_i|A)}{P(F_i|B)}$$

It is known[3] that satisfying one of the two relations $Q(x) > 1$ and $Q(x) < 1$ is equivalent to locating x on one side or the other of a hyperplane Δ in n-dimensional space, and that satisfying the relation $Q(x) = 1$ is equivalent to locating x on Δ. The degree of confidence associated with the corresponding decisions depends upon the distance of x from Δ, the confidence being high when x is far from Δ and low when x is near

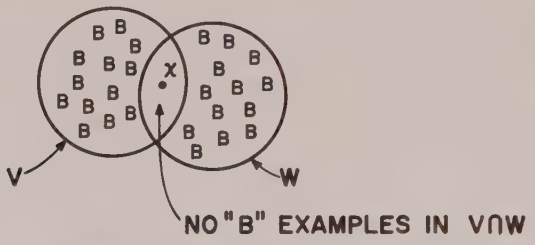


Fig. 4. A Typical Condition for Accepting the Hypothesis $K(V, W, B) \leqq \gamma'$ (for Deciding that $x \varepsilon A$).

to Δ. This simple situation is sometimes taken as an ideal when evaluating features. A set of features is commonly considered to be "powerful" if the two classes can be separated by a simple hypersurface and if the average distance of an object from the hypersurface is large.

Now the gerrymandering procedure will work when $P(F_i|A) = P(F_i|B)$ for all i and $P(A|U) = P(B|U)$, i.e., when x lies *on* the hyperplane (see, for example, Universes I and II). But the gerrymandering procedure, will not work well when x is far from Δ, since then the features do not tend to be weak. These observations suggest that the invention of powerful features is not a necessary first step in the development of an automatic classification system employing the gerrymandering procedure. However, this procedure must be supplemented by other schemes such as the hyperplane rule because features will not always be weak.

ACKNOWLEDGMENTS

This research was sponsored by the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under contract AF 33(615)-

1764. Further reproduction is authorized to satisfy needs of the U. S. Government.

This work has benefited greatly from the suggestions and criticisms of Saul Amarel, Jack Sklansky, Robert Winder, and Arthur Zingher.

REFERENCES

1. Sebestyen, G. "Pattern Recognition by an Adaptive Process of Sample Set Construction," *IRE Transactions on Information Theory*, **8**, No. 5, pp. 82–91, September, 1962.
2. Wald, A. *Statistical Decision Functions*, John Wiley and Sons, New York, Chapter I, 1950
3. Nilsson, N. J. *Learning Machines*, McGraw-Hill, New York, pp. 47–50, 1965.

C. R. LEGÉNDY *

United Aircraft Research Laboratories

East Hartford, Connecticut

How Large are Hebb's Cell Assemblies?

ABSTRACT

Hebb's cell assemblies, if they exist, probably contain at least several hundred or several thousand neurons. This suggestion is founded on the premise that cell assemblies must be readily interconnectible, and they are so only if interconnecting them requires merely reinforcing of already existing synapses, rather than a search for internuncial neurons or the growing of new fibers. A simple probability theoretic calculation (implicitly containing the assumption of initially and locally random connections between neurons), together with some reasonable figures for the number of synapses on the average neuron, etc, yields the numerical estimate quoted above.

Hebb[1] proposed that the effect of experience on the cortex may be to organize neurons into *cell assemblies*. Cells in these assemblies contain information in form of the sheer fact of belonging together, the physical manifestation of their alliance being a wealth of connections, making the cell assemblies capable of self-reinforcing, reverberatory activity.

The existence of reverberating circuits in the brain has been established for a long time[2], well before the work of Hebb. Recent microelectrode studies of the visual cortex of the cat by Hubel and Wiesel[3] revealed no suggestion of cell assemblies in Brodmann's areas 17, 18 and 19; on the other hand, it can be argued that the visual detection mechanisms discovered by Hubel and Wiesel are apparently innate, whereas evidence for or against Hebb's hypothesis must come from those structures whose development is prevented by early sensory deprivation.[4,5]

In the following, I would like to add a piece of "theoretical data" in favor of the cell assembly concept by pointing out, using probability theory, that cell assemblies need not be intolerably large before they have so many afferents and efferents that direct synaptic contact exists, with great certainty, between all cell assemblies within the same cortical area. (The calculation below yields for the size at which this happens one of the order of 10^3–10^4 neurons.)

* Present address: Westinghouse Cambridge Laboratory, Cambridge, Massachusetts.

The observation makes the concept of cell assemblies attractive because in view of it cell assemblies can be functionally interconnected very fast. Even though we do not know how long it takes to reinforce a synapse, we can safely assume that it takes a much shorter time than it does to make contact either through a chain of internuncial cells[2] or by means of tropistic growth.[6] On the other hand, important sensory events often do not give the organism a second chance to observe them, and it stands to reason that Nature would make its mechanisms for recording such events as fast as it can. In fact, the human brain is known to be able to comprehend and retain complex events on a single contact.

For the purposes of obtaining numerical results, I shall make the assumption that the members of cell assemblies are randomly scattered throughout the cortical area containing them, so that cell assemblies make contacts evenly and the chances of a given cell being in contact with a given assembly do not depend on the location of the cell within the area.

It is necessary to note that if the cell assemblies were innate structures, such a "random connection scheme" would be quite an inefficient way of ensuring contacts. The efficient way then would be simply to connect the assemblies that must be connected, from birth. This is indeed the way in which Nature connects the various functionally and anatomically distinct cortical and subcortical structures. However, the very essence of the cell assembly concept is that cell assemblies are not innate structures, therefore the brain does not know *a priori* which cells are to assemble with which cells; much less does it know which assemblies are to require contact with which assemblies. Hence the assumption of initially and locally random connections.

I shall also assume that connections within the cell assembly only take up a small fraction of the synapses of the members; therefore, to a reasonably good approximation, the number of cells in synaptic contact with the assembly is given by the product of the number of cells in the assembly and the average number of cells in synaptic contact with each cell.

The calculation I propose is then extremely simple:

Let us say that some cell assembly has n_1 members, each member is in synaptic contact with S cells in the cortical area considered, and this cortical area contains N cells altogether. Then, by the above assumptions, any cell outside the assembly but within the cortical area has a proba-

bility $n_1 S/N$ of having synaptic contact with the assembly, and a probability $1 - (n_1 S/N)$ of not having such synaptic contact. If there is another cell assembly in the same cortical area, having, say n_2 members, the probability that neither of these is in synaptic contact with the first assembly is $P = [1 - (n_1 S/N)]^{n_2}$; this then is the probability that the two assemblies are not in synaptic contact.

The latter formula can be written in a very convenient form by recalling that, if the quantity α is much greater than 1, the expression $[1 - (1/\alpha)]^{\alpha}$ is approximately independent of α and equal to $1/e = 1/2.7182\ldots$, and noting that for the cases of interest $N/n_1 S$ is much greater than 1. For the formula for P can be re-written as $P = [1 - (1/\alpha)]^{\alpha n_1 n_2 S/N}$, where for emphasis, α has been written instead of $n_1 S/N$ in some places; thus we have

$$P = e^{-n_1 n_2 S/N} \tag{1}$$

This probability is small, i.e. the two assemblies most probably are in synaptic contact, if $n_1 n_2 S/N$ is much greater than 1. Computation shows that when $n_1 S/N$ is not much smaller than 1, the probability P is even smaller than shown by the approximation (1).

The result is that *the cell assemblies can be expected to be in mutual contact if they all have well over* $\sqrt{N/S}$ *members.* With this information at hand, we return to the question of this talk: How large are Hebb's cell assemblies?

We reason: the brain must be able to interconnect the elementary building blocks of its mental entities whenever information received from the sense organs dictates that it should. Assuming that Hebb's cell assemblies are the elementary building blocks in question, we conclude that Hebb's cell assemblies must be large enough to be in mutual contact; thus they must contain well over $\sqrt{N/S}$ neurons. (We remark that the notion that any given pair of single neurons within a cortical area can be expected to have direct synaptic contact is a misconception; contact between single neurons can be expected to exist only through internuncial neurons.)

For instance, if all assemblies have $10\sqrt{N/S}$ members, the formula (1) gives $P = e^{-100} \cong 10^{-43}$, i.e. the odds in favor of mutual contact are about 10^{43} to 1. On the other hand, if all have only $0.1\sqrt{N/S}$ members, $P = e^{-0.01} \cong 1 - 0.01$, i.e. the odds are turned to 100 to 1 against mutual contact. If the number of cells in the cortical area is 10^8 and

each makes synapses with 10^3 cells, $10\sqrt{N/S}$ is about 3000. It is easy to show that the expected number of synapses connecting any two assemblies in the latter case is 100. (In general it is $n_1 n_2 S/N$.)

By the way, it is possible for the brain to interconnect cell assemblies smaller than the above, without growing new fibers or searching for internuncial neurons, if the number of assemblies to be interconnected is not 2, but some larger number, say, a few dozen. I hope to elaborate on this point in a separate communication.

It is a special privilege to thank Dr. W. S. McCulloch for several illuminating discussions on the subject and his kindness of reading the manuscript.

REFERENCES

1. Hebb, D. O., *The Organization of Behavior* (John Wiley and Sons, Inc., New York, 1949).
2. Lorente de Nó, R., J. Neurophysiol., **1**, 207–244 (1938).
3. Hubel, D. H , and Wiesel, T. N. J. Neurophysiol. **28**, 229–289 (1965).
4. von Senden, M., *Raum- und Gestaltauffassung bei operierten Blindgeborenen vor und nach der Operation* (Barth, Leipzig, 1932).
5. Riesen, A. H., Science, **106**, 107–108.
6. Kappers, C. U. A., J. Comp. Neurol., **27**, 261 (1917).

RICHARD BELLMAN

University of Southern California

Los Angeles, California

ROBERT S. ROTH

Structures Department, Avco Corporation

Wilmington, Massachusetts

A Technique for the Analysis of a Broad Class of Biological Systems

INTRODUCTION

The mathematical modelling of complex biological systems has become increasingly important with the development of both analog and digital computers. With larger memory capacity and more sophisticated numerical techniques, mathematicians and biologists should be able to study more closely some of the intricate biological processes occurring in nature.

This paper will be restricted to the examination of a class of biological systems whose dynamic behavior may change radically at certain critical times. A somewhat similar biological system was studied by Heinmets[1] in which he set up a system of nonlinear differential equations describing the process of induced enzyme synthesis. Using an analog computer he was able to solve the equations when both the initial conditions and system parameters were assumed. By choosing the parametric constants very carefully from experimental results[2] he was able to explore, mathematically, the basic mechanism of enzyme synthesis.

The mathematical models, such as the one derived by Heinmets, are predicated on the assumption that all subprocesses began operation at the initial time zero, vary continuously over the given time span, and do not change their basic behavior pattern. By this we mean that the constants of the differential equations are invariant over the time span.

It is the basic intention of the paper to develop a mathematical technique whereby the assumptions of the above paragraph may be relaxed; that is to say, all subprocesses need not begin at time zero, need not vary continuously over the entire time span and may change their basic behavior patterns at certain critical, but unknown times. It is hoped that this technique may enable biologists to better model their system.

As will be shown in the development, the experimental data will play a central role in the theory. In this philosophy, the data will not be used as a verification of the theory but will help form its basic structure.

The Mathematical Model

Let us consider a biological system which will operate over a time span $[0, T]$. Let the system be comprised of N subprocesses denoted by $x_1(t), x_2(t), \ldots x_N(t)$.

The Differential Equations

The set of differential equations (the model) used to describe the time dependent behavior of the subprocesses of the biological system is assumed to have the following form

$$\dot{x}_i(t) = \sum_{\substack{k,j=1 \\ i=1,N}}^{N} \left(a_{ijk}x_j(t)\, x_k(t) + b_{ij}x_j(t)\right) \tag{1}$$

where $(\dot{})$ is the derivative with respect to time t and $\{a_{ijk}, b_{ij}\}$ are the coupling coefficients of the system. The set of equations (1) represents an abstract generalization of the Heinmets model in Reference 1.

The Critical Times

Let us assume that the entire time interval $[0, T]$ is divided into P subintervals such that in any subinterval no equation constant changes, that is, there exists, a set of critical times $\{t_p\}$ at which at least one equation constant suddenly changes its value. (A sudden change in an equation constant indicates that the solution will be continuous at that point but its velocity will change.) These critical times are considered unknowns of the problem and must be determined.

The Experimental Data I

As mentioned above, the experimental data will be used in developing the structure of the theory. Fundamentally, the equation constants and the initial conditions of the system will be determined in such a way as to "best" fit the data in the least square sense. Let $d_i(t_j)$ be the appropriate observation of subprocess x_i at time t_j.

A measure of the error between the solution of the problem and the observations taken at Nk times within the time interval is given as

$$E = \sum_{k=1}^{Nk} \left\{ \sum_{i=1}^{N} (x_i(t_k) - d_i(t_k))^2 \right\} \tag{2}$$

The Mathematical Problem

Solve the system of differential equations

$$\left.\begin{aligned} \dot{x}_i(t) &= \sum_{k,\,j=1}^{N} (a_{ijk} x_i(t)\, x_j(t) + b_{ij} x_j(t)) \\ \dot{x}_i(0) &= \bar{X}_i \quad i = 1, \ldots, N \end{aligned}\right\} \tag{3}$$

subject to the conditions, that the initial conditions $\bar{x}_i$, the equation constants a_{ijk}, b_{ij} and the critical times T_l are chosen such that

$$E = \sum_{k=1}^{Nk} \left\{ \sum_{i=1}^{N} [x_i(t_k) - d_i(t_k)]^2 \right\} \tag{4}$$

is a minimum over the entire time span $[0, T]$.

Technique of Solution

We shall consider the solution of the set of equations (2) and (3) in two parts. First we solve the equations in the special case of no critical times within the interval [0, T], then we shall consider the problem of finding the critical times.

The technique of solution will be that of segmental differential approximation introduced in 1964 by Bellman, Gluss and Roth[3-6]. It has been applied to problems of mechanics and neurology by Roth[6,7] and recently to the problem of the analysis of a metabolic process[8]. The

technique is a blend of differential approximation[9] and dynamic programming[10], and uses the ideas of quasilinearization[9].

Differential Approximation

Consider the problem of solving the system of equations

$$\left.\begin{aligned} \dot{x}_i(t) &= \sum_{k,\,j=1}^{N} [a_{ijk}x_j(t)\,x_k(t) + b_{ij}x_j(t)] \\ x_i(0) &= \bar{x}_i \quad i = 1, \ldots, N \end{aligned}\right\} \tag{5}$$

where $\bar{x}_i$, a_{ijk}, b_{ij} must be chosen such that the error

$$E = \sum_{i=1}^{Nk} \left\{ \sum_{j=1}^{N} (x_j(t_i) - d_j(t_i))^2 \right\} \tag{6}$$

$$0 < t_1 < t_2 < \ldots < t_k < t_{Nk} \leqq T$$

is minimized.

In the set of equations (5), the constants a_{ijk}, b_{ij} can be thought of as functions of time t satisfying the differential equations,

$$\left.\begin{aligned} \dot{a}_{ijk} &= 0 \\ \dot{b}_{ij} &= 0 \\ a_{ijk}(0) &= \bar{a}_{ijk} \\ b_{ij}(0) &= \bar{b}_{ij} \end{aligned}\right\}. \tag{7}$$

Then the entire system has the form,

$$\dot{x}_i = \sum_{k,\,j=1}^{N} [a_{ijk}x_j(t)\,x_k(t) + b_{ij}x_j(t)] \tag{8}$$

$$\dot{a}_{ijk} = 0$$

$$\dot{b}_{ij} = 0$$

$$x_i(0) = \bar{x}_i$$

$$a_{ijk}(0) = \bar{a}_{ijk}$$

$$b_{ij}(0) = \bar{b}_{ij} \quad i = 1, \ldots, N$$

In convenient vector notation

$$\left.\begin{aligned} &\dot{\bar{y}} = \bar{G}(\bar{y}) \\ &\text{where} \\ &\bar{y} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \\ a_{111}(t) \\ \vdots \\ a_{NNN}(t) \\ b_{11}(t) \\ \vdots \\ b_{NN}(t) \end{bmatrix} \end{aligned}\right\} \tag{9}$$

If there are N basic unknowns in the problem, and all coupling coefficients are present, then system (9) represents a set of differential equations whose size is given as

$$\frac{N^2(N+1)}{2} + N^2 + N = Nc \tag{10}$$

Quasilinearization

The system given by (9) is nonlinear and is in a form to apply the technique of quasilinearization[9]. Let us introduce a sequence of functions $\bar{y}^{(1)}(t), \bar{y}^{(2)}(t), \ldots, \bar{y}^{(n)}(t)$ such that if $\bar{y}^{(1)}(t)$ is a first approximation to the solution of the system, then the n^{th} approximation is found by the solution of the linear system

$$\dot{y}_i^{(n)}(t) = G_i\left(\bar{y}^{(n-1)}(t)\right) + \sum_{j=1}^{Nc} \frac{\partial G_i(\bar{y}^{(N-1)})}{\partial y_j} [y_j^{(N)}(t) - y_j^{(N-1)}(t)] \tag{11}$$

where Nc is the number of equations in the set. Such an iteration process, a Newton Ralphson analog in function space, is a second order convergent process[11].

Numerical Procedure

The equations (11) form a set of first order, linear ordinary differential equations. Any solution can be written in the form,

$$\bar{y}^{(N)}(t) = \bar{P}(t) + \sum_{k=1}^{Nc} C_k \bar{H}^k(t) \tag{12}$$

where $\bar{P}(t)$ is the particular vector solution of (11) and $\bar{H}^K(t)$ are the homogeneous vector solutions.

Both solutions $\bar{P}$ and $\bar{H}^K$ can be generated numerically by the following scheme:

For the particular solution, solve the set of equations

$$\dot{y}_i^{(n)}(t) = G_i\left(\bar{y}^{(N-1)}(t)\right) + \sum_{j=1}^{Nc} \frac{\partial G_i}{\partial y_j}\left(\bar{y}^{(n-1)}(t)\right) \cdot \left[y_j^{(n)}(t) - y_j^{(N-1)}(t)\right] \tag{13}$$

$$i = 1, \ldots, Nc$$

with the initial condition, $y_i^{(n)}(0) = 0$ for all i.

Each homogeneous solution is the solution of the set

$$\dot{y}_i^{(n)}(t) = \sum_{j=1}^{Nc} \frac{\partial G_i\left(\bar{y}^{(n-1)}(t)\right)}{\partial y_j} y_j^{(n)}(t) \tag{14}$$

with the initial conditions given as follows:

$$\bar{H}^{(1)}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \bar{H}^{(2)}(0) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \ldots \bar{H}^{(Nc)}(0) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{15}$$

Once the particular and homogeneous solutions are calculated, the constants C_k of equation (12) must be found such that the solutions x_i best fit the observed data in the least square sense. Substituting the expression (12) into the measure of the error, (6), gives

$$E = \sum_{l=1}^{Nk} \left\{ \sum_{i=1}^{Nc} \left[\left(P_i(t_l) + \sum_{k=1}^{Nc} C_k H_i^k(t_l) \right) - d_i(t_l) \right]^2 \right\} \tag{16}$$

The quadratic form E may be minimized by setting

$$\frac{\partial E}{\partial C_m} = 0$$

The normal equations for determining C_k are

$$\sum_{l=1}^{Nk} \left\{ \sum_{i=1}^{Nc} \left[P_i(t_l) + \sum_{k=1}^{Nc} C_k H_i^k(t_l) - d_i(t_l) \right] H_i^M(t_l) \right\} = 0$$

$$M = 1, 2, \ldots, Nc$$

or

$$\sum_{k=1}^{Nc} C_k \left\{ \sum_{l=1}^{Nk} \sum_{i=1}^{Nc} H_i^k(t_l)\, H_i^M(t_l) \right\}$$

$$+ \left\{ \sum_{l=1}^{Nk} \sum_{i=1}^{Nc} [P_i(t_l) - d_i(t_l)]\, H_i^{\mathbf{M}}(t_l) \right\} = 0 \tag{17}$$

if we define the matrix A_{kM} as

$$A_{kM} = \sum_{l=1}^{Nk} \sum_{i=1}^{Nc} H_i^k(t_l)\, H_i^M(t_l) \tag{18}$$

and the vector by b_M as

$$b_M = \sum_{l=1}^{Nk} \sum_{i=1}^{Nc} (P_i(t_l) - d_i(t_l))\, H_i^M(t_l) \tag{19}$$

then C_k is determined by the equation

$$\sum_{k=1}^{Nc} A_{kM} C_k + b_M = 0 \tag{20}$$

$$M = 1, \ldots, Nc$$

Once C_k is found, the n^{th} approximation to the solution of (1) is given by Eq. (12), namely

$$\bar{y}^n(t) = \bar{P}(t) + \sum_{k=1}^{Nc} C_k H^k(t) \tag{21}$$

Having $\bar{y}^n(t)$, the next iteration function $\bar{y}^{n+1}(t)$ can then be generated. The process is continued until proper convergence is attained. The process is begun by choosing a set of general initial conditions and directly integrating Eq. (1).

The convergence criterion is determined by examining the sequence of C_i's generated by Eq. (20) for each iteration. A careful examination of the C_i's will show that they are the generalized initial conditions of the system, that is, the first N are the physical initial conditions and the remaining set represent the physical constants.

The problem is considered solved when that iteration is reached for which the generalized initial conditions properly converge.

Dynamic Programming

We are now in a position to examine the problem of determining the critical times at which the system will change its dynamic behavior. To be precise, let $\{t_p\}$ be a set of times within the interval $[0, T]$ which

will include the critical times as a subset $\{t_{pk}\}$. The problem is to determine $\{t_{pk}\}$.

It is clear from the last section that if a set of data is observed during a time interval t_i to t_j, and the form of the differential equations is specified, then the generalized initial conditions (the initial conditions plus system parameters) and the associated error can be found. Also if the set of times $\{t_p\}$ is the subset $\{t_{pk}\}$, then the error made by fitting the data over the entire time span should be a minimum. This suggests a method for finding the subset $\{t_{pk}\}$ namely choose the set $\{t_p\}$, determine the associated errors for all combinations of time intervals, and the search for those consistent times for which the total error is a minimum. If the set $\{t_p\}$ is large, the search procedure, even on a machine, may be enormous. The technique of dynamic programming provides an efficient procedure by which to make the search.

Basically, the idea behind the search technique of dynamic programming is this: if within the time span $[0, T]$ p times have been found to minimize the entire error, then the proper choice of the $(p + 1)$ time must further minimize the error, and its value will not affect the values of the p previously chosen times in any way.

Let us define:

$E(t_j, t_k)$ = the error calculated when the system of differential equations is fit to data observed between times t_j and t_k where t_j, t_k are members of the set $\{t_p\}$.

Then for any subset of times $\{t_{pk}\}$ of $\{t_p\}$ spanning the time from 0 to T, the total error is given by

$$E(t_{p_1} \ldots t_{pN}) = E(0, t_{p_1}) + E(t_{p_1}, t_{p_2}) + \ldots + E(t_{p_{N-1}}, t_{p_N}) + E(t_{p_N}, T).$$

The critical times $\{t_{pk}\}$ is the smallest subset of $\{t_p\}$ which will minimize E.

These critical times are found by utilizing the fundamental equation of dynamic programming. The derivation is briefly outlined below. The solution of the equation will yield the critical times of the system.

Consider the function

$$E(0, t_{p_1}) + E(t_{p_1}, t_{p_2}) + \ldots + E(t_{p_N}, T) \tag{22}$$

The critical times $\{t_{pk}\}$ are those for which

$$E = \min_{\substack{(\text{all } t_{p_k}) \\ k=1, \ldots, N}} [E(0, t_{p_1}) + E(t_{p_1}, t_{p_2}) + \ldots + E(t_{p_{N-1}}, t_{p_N}) + E(t_{p_N}, T) \tag{23}$$

Because expression (23) is linear,

$$E = \min_{\left(t_{p_k},\ k=1,\ N\right)} \{E(0, t_{p_1}) + E(t_{p_1}, t_{p_2}) + \ldots E(t_{p_{N-1}}, t_{p_N})\} + \min_{t_{p_N}} E(t_{p_N}, T)$$

$$= \min_{(t_{p_N})} \left[\min_{\left(t_{p_k},\ k=1,\ N-1\right)} \{E(0, t_{p_1}) + \ldots + E(t_{p_{N-1}}, t_{p_N})\} + E(t_{p_N}, T) \right] \quad (24)$$

The problem posed by the bracketed quantity in (24) is precisely the same as (23) with the exception that the set of points $\{t_{pk}\}$ have been reduced by one.

Define

$F_M(t)$ = the minimum error resulting from fitting differential equations to data observed over the time interval $[0, t]$ if M critical times are allowed.

Then by the arguments used in defining (24),

$$F_M(t) = \min_{\{t_q < t\}} [F_{M-1}(t_q) + E(t_q, t)]$$

$$F_1(t) = E(0, t); \quad t_q \varepsilon \{t_p\} \quad (25)$$

This is the fundamental equation of dynamic programming[10].

For the solution we are interested in

$$E = F_M(T) \quad (26)$$

where we are allowing M critical times to occur. $F_M(T)$ is calculated by the recursion notation (25) and the critical times are found as by-products.

Numerical Examples

Two numerical examples have been chosen from Reference (6), which will illustrate the ideas of differential approximation and the method of finding the critical times.

a. *The Van der Pol Equation.* Let the system be governed by a single Van der Pol equation in the time interval $[0, T]$. The Van der Pol equation is

$$\ddot{x}(t) + \lambda_1[1 - x(t)^2]\,\dot{x}(t) + \lambda_2 x(t) = 0$$

$$x(0) = c_1 \quad (27)$$

$$\dot{x}(0) = c_2$$

Using observations * on $x(t)$, determine λ_1, λ_2, C_1 and C_2. The equation (27) may be rewritten

$$\begin{aligned} \dot{x} &= u & x(0) &= c_1 \\ \dot{u} &= -[\lambda_1(1 - x^2)\, u + \lambda_2 x] & u(0) &= c_2 \\ \dot{\lambda}_1 &= 0 & \lambda_1(0) &= c_3 \\ \dot{\lambda}_2 &= 0 & \lambda_2(0) &= c_4 \end{aligned} \tag{28}$$

The quasilinear equations are

$$\begin{aligned} \dot{x}^{N+1} &= u^{N+1} \\ \dot{u}^{N+1} &= -\{[\lambda_1^N(1 - (x^N)^2)\, u^N + \lambda_2^N x^N] \\ &\quad + [-2x^N\lambda_1^N u^N + \lambda_2^N]\,(x^{N+1} - x^N) \\ &\quad + \lambda_1^N[1 - (x^N)^2]\,(u^{N+1} - u^N) \\ &\quad + [(1 - (x^N)^2)\, u^N]\,(\lambda_1^{N+1} - \lambda_1^N) \\ &\quad + [x^N]\,(\lambda_2^{N+1} - \lambda_2^N)\} \end{aligned} \tag{29}$$

The results are shown in Table I.

Table I: Results of Numerical Studies

Initial *Approximation*	1st *Iteration*	2nd *Iteration*	3rd *Iteration*	4th *Iteration*
0.9968	0.9924	0.9945	1.0137	1.0016
—0.0813	—0.1271	—0.2005	—0.1357	—0.1993
10.1000	9.0068	7.9824	9.4799	8.1557
1.1000	0.9545	0.8385	0.9076	0.8339

5th *Iteration*	6th *Iteration*	7th *Iteration*	8th *Iteration*	True *Values*
0.9840	0.9913	0.9943	0.9977	1.0000
0.0073	—0.1062	—0.1130	—0.1252	0.0000
10.7041	9.5735	9.6453	9.5563	10.0000
1.0480	0.9724	0.4830	0.9480	1.0000

* In the numerical cases, the observations are generated by solving equation (27) for a given set λ_1, λ_2, c_1 and c_2

b. *Suddenly Changing Frequencies and Amplitudes of a Harmonic Oscillation.* Let the system be governed by the equation,

$$\ddot{x}(t) + K(t)\,x(t) = 0$$
$$x(0) = c_1$$
$$\dot{x}(0) = c_2 \tag{30}$$

and let the system be observed over a period of time as recorded in Table II. Then what are the initial conditions, the precise form of $K(t)$,

Table II. Harmonic Motion, Nonlinear Spring Data Record

Time	*obs.*	*Time*	*obs.*
0.000	0.0000	4.867	1.7925
0.157	0.5355	5.024	1.3985
0.314	0.9296	5.181	—.9265
0.471	0.6880	5.338	—1.9722
0.628	0.6384	5.495	—.2948
0.785	0.0643	5.652	1.7897
0.941	—.5342	5.809	1.4030
1.099	—.9290	5.996	.5433
1.256	—.9692	6.123	—.3010
1.413	—.3119	6.280	0.0000
1.570	—.0659	6.437	0.0747
1.727	0.5329	6.584	0.1525
1.884	0.9284	6.752	0.2993
2.051	0.9696	6.908	0.2946
2.198	0.6408	7.065	0.3516
2.355	0.0675	7.222	0.4038
2.512	—.5315	7.379	0.4456
2.669	—.9278	7.536	0.4745
2.826	—.9700	7.693	0.4936
2.983	—.6421	7.850	0.5001
3.140	—.0691	8.007	0.4936
3.297	1.7981	8.164	0.4745
3.454	1.3893	8.321	0.4456
3.611	—.9380	8.478	0.4038
3.768	—1.9700	8.635	0.3516
3.925	—.2821	8.792	0.2946
4.082	1.7953	8.949	0.1525
4.239	1.3939	9.106	0.0747
4.396	—.9322	9.263	0.0001
4.553	—1.9711	9.420	End
4.710	—.2885		

and the critical time intervals which will produce a solution best fitting the data in the least square sense.

For details of the differential approximation see Reference (7).

The time span was between $t = 0$ and $t = 9.42$. The set $\{t_p\}$ was taken to be

$$t_p = 1.57\,p : p = 1, 2, \ldots 6$$

The number of allowable system changes was set at 3. The solution is found in Table III.

Table III

Beginning Time	Final Time	General Initial Conditions c_1	c_2	K
0.00	3.14	0.615×10^{-6}	3.99	16.00
3.14	6.28	—0.03	16.01	64.06
6.28	9.42	—0.002	0.500	1.001

Experimental Data II

The application of this technique to a biological system, such as the Heinmets model for induced enzyme synthesis, depends crucially on the available experimental data. For a larger system with many subprocesses, it is often impossible to measure each subprocess. By a careful redefinition of the error measurement (6) an approximate solution to the problem may be found using only partial information. Conversely if certain information is known about the system, the set of unknowns may be reduced.

The accuracy of the data is important. If the system is observed over a long period of time, perhaps less accuracy is required for each observation. Thus the technique can be used to formulate criteria for the proper design of experiments.

The mathematical model forms the basis of the technique, yet in practice such a model may be unknown. The data may be fit to many models and the one producing the minimum error may be considered the "best".

CONCLUSIONS

The underlying reason for analyzing a biological system, or any system, is to try to understand the system structure. The technique described above is intended to select the most reasonable mathematical

structure which will fit the observed behavior of the system. To the practicing biologist having the mathematical model may not be enough and the final step is to explain the newly found mathematical model in physical terms. This task may be quite formidable and will vary with the problem but, after all, the development of a mathematical theory to explain a physical phenomenon is still an art, its solution is a science.

REFERENCES

1. Heinmets, F. "Analog Computer Analysis of a Model-System for the Induced Enzyme Synthesis," *J. Theor. Biol.* (1964), **6**, 6–75.
2. Pardee, A. B., and Prestidge, L. S. (1961) *Biochem. Biophys. Acta.*, **49**, 77.
3. Bellman, R. E., Gluss, B., and Roth, R. S. "Segmental Differential Approximation and the Black Box Problem." The Rand Corporation, RM-4269, PR (Oct. 1964).
4. Bellman, R. E., Gluss, B., and Roth, R. S. "On the Identification of Systems and the Unscrambling of Data: Some Problems Suggested by Neurophysiology," The Rand Corporation, RM-4266-PR, (Oct. 1964).
5. Bellman, R. E., Gluss, B., and Roth, R. S. "Identification of Differential Systems with Time Varying Coefficients," The Rand Corporation, RM-42-88-PR (Nov. 1964).
6. Bellman, R. E., Gluss, B. and Roth, R. S. "On the Identification of Systems and the Unscrambling of Data: Some Problems suggested by Neurophysiology," *Proc. Nat. Acad. Sci. U.S.* **52**, No. 5, 1239–1240 (1964).
7. Roth, R. S., "The Unscrambling of Data: Studies in Segmental Differential Approximation," Avco/RAD, TM-65-22 (May 1965). Also J. Math. Anal. Appl. **14**, 5–22 (1966).
8. Bellman, R. E., and Roth, R. S. "Segmental Differential Approximation and Biological Systems: An Analysis of a Metabolic Process," *J. Theor. Biol.* (1966), **11**, 168–176.
9. Bellman, R. E., Kagiwada, H., and Kalaba, R. "Quasilinearization, System Identification, and Prediction," The Rand Corporation, RM-3812-PR (Aug. 1963).
10. Bellman, R. E. "Dynamic Programming," Princeton Univ. Press., 1957.
11. Kalaba, R. "On Nonlinear Differential Equations. The Maximum Operation and Monotone Convergence," *J. Math. Mech.* **8**, 519–574 (1959).

F. H. GEORGE

University of Bristol and Educational and Scientific Developments Ltd.

Bristol, England.

Hypothesis Confirmation on a Digital Computer

This paper is intended to outline the problem of both formulating and confirming hypotheses on a digital computer, and relating the results to heuristic programming. This is a first step towards tackling the problem of using a computer as a hypothesis maker in the context of artificial intelligence, where the total picture is one of a computer program for both "on line" and "off line" computing. This type of program must allow inductive and deductive reasoning to occur and will deal in ordinary language (e.g. a natural language such as English) for question-and-answer purposes. (George, 1965, 1966).

There are many different approaches to this problem, but this particular approach does not at this stage make use of the corporate knowledge of probability theory and the usual arguments about the basis of induction. This is because it is assumed that inverse probability, Bayes Theorem, and other well defined aspects of probability are available to the computer via the programmer, and the computer's problem here is to recognize which method is appropriate to which problem; although the difficult problem of recognition, and indeed risk analysis too, will not be discussed here. We shall concentrate here on formulating inductive inferences (hypotheses) based on concepts or terms. This is a process which is essential to the logical capacity of a computer dealing in natural language. First then we shall look at what we have called concepts.

CONCEPTS

We think of classes as being composed of members, so we say $x \in a$ where x is a member (or an individual) of (belonging to) the class a. So we shall say classes are depicted by lower case letters: —

$$a, b, \ldots, n$$

with or without suffixes, and individuals by lower case letters at the end of the alphabet: —

$$x, y, \ldots, w$$

with or without suffixes. We must next say that an individual is composed of a set of properties, each property being a member of its own class, so while we write $x \in a$ or $y \in b$ we also write: –

$$x = a \cdot b \cdot c \ldots n$$

or in Polish notation

$$x = AA \ldots A\, ab \ldots n \tag{1}$$

Where "·" or A is *conjunction* and where we mean x to be, say, a "green, round, long, ..., elephant".

We now say that concepts are represented by classes and named by the class name, which are sometimes called "terms". So "Scots", "Irish", "tall people", etc, are all terms, and the structures into which these terms can be fitted are the range of statements capable of being made in a Indo-European subject-predicate language such as English. We shall in fact in this article limit our analysis to the syllogistic forms:—

"All x is a"

"No x is a"

"Some x is a"

"Some x are not a"

This loss of generality allows us to concentrate our thoughts to a greater extent on the methods being used, where we can easily see how to generalize the argument to other linguistic forms, where semantic rules may become relevant. Let us now look at hypothesis formation.

HYPOTHESIS FORMATION

The flow chart in figure 1 shows the steps envisaged as building hypotheses from concepts. We must try to describe the actual process.

We assume we have some existing concepts $C_1, C_2, \ldots, C_n$ which are named by words or phrases and are, in the sense of the syllogism, simply called terms. Our problem now becomes one of substituting relevant terms, in turn, in to the relevant statement forms.

Our criteria of relevance present a major problem, but we certainly mean to include those H's involving the two terms C_i and C_j of the question. We will also wish to include, as needed, all terms which are *associated with* C_i and C_j. This carries the (relevance) search systematically beyond the immediate and obvious.

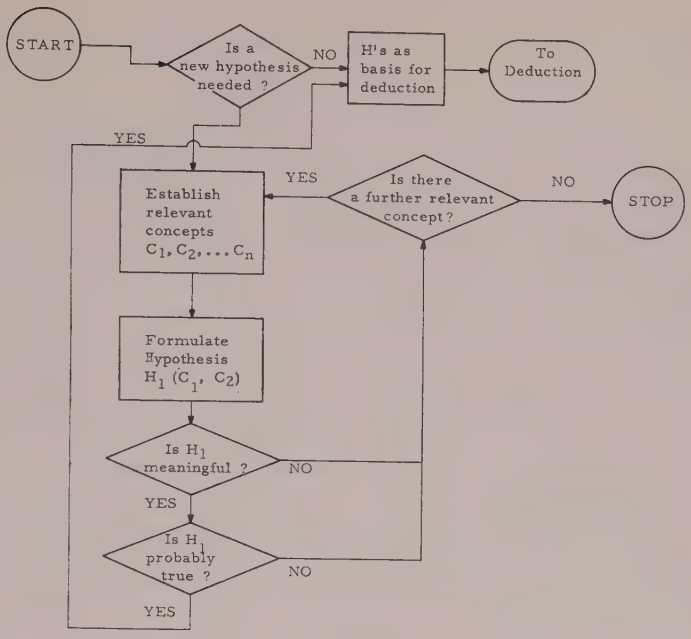


Fig. 1

We can now write into the frame "all x is a"

"all x is b"

"all y is b"

"all x is c"

"all y is c"

and so on and so forth. We have of course assumed for the sake of this article that the concepts themselves are given. In fact they can be acquired through language or formed by conjoining properties as in Eq. (1) above. When given in linguistic form however, it is perhaps easier to think of the process as one of adding new hypotheses to our axiom set.

We must say next that our criteria for meaningfulness lack generality. We can use formation rules in some cases to rule out obviously stupid H's, and this includes such statements (H's) as

"All Alecs are Jocks" or

"All Scots are Irish", etc.

but we cannot, with certainty, obliterate all meaningless combinations.

We now need to go through the process on the computer of confirming, or infirming these hypotheses $H_1, H_2, \ldots, H_n$, and this takes us over to deduction.

First of all we meet the question of consistency in the same way as that raised in any axiomatic system. We can of course search through all the H's stored in our computer and look for examples of H_i and H_j. But this, while possible in "off line moments", is probably less important than the ability, when the H's are needed for deduction, of taking H_i, H_j and seeing whether a consequence C_k can be derived and H_j tested for consistency at the same time.

It goes without saying that in the process of taking pairs of H's for deductive consequences that we shall select pairs of H's with at least one common term. The following chart shows, in general terms, the process envisaged. In fact, statements relevant to the hypotheses under test may not contain either term in H, but be linked by common terms.

We have still to show that testing the statements (H's) for truth is the central feature of the usefulness of this total process. We will now turn to precisely this matter.

CONFIRMATION

We have at least one hypothesis H_1, and are asked to confirm it. We need, of course, factual support (Kemeny and Oppenheim, 1952). It is the "facts" with which we support hypotheses, and it is these so-called facts that lend credibility to them. We then will wish to say that the hypothesis H_1 is confirmed to the extent p (with some sort of measure included where this is possible) by facts F_j, i.e.

$$C(H_1; F_1, F_2, \ldots, F_j) = p \tag{2}$$

How then are we to realise this situation in a computer? It is clear that we are involved now in a logical and testing process that incorporates the flow charts of both figures 1 and 2. A version of this is shown in figure 3.

The test for meaningfulness is, for linguistic propositions, that they should exclude such formats as "is Scot an Alec?" (a class being a member of an individual); the contents of box 12 depend on a relevance test that involves the use of *the* terms in the question, and terms related associatively in memory. We assume in the case of box 16, that events not proved false, are, in fact, true (to some extent).

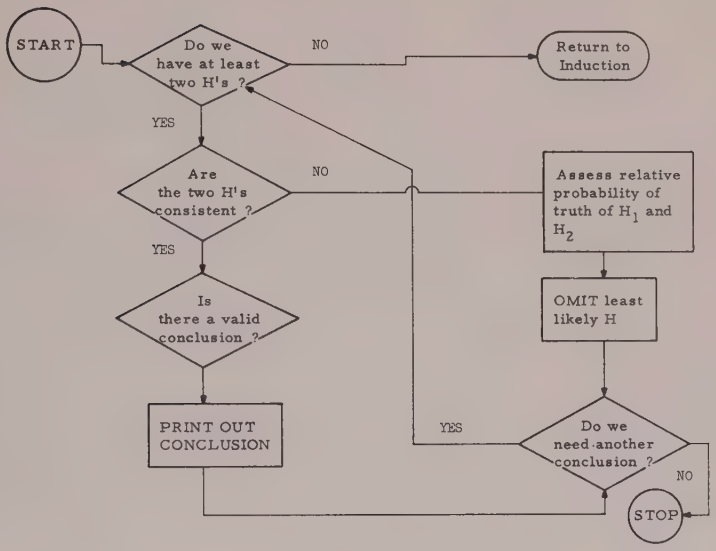

START
Do we have at least two H's ?
NO
Return to Induction
YES
Are the two H's consistent ?
NO
Assess relative probability of truth of H_1 and H_2
YES
Is there a valid conclusion ?
OMIT least likely H
PRINT OUT CONCLUSION
YES
Do we need another conclusion ?
NO
STOP

Fig. 2

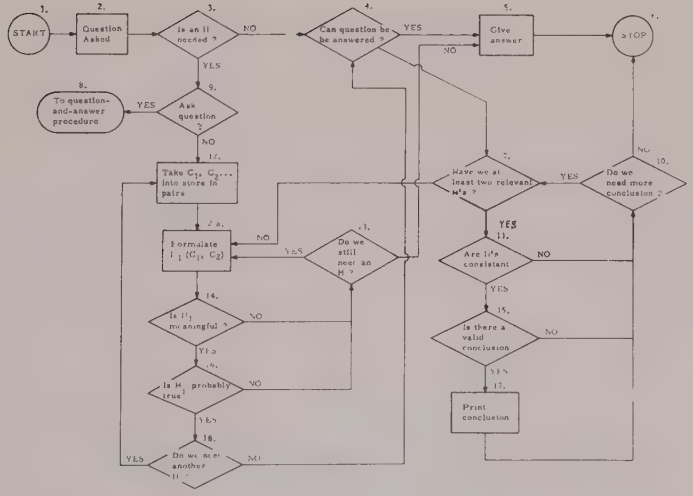

1.
START
2.
Question Asked
3.
Is an H needed ?
NO
YES
4.
Can question be answered ?
YES
NO
5.
Give answer
6.
STOP
8.
To question-and-answer procedure
YES
9.
Ask question ?
NO
10.
Do we need more conclusion ?
NO
YES
11.
Are H's consistent ?
NO
YES
12.
Formulate H_i (C_1, C_2)
NO
YES
13.
Do we still need an H ?
14.
Is H_i meaningful ?
NO
YES
15.
Is there a valid conclusion
NO
YES
16.
Is H_i probably true ?
NO
YES
17.
Print conclusion
18.
Do we need another H ?
YES
NO

Fig. 3

We can now in the computer merely show that the statements F_1, $F_2, \ldots, F_j$, are relevant to H_j, and are what we might call "evidential statements". Let us look at some commonsense examples. The statement "All Scots people are tall" (3) is supported by the statements "Jock is a Scot and Jock is tall" (4), "I have met no Scot who is not tall" (5), etc. We can now see that both consistency and relevance are factors which influence our search.

H_1, H_2, if taken in pairs, imply a conclusion which may or may not be valid, but H_1 and H_2 must be consistent and certainly if we are looking for evidential statements with respect to some state C_1, their relevance is vital.

We often say that given a conclusion we can find its premises, and here, as in the test for confirmation, we are really interested in ordinary deduction.

So if (3) is supported by (5), it is certainly the case that (5) implies (3), inductively, and given the truth of (3), (4) and (5) follow deductively. Therefore our test is, in principle, straightforward. We need consistency, relevance, and the ability to show that a statement supports inductively another statement by virtue of deductively following from it. This allows us now to effectively program our computer.

LANGUAGE

A computer message or input may take the following general form

$$(a_1 a_2 \ldots a_n)\,(AB \ldots N)\,(xy \ldots w)$$

the letters $a_1, a_2, \ldots, a_n$ are digit spaces that prescribe the features of the message: –

$a_1 = 1 \rightarrow$ linguistic
$a_1 = 0 \rightarrow$ non-linguistic input
$a_2 = 1 \rightarrow$ statement
$a_2 = 0 \rightarrow$ question
$a_3 = 0 \rightarrow$ present tense of verb
$a_3 = 1 \rightarrow$ future tense of verb
$a_3 = 2 \rightarrow$ past tense of verb
etc.

where $\rightarrow$ means "implies"

this is then followed by a statement, an example of which we reproduce in Polish notation, e.g.

$$Mxa$$

means, for $a_1 = 1$, $a_2 = 0$, and $c = 0$, "is x a member of a?".

So by choosing operators as representative of verbs (including adverbs) and connectives, and by choosing individual variables as classes, which represent nouns (including pronouns) or properties which are represented by adjectives, we can reduce complicated English sentences to simple subject-predicate statements.

If the operator M means "... is a member of ...", then if x = Jock and a = Scots people, then Mxa, prefixed by 10, means "is Jock a Scot?. If we now use this question to initiate our logical program, we find, following the flow chart of figure 3, that this could be followed by a simple search that elicits the response $11Mxa$ = "Jock is a Scot", assuming here that what is known (S_1) is stated, and the program is not concerned with the more complicated matter of "bluffs" and the like.

If the message $11Mxa$, which in machine code (binary) could be as follows: –

$$1100110100011000000000100 \qquad (6)$$

or broken down into stages is

(11)	(001101)	(00)	(011000)	(00)	(000001)	(00)
P	Q	R	S	T	U	V

where

P = linguistic digit and question or statement digit

Q = code for M

R = suffix (0 here) for M

S = code for x

T = suffix (0 here) for x

U = code for a

V = suffix (0 here) for a

Words can, in machine code, be of any finite length depending upon the length of the statement, but if we restrict ourselves to simple subject-predicate forms, we might seldom expect them to be very much longer than (6) above.

Now we come to the inferential process which wholly depends upon the other existing acceptable statements which are already in store. So if we have in store: –

S_2 = "Jock is tall"

S_3 = "All Scots are tall"

S_4 = "Jock is blue-eyed"

S_5 = "Most Scots are blue-eyed"

S_6 = "Jock lives is Kilmarnock"

S_7 = "Kilmarnock is in Scotland"

S_8 = "Almost all people living in Scotland are Scots"

The computer can now from S_2, S_3, ..., S_8 prescribe weightings for the statement S_1

Firstly S_1 is *not* in store

Secondly S_2, S_3, ..., S_8 are consistent since no two of the statements S_2, S_3, ..., S_8 imply anything that is inconsistent with any other statement.

e.g. if S_9 = Alec is a Scot
and S_{10} = Alec is short

then we (the computer) would have to question S_3, S_9 or S_{10}.

Furthermore, we have no statement that supports S_1, since there is no statement such as: –

S_{11} = Jock is short

But neither is there a clear deductive case which would exist if there was a statement

S_{12} = Only Scots people are tall

So we are left with an uncertain inference which could read

S_{13} = Jock is probably a Scot,

but we would like to strengthen that by some actual measure, or at least by modalities allowing of the use of *possibly* or *almost certainly* for *probably*, or indeed more subdivisions besides.

In fact, of course, in the light of no inconsistency, and lacking evidence of S_1, then S_6 and S_7 which yield

S_{14} = Jock lives in Scotland

taken with S_8 is the basis for saying "almost certainly".

We should mention that there is a problem for induction in computer terms, since we are really averring that inductive inferences, although formulated in language (off line) are essentially about processes being monitored (on line) and if we exclude on line procedures from the off line computer we make statements related in linguistic terms from other humanlike sources the only basis for induction.

HEURISTICS

We now need to say something about heuristics, since it is with heuristics in mind that programs have been written along the lines suggested in this article. Heuristics, it will be remembered, are generalized "rules of thumb" (inductions) that are also guides to intelligent activity in an environment, and which must be used when either algorithms are unavailable or uneconomic.

In formulating hypotheses (another name for heuristics), we can, when an appropriate goal is presented for example in the form of a question (Fig. 3), be formulating heuristics. Make the computer's environment a game and then a question about strategies in the game leads directly to the formulation of a heuristic, based on existing concepts. Where such heuristics are not useful (do not improve one's game) then new concepts must be available for hypothesis formation, or information must be obtained from external sources. Both are the basis of heuristics arrived at by uncertain inference or induction.

Finally it must be said that under the conditions where our scheme is specifically heuristic-forming, we need to develop the recognition problem referred to at the beginning of this note. We need now, in games parlance, to recognize 1-person, 2-person, ..., n-person, determinate, indeterminate, with or without coalition, games, etc. Then we have a similarity criterion which allows the formation of particular and "most-likely-to-be-relevant" heuristics. We still need of course some form of recursion formula or meta-rules which allow the more precise "fitting" of a heuristic to a situation. Let us give a simple example.

If in a game G_1, a rule R_1 says "always fill in the centre square first" then, assuming R_1 was part of an algorithm or a well-confirmed heuristic, then only after R_1 was used and failed in a new "similar" game G_2, should our meta-rule operate and this would – subject to game rule

constraints – be applied as "always fill in ... first". The goal offered by the game constraints and the subgoals which can be derived by precisely our logic inference programs described in this article, will suggest which variation is most likely to prove successful e.g. "corner square", "near centre, but not centre, square", etc.

This brief statement is all that space permits in the linking together of inference-making and hypothesis confirmation to heuristics and their formation and use.

Finally let it be said that many programs have been written for the computer along these lines and have at least held out high hopes of success in many fields of application. These programs have in practice gone beyond syllogistic reasoning and considered inferences made on factual as well as formal bases.

REFERENCES

George, F. H. (1965) Computer Applications in Decision Making and Process Control. Paper given at International Symposium on Long Range Planning for Management, Paris, 20–24 September 1965.

George, F. H. (1966) Language, Logic and Computers. Paper to be published in "*Automata Theory and Learning Systems.*" Academic Press. New York.

Kemeny, J. G., and Oppenheim, P. (1952) Degree of Factual Support *Philos. Sci.* **19**, 307–324.

Newell, A., Shaw, J. C., and Simon, H. A. (1963) The Theorem Proving Machine. In "*Computers and Thought.*" (Eds Feigenbaum, A., and Feldman, J.). McGraw-Hill. New York.

SECTION IV

Engineering Models and Applications

K. STEINBUCH

and E. SCHMITT

Institut für Nachrichtenverarbeitung und Nachrichtenübertragung

University of Karlsruhe,

Karlsruhe, Germany

Adaptive Systems Using Learning Matrices

ABSTRACT

Of late, learning machines are gaining in practical importance for processes in which the governing mathematical rules are a priori unknown, and/or vary with time. Such systems change their behaviour on the basis of their past experience. During the adaptation stage a mathematical relationship between the input and the output of the adaptive system is established by changing the parameters of an adjustable transformation unit in accordance with a selected criterion of performance.

The adjustable transformation unit can be realized using Learning Matrices. Depending on the complexity of a Learning Matrix structure, binary or non-binary sets of input patterns can be transformed to outputs which may either be represented in an "one out of *m*-code" or in any binary code or in a set of non-binary output components. In general, the latter transformation may be arbitrary and non-linear.

It is an inherent and advantageous property of Learning Matrix systems that the data are processed in parallel and that the systems are capable of operating in various modes. In the training mode (operating mode 1) some definite training procedure may be implemented. In special cases, adaptation can be performed automatically too, either in one step, or by an iterative procedure using some averaging technique or by other iterative methods.

Learning Matrix systems have been realized by using adaptive hardware and by digital computer simulation. As one special example, a hierarchical adaptive system which is capable of predicting numbers of binary sequences whose statistical properties are not restricted is described.

ADAPTIVE SYSTEM MODEL

In recent years much effort has been made to develop technical systems which we generally refer to as "adaptive systems," "learning systems" or "self-organizing systems." Although many definitions of

such systems exist, a general description including all of them is yet lacking.

Adaptive systems have practical importance for processes in which the governing mathematical laws are a priori unknown, and/or vary with time. Their purpose is to achieve a practical goal, for example control a non-linear process, or recognize variable input patterns or forecast the weather etc.

In order to explain the principal function of an adaptive system, it is convenient to consider the model in Figure 1. The relevant data pertaining to the physical conditions of the environment to be adapted are transduced by the sensor unit. It should be noted that certain difficulties are associated with the selection of suitable characteristic data of the process. As this problem is beyond the scope of this paper it will not be dealt with in greater detail.

In order to transform the input signals into the respective output signals a transformation unit is needed, the parameters of which are adjustable. Further, an adaptation algorithm must be provided which

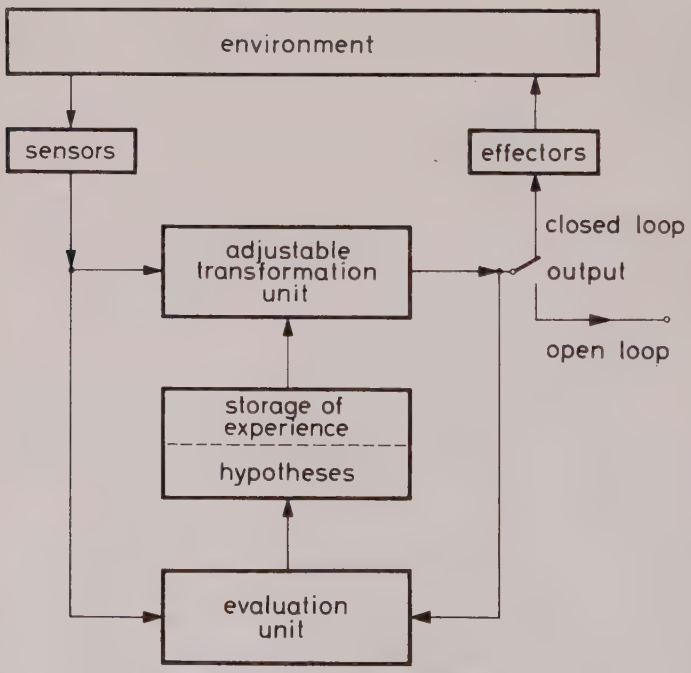


Fig. 1. Adaptive System Model

should be capable of making adjustments depending on the instantaneous physical situation of the environment and the past experience stored in the system (storage of experience).

The decisions for adjusting the particular elements and for storing or eliminating past experience are controlled by the evaluation unit in such a way that the performance of the system would be improved successively during adaptation, until the desired transformation is established. Out of several possible adaptation criteria one may be selected corresponding to the instantaneous situation. Further, facility must be provided to store past experience, including the possibility "to forget," in cases where the processes are non-stationary. All these possibilities foreseen by the designer will be henceforth called "hypotheses." The larger the number of hypotheses, the higher would be the probability to find a desired solution of the problem provided that convergence can be achieved. Lastly a suitable evaluation criterion for the performance of the system has to be chosen.

In some applications the output of the adaptive system is fed back through effectors (closed loop system), for instance in closed loop control systems. In order to effect a goal directed change of the environment, suitable parameter changes selected by the hypotheses unit must be carried out internally (test signals) depending on the past experience and the state of the external system to be controlled.

On the other hand there exist processes in which the output has no influence on the environment, for instance as in the case of pattern recognition, weather forecasting, or similar applications (open loop system).

It should be noted that this system stands for a general model. In practical systems sometimes the particular units of the scheme are integrated so that they are not distinguishable from one another. In some cases the storage of experience may be contained in the adjustable transformation unit.

LEARNING MATRICES AS ADJUSTABLE TRANSFORMATION UNIT

Adjustable transformation units can be realized by Learning Matrices. The principle of Learning Matrices, and some of their applications have been published in earlier papers.[1,2,3,4]. The essential features concerned with adaptive networks will be described below.

The structure of a single Learning Matrix may be represented as in Figure 2. It consists essentially of adjustable connecting elements ("weights") arranged in the form of a matrix. The columns and rows of the matrix are connected with input and output units which come into action during the various operating modes. Three operating modes are possible.

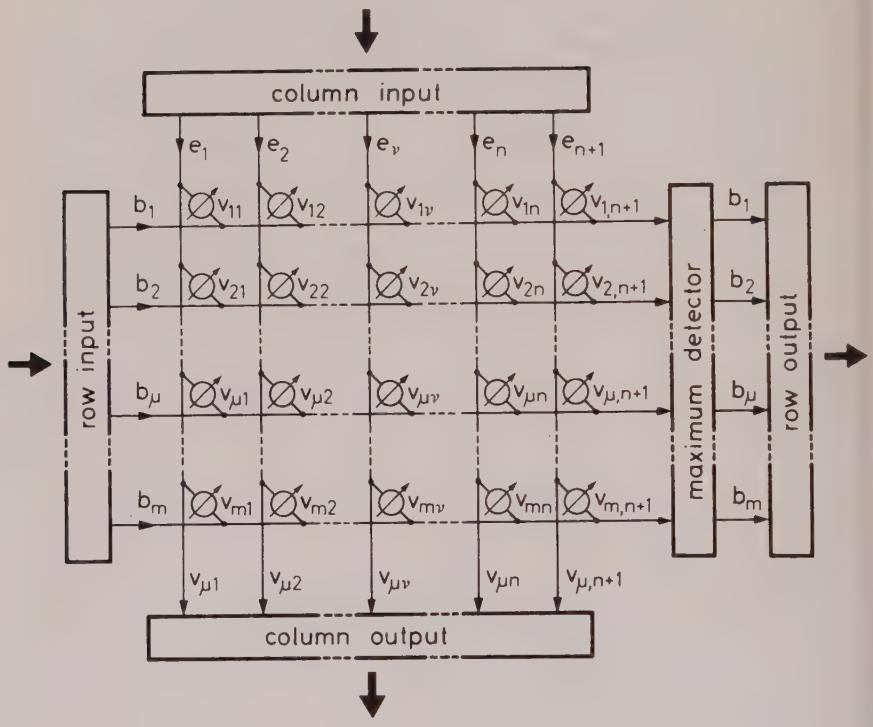


Fig. 2. Schematic of the Learning Matrix

1. Operating mode 1 (training mode)

The input signals $e_1 \ldots e_n$ derived from the sensors are applied to the columns (Fig. 3). They may be binary or non-binary.

In the latter case, we assume that their values are such that

$$-1 \leqq e_\nu \leqq +1 \tag{1}$$

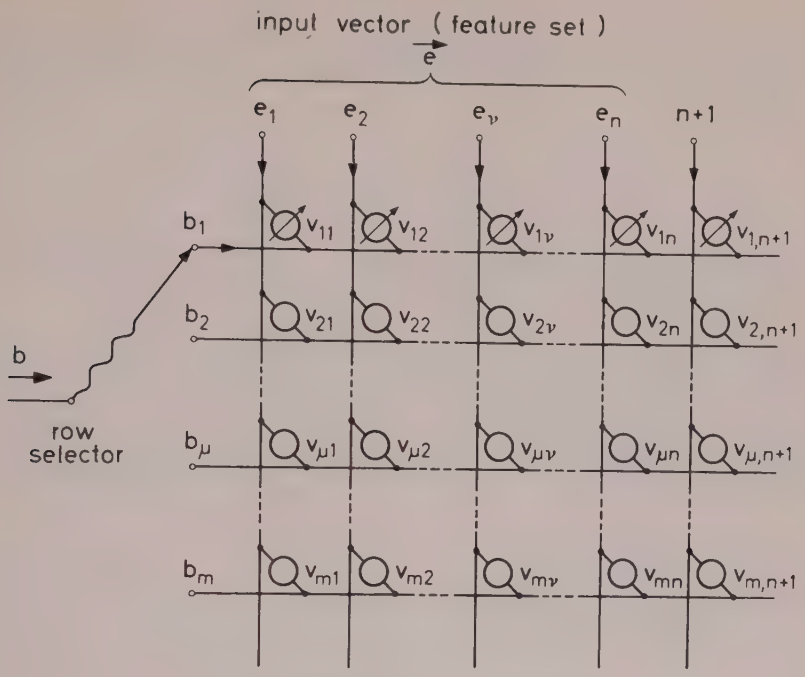


Fig. 3. Operating Mode 1 (Training Mode)

If the inputs are binary, then they may take the values $+1$ or -1 directly.

Further, we suppose that the adjustable connecting elements may be varied between the limits -1 and $+1$; i.e.

$$-1 \leqq v_{\mu\nu} \leqq +1 \tag{2}$$

During the training mode the "weights" of one particular row are adjusted simultaneously. In order to achieve this, a certain row is chosen by the row selector, and the elements are varied by means of coincidence of row and column signals. There are various procedures possible to adjust the elements, based on certain criteria, which will be described later. Each of the possible input sets is assigned to a particular row; each row corresponds to a certain "class".

Besides the n columns corresponding to the n components of the input, it may be seen from Figure 3 that there is an auxiliary column $n + 1$ of connecting elements. By means of this, the performance of the system in

classifying the various input sets is improved. In the case of binary signals the $(n + 1)^{st}$ column is not necessary, as other means for classifying are provided.

2. Operating mode 2 (classifying mode)

During the operating mode 2 which would be also called classifying mode it is required that the input sets be classified to the respective classes represented by the "weight sets" contained in the various rows.

As is shown in Figure 4 the n components of the first input signal $e^{(i)}$ are applied to the columns 1 ... n, and a constant signal +1 to the column $n + 1$. In each of the rows a linear function of the input $e^{(i)}$ is formed by multiplying the "weights" in the columns with the corres-

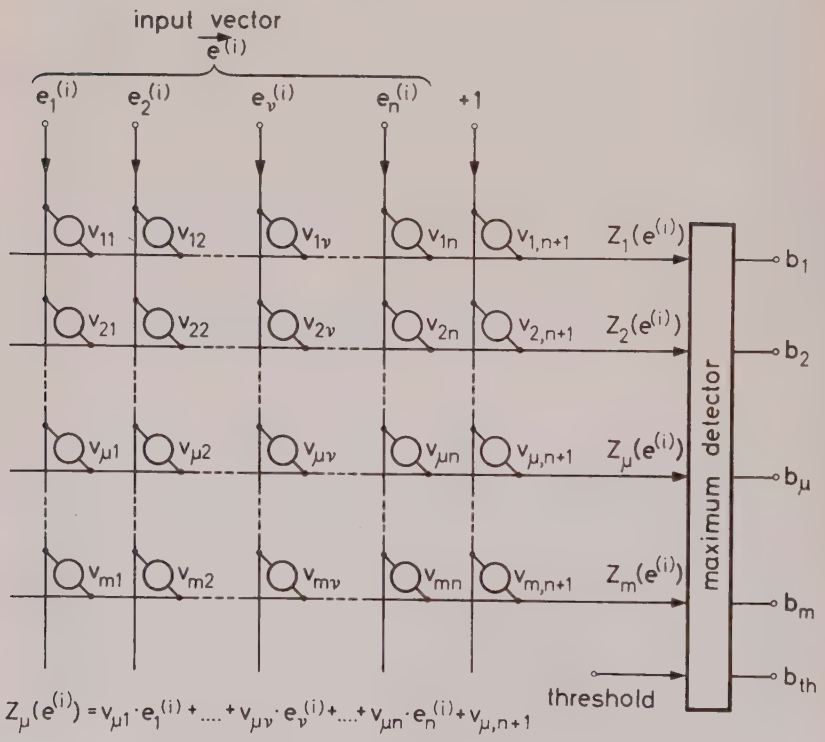


Fig. 4. Operating Mode 2 (Classifying Mode)

ponding signal components and summing up along the rows. Thus we obtain a set of linear functions $Z_1(e^{(i)}) \ldots Z_m(e^{(i)})$ at the outputs of the rows, namely

$$\begin{aligned} Z_\mu(e^{(i)}) &= v_{\mu 1} \cdot e_1^{(i)} + v_{\mu 2} \cdot e_2^{(i)} + \cdots + v_{\mu \nu} \cdot e_\nu^{(i)} + \cdots \\ &\quad + v_{\mu n} \cdot e_n^{(i)} + v_{\mu, n+1} \\ &\quad \mu = 1, \ldots, m \end{aligned} \tag{3}$$

The maximum detector selects and indicates the row with the greatest sum.

It may be seen that any input set $e^{(i)}$ can be represented by an n-dimensional vector in the Euclidean space.

$$e^{(i)} = \{e_1^{(i)}, e_2^{(i)}, \ldots, e_\nu^{(i)}, \ldots e_n^{(i)}\} \tag{4}$$

Similarly a "weight set" may also be represented by a vector v_μ, however, with $n + 1$ components.

$$v_\mu = \{v_{\mu 1}, v_{\mu 2}, \ldots, v_{\mu \nu}, \ldots, v_{\mu n}, v_{\mu, n+1}\} \tag{5}$$

Further, we define a reduced "weight vector" v_μ with n components

$$v_\mu^* = \{v_{\mu 1}, v_{\mu 2}, \ldots, v_{\mu \nu}, \ldots, v_{\mu n}\} \tag{6}$$

The input sets and the n-dimensional "weight set" v_μ^* may also be thought of as being represented by the endpoints of the vectors in the n-dimensional space.

Classification is obtained as a result of partitioning the n-dimensional space into regions by a set of hyperplanes which satisfy the following equations:

$$\begin{aligned} &Z_1(e^{(i)}) - Z_2(e^{(i)}) = 0 \\ &Z_1(e^{(i)}) - Z_3(e^{(i)}) = 0 \\ &\vdots \\ &Z_1(e^{(i)}) - Z_m(e^{(i)}) = 0 \end{aligned} \tag{7}$$

$$\begin{aligned} Z_2(e^{(i)}) - Z_3(e^{(i)}) &= 0 \\ &\vdots \\ Z_2(e^{(i)}) - Z_m(e^{(i)}) &= 0 \\ &\vdots \\ Z_{m-1}(e^{(i)}) - Z_m(e^{(i)}) &= 0 \end{aligned} \tag{7}$$

There are $\frac{m \cdot (m-1)}{2}$ such equations (m = number of classes) of which, however, not all are necessarily linearly independent of one another. Further, under certain circumstances, some of the hyperplanes represented by Eq. (7) may be redundant.

In the case of n-dimensional input sets the separating hyperplanes have $(n-1)$ dimensions. Thus, for two-dimensional inputs straight lines are obtained; for three-dimensional inputs we obtain planes.

In vector representation, the Eq. (7) can also be written in the form:

$$\begin{aligned} e^{(i)} \cdot (v_1^* - v_2^*) &= (v_{2,n+1} - v_{1,n+1}) \\ e^{(i)} \cdot (v_1^* - v_3^*) &= (v_{3,n+1} - v_{1,n+1}) \\ &\vdots \\ \text{etc.} & \end{aligned} \tag{8}$$

For all cases where the difference of the $(n+1)^{st}$ "weights" in Eq. (8) is not equal to zero, the separating hyperplanes do not pass through the origin of the Euclidean space. The function of the auxiliary "weights" in the $(n+1)^{st}$ column is therefore explained. By giving them appropriate values we may obtain a lateral shift of the separating hyperplanes.

As a special case we consider minimum distance classification. Here, an input set is categorized to such a class that the distance between the points corresponding to the input set and to the representative of the class is a minimum. In this case it is necessary to adjust the "weights" of the $(n+1)^{st}$ column to values such that

$$v_{\mu,n+1} = -\frac{1}{2}|v_\mu^*|^2 \tag{9}$$

$$\mu = 1, \ldots, m$$

To illustrate this special case we consider two-dimensional input sets which are to be classified into four classes. We assume that the class representatives exist as weight vectors stored in four rows of a Learning

Matrix. In Figure 5 the points $V_1^*, \ldots, V_4^*$ denote the endpoints of the weight vectors $\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_4^*$ in the two-dimensional space. In this case the separating hyperplanes are the perpendicular bisectors of the lines joining the points $V_1^*, V_2^*, V_3^*, V_4^*$. The classes are formed as regions bordered by these separating lines. Input sets are classified according to the regions in which the endpoints of the vectors are located. As may be seen in Figure 5 the separating straight line (indicated by the dotted line S_{24}) between class 2 and class 4 is redundant. This occurs generally when classes are not adjoined.

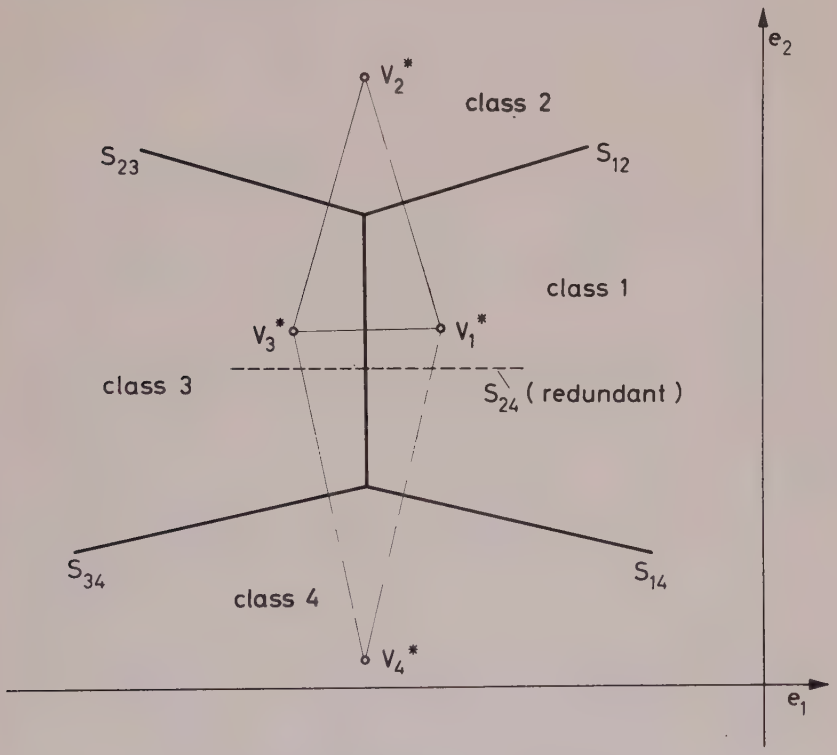


Fig. 5. Minimum Distance Classification

Further, in some applications it would not be expedient to classify input sets which lie too far away from the nearest class representative. Such input sets can be recognized and indicated by means of an additional variable threshold input to the maximum detector (Fig. 4). To do this, we

can arrange that, in case of binary inputs, the row output b_{th} delivers a signal if a predetermined Hamming distance between an input set and nearest representative is exceeded.

3. Operating mode 3 (read out)

In this mode of operation the components of the representative vector stored in any row are read out at the column outputs (Fig. 6). The components of a certain row are either binary or non-binary. As it will be

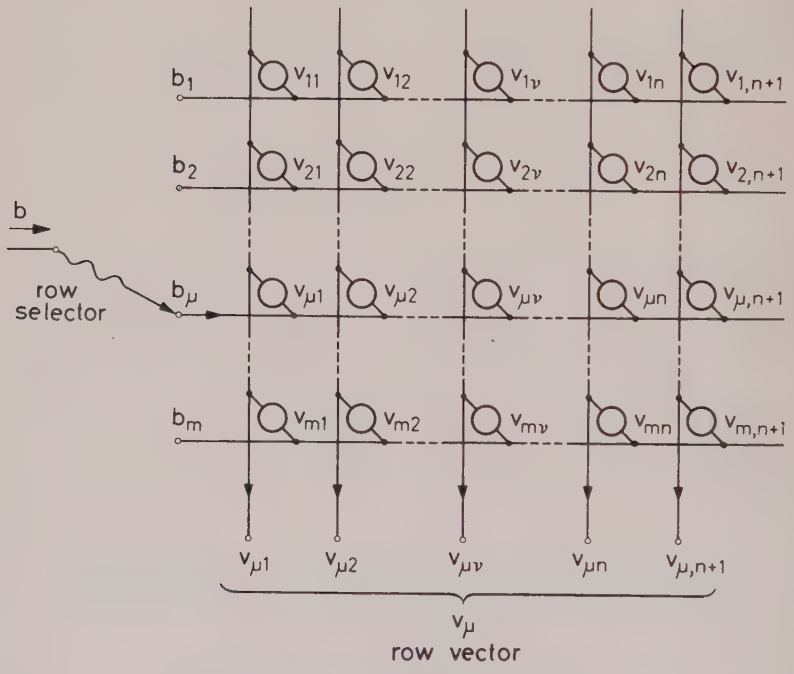


Fig. 6. Operating Mode 3 (Read Out)

shown later, even in the case of binary input sets the connecting elements may assume non-binary values. If, in spite of this, binary signals are desired at the column outputs (Fig. 6) a suitable transformation must

be performed. For instance according to the following relationship

$$\text{column output} \begin{cases} v'_{\mu\nu} = +1, & \text{if} \quad v_{\mu\nu} > 0 \\ v'_{\mu\nu} = -1, & \text{if} \quad v_{\mu\nu} < 0 \\ \text{undecided}, & \text{if} \quad v_{\mu\nu} = 0 \end{cases}$$

4. Learning Matrix-Dipole

The various operating modes of the Learning Matrix provide us with means to construct comprehensive systems for performing adjustable transformations. A so-called Learning Matrix-Dipole consisting

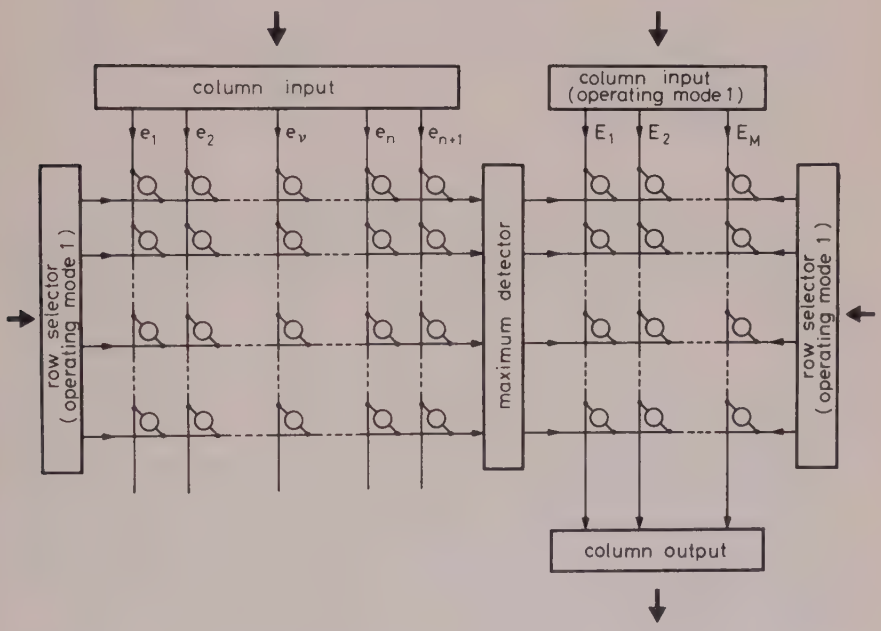


Fig. 7. Schematic of Learning Matrix-Dipole

of two matrices is shown in Figure 7. The connecting elements of a given row of the two matrices are adjusted simultaneously, whereby, the adaptation laws in the operating mode 1 may be different for the two matrices. When the left matrix operates in the classifying mode, the

right one operates in the read-out mode. Thus any binary or non-binary input set of n components can be transformed into any desired output set of M components. This is performed in two stages: In the left matrix the n-dimensional input set is transformed into a binary output in an "one out of m-code," the right matrix transforms this "one out of m" binary input into a binary or non-binary M component output. In this way non-linear transformations may also be effected.

In the last years Learning Matrices have been realized by suitable hardware and by computer simulation. The most appropriate elements for realizing the connecting elements are tape-wound magnetic cores up to now.

It should be observed that the great advantage of Learning Matrix networks in hardware is the fact that during all operating modes the data are processed in parallel. For this reason a Learning Matrix in hardware operates very much faster than a simulated network of this type by a digital computer.

5. Adapting techniques

During the training mode the connecting elements of the Learning Matrix are adjusted. This may be done by various procedures.

6. Binary input sets

a) one step adjustment

A trivial case is that of one step adjustment. Here, the connecting elements are set in one step to the binary values $+1$ or -1 depending on the binary inputs. This procedure is useful if the representative of the desired class is known and available to be applied to the input.

b) stepwise adjustment

If this is not the case, then a one step adjustment cannot be applied. The following averaging technique is useful, if the endpoints of the input vectors are clustered around a mean value. During every adjustment step, instead of setting the connecting elements to their full values at once, they are shifted by a small amount Δv in the appropriate direction. Thus, the "weights" of the matrix are made to act as up

and down counters with boundaries at $+1$ and -1. The rate at which the boundaries are achieved is dependent on the value of Δv which may be controlled according to a suitable criterion ("h-input," see ref. 3). For instance, the input sets belonging to a class may be statistically distributed around a mean value. Then, after some adjusting cycles the components of this mean value set would be formed in the connecting elements of the corresponding row.

It must be noted that the lengths of the row vectors are not necessarily equal. In order that the input set be classified correctly in the operating mode 2, the classification must be invariant with respect to the vector length. This can be achieved by increasing all the length to the maximum value $|\boldsymbol{v}_\mu^*| = \sqrt{n}$ (as all components are either $+1$ or -1). In this case an input set would be classified to a class to whose representative it bears the smallest angle.

7. Non-binary input sets

a) One step adjustment

If, as in the case of binary input sets, the class representative is available at the input, adaptation can be performed in one step[5]. Here, it is suitable to give the $(n+1)^{st}$ weights the value in Eq. (9) so that the matrix may act as a minimum distance classifier.

b) Standard average adjustment

For stationary processes, under the condition that the input sets are clustered around a mean value, a standard average technique is useful to perform adaptation. The class representative would be obtained by presenting the cluster of input sets in several steps.

The $(N+1)^{st}$ adjustment of the μ^{th} row vector $\boldsymbol{v}_\mu$ would then be

$$\boldsymbol{v}_\mu^*(N+1) = \frac{1}{N+1} \cdot [N \cdot \boldsymbol{v}_\mu^*(N) + \boldsymbol{e}(N+1)] \tag{10}$$

where the N's represent the number of adjustments at which the row μ was selected.

If all clusters have equal diameters, minimum distance classification may be used, whereby equidistant hyperplanes constitute the separating surfaces. Otherwise the adjustment of the $(n+1)^{st}$ "weights"

may facilitate the required shift of the separating hyperplanes to be achieved.

c) Weighted average adjustment

In the case of non-stationary processes, a weighted average technique should be employed in order that events occurring at earlier stages have a weaker effect on an adjustment. For example, the following adjustment law can be applied:

$$\boldsymbol{v}_\mu^*(N+1) = \alpha \cdot \boldsymbol{v}_\mu^*(N) + (1-\alpha) \cdot \boldsymbol{e}(N+1) \qquad (11)$$

$$(0 < \alpha < 1)$$

In this case the $(N+r)^{\text{th}}$ input set is weighted α^{-r} times as strong as the N^{th} input set. By setting $\alpha = \dfrac{N}{N+1}$ Eq. (11) reduces to Eq. (10).

A more versatile system may be obtained by controlling the parameter α in Eq. (11) with a suitable evaluation function.

BINARY PREDICTION USING A LEARNING MATRIX-DIPOLE

As an application of an adaptive system (open loop) illustrated in the model of Figure 1 we consider a device capable of predicting binary sequences. In this system a Learning Matrix-Dipole is used as an adjustable transformation unit.

The complexity of a predicting system depends on the statistical properties of the binary sequences to be processed. We assume that the binary sequences are represented by non-homogeneous Markovian processes, i.e. the lengths (degrees) and the conditional probabilities of the Markovian chains vary with time. In doing this we include all other statistical processes in our system.

An adaptive system designed to process such sequences should adapt itself to

a) the varying lengths of the Markovian chains, and

b) the varying conditional probabilities.

Figure 8 shows a system capable of performing these requirements. It consists, essentially, of a Learning Matrix-Dipole structure. The requirement a) is satisfied by the left matrix which is divided into a set

of subsystems $P_1, P_2, \ldots, P_n$. Markovian chains of length one are adapted in the subsystem P_1, that of length n in the subsystem P_n. The evaluated past experience of the subsystems are stored in the last column ex_{n+1}.

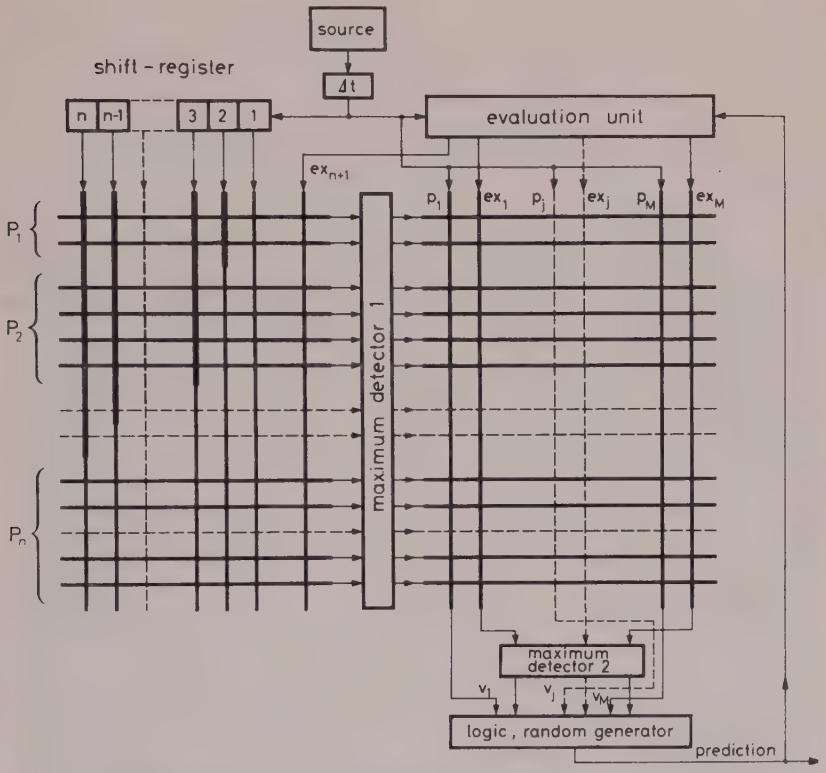


Fig. 8. Binary Prediction using a Learning Matrix Dipole

Requirement b) can be satisfied by the right-hand matrix which calculates the conditional probability of the oncoming binary number according to several procedures. A column pair (j) is provided for each procedure, whereby in the first column (p_j) the conditional probability is calculated and in the second (ex_j) the experience corresponding to the respective procedure is stored.

The various subsystems of the left matrix and the different procedures in the right one can be looked upon as a set of hypotheses. The hypo-

theses of the left matrix are primary, and those of the right are secondary; i.e. for prediction, first, a subsystem, and subsequently a column pair are selected such that the "best" success will be achieved.

The input signals are emitted as a binary sequence from the source. Δt represents some time delay of the oncoming number with respect to its prediction. The input signals to the left matrix are derived from the parallel outputs of a shift-register in which the preceding set of n binary digits are stored. This set constitutes the Markovian chains of variable length 1 ... n. Markovian chains which occur the first time are adapted into the respective subsystems P_1, ... P_n in one step, by automatically selecting a free row of each subsystem, as described in (3). Thus, each row represents a certain Markovian chain. In the same row of the right matrix the conditional probabilities of this chain are formed according to the various procedures which differ from one another in the step amount Δv by which the connecting values are varied. In the first column (p_1), the step-width Δv has a small value which would be useful in the case of stationary processes. In the last column (p_M), Δv has a large value, so that very strong nonstationary processes may be adapted. The adjustment of the connecting elements is as follows. If the number following the Markovian chain is $+1$ the connecting value is increased by Δv, if it is -1 it is decreased by Δv.

The performance of the various hypotheses consisting of the subsystems in the left matrix and the M procedures for calculating the conditional probability in the right matrix is evaluated by the evaluation unit when the predicted input signal occurs. The results of the evaluations are stored as experience in the column ex_{n+1} of the left matrix as well as in the columns ex_1 ... ex_M of the right matrix. This is achieved by increasing the respective connecting values by a fixed amount Δv in the case of correct prediction, and decreasing them in the case of incorrect prediction. Thus, the connecting element with the highest value determines the "best" hypothesis. By variation of the step-width Δv, past experience can be eliminated more or less rapidly (as the connecting values are bounded at ± 1).

For prediction of the next number, the left matrix works in the classifying mode in which a row, corresponding to the Markovian chain at the input, and to the best subsystem, is selected by the maximum detector 1. The right matrix operates in the read-out mode (operating mode 3). Here, the best hypothesis is selected by the maximum detector 2.

The corresponding weight value v_j is used to predict the number most probably emitted by the source. The prediction would be:

$$+1, \quad \text{if} \quad v_j > +\delta$$

$$-1, \quad \text{if} \quad v_j < -\delta$$

$$\text{indeterminate, if} \quad -\delta \leqq v_j \leqq +\delta$$

(δ represents a certain dead zone)

In the last case a binary number would be generated at random.
Figure 9 shows a realized system of the described structure in hardware. The input signals may be generated by pushing the buttons "0" or

Fig. 9. Binary Number Predictor

"1", or read in by a tape reader. We conducted a large number of experiments with this system and we investigated stationary as well as non-stationary binary sequences consisting of Markovian chains 1 ... 4. To obtain a measure for the performance of the system, the cumulative entropy and the cumulative redundancy of the binary sequences were calculated[6].

An experiment with 121 binary sequences was started containing 700 digits each. The processes were mostly nonstationary. The cumulative redundancy revealed that 73% of the digits may be predicted correctly by an ideal system. With the system in Figure 9, consisting of four subsystems and three procedures for adapting the conditional probabilities, we obtained 69.8% correct predictions.

REFERENCES

1. Steinbuch, K. Die Lernmatrix, *Kybernetik* **1**, 36–45, January 1961.
2. Steinbuch, K., and Piske, U. A. W. Learning matrices and their applications, *IEEE Trans. on El. Computers*, EC-**12**, 846–862, December 1963.
3. Steinbuch, K. Adaptive networks using learning matrices, *Kybernetik*, **2**,148–152, February 1965.
4. Steinbuch, K., and Widrow, B. A critical comparison of two kinds of adaptive classification networks, *IEEE Trans. on Electronic Computers*, EC-**14**, 737–740, October 1965.
5. Müller, P. Eigenschaften und Aufbau von Lernmatrizen für nichtbinäre Signale, *Kybernetik* **2**, 103, September 1964.
6. Schmitt, E. Untersuchungen an Binärprädiktoren, insbesondere bezüglich ihrer Anpassungsfähigkeit und ihrer Vorhersageleistung gegenüber Versuchspersonen, *Kybernetik*, **2**, 93–102, September 1964.
7. Nilsson, N. J. "Learning machines", McGraw-Hill, New York (1965).

HANS MARTIN LIPP

Institut für Nachrichtenverarbeitung und Nachrichtenübertragung,

Technical University of Karlsruhe,

Karlsruhe, Germany.

An Application of the Learning-Matrix Dipole

The learning-matrix ([1]) is a special adaptive network in which appropriate (mostly magnetic) components realize adjustable connections between the column wires and the row wires of a two-dimensional matrix.

The crosspoint of each column wire and each row wire contains one connection-element, the value of which is adjustable as a function of the column and row signals. All row wires are terminated by a common unit the maximum detector, which determines the row with the highest signal level (see Figure 1).

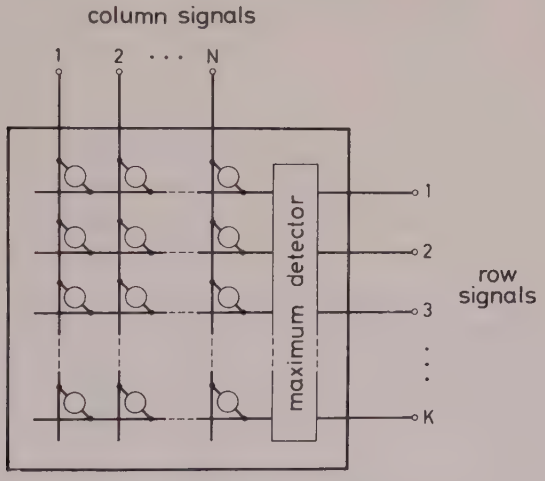


Fig. 1. Schematic of the Learning-Matrix

Using the learning-matrix three modes of operation are to be distinguished (the learning-matrix and its function will be described in detail in another paper ([2]); therefore the following explanations will only show the operation in principle).

1. Operating mode 1, also called training mode.
   Signals will be applied to all column wires and to one of the row wires. The row signals are binary ones, while the column signals either may be binary or nonbinary. All those connection elements, selected by coincidence of column and row signals, will change their values by a certain amount in a given tendency (e.g. both signals are binary 1, the connection value enlarges; the row signal binary 1, the other binary 0, the value is decreased). The connection value is limited by an upper and a lower boundary, and therefore the variance is less than or at most equal to the difference of these two boundary values.
   Repeating the formation of the crosspoints with slightly differing column signals and the same row signal, that row contains at last a weighted average as a representative of that class the applicated column signals define.
   After having formed the connection elements two other operating modes are possible.
2. Operating mode 2. During this operating mode only column signals are applied to the learning-matrix. Via the connection elements they produce row signals, which are a function of the values of the column signals and the connection elements. The signal of each row gives a measure of the "similarity" between the applied signals and its representative ("similarity" is in the case of binary row signals related to the Hamming distance of two code-words). The maximum-detection unit indicates the row with the largest similarity signal.

   In a simplified description this operating mode of the learning-matrix equals in the case of binary signals the function of a translator, which transforms any binary code of N places into a $\binom{K}{1}$-code, if the matrix contains N columns and K rows.

   In contrast to common translators, the transformation rule is variable (by using operating mode 1) and not fixed.
3. Operating mode 3. Whereas in operating mode 2 only the column signals are given, in this mode a signal to one row only is applied. The connection elements now produce column signals proportional to their values (either binary or nonbinary). This way, the representatives of each class are available at the column wires (with nondestructive read-out).

   In a simplified description this operating mode of the learning-matrix equals in the case of binary signals the function of a common

translator, which transforms a $\binom{K}{1}$-code into any binary code of N places with variable transformation rule.

It is possible to interpret the two transformations described as inverse operations. The application of these inverse operations one after the other enables us to assign, for example, to any binary code of N places any binary code of M places, using two learning-matrices with N or respectively M column wires and the same number of K row wires. The value of N and M need not be equal and may differ considerably.

Figure 2 shows this translator system using two row-coupled learning-matrices. This systems is called learning-matrix dipole.

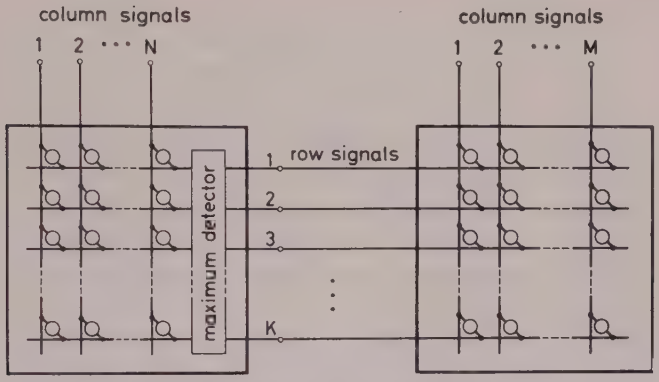


Fig. 2. Schematic of the Learning-Matrix Dipole

One of the training methods possible for establishing the transformation rule is the following. (Further on, only the special case of binary signals will be considered: During the first period, the left-hand learning-matrix will be operated in operating mode 1, that means column and row signals have to be applied at the same time. This first step results in a code transformation from a binary code of N places into a suitable $\binom{K}{1}$-code. In the second period the left-hand matrix operates in mode 2 and assigns to each code word of N binary places a row signal. The right-hand learning-matrix is in operating mode 1, and adjusts, according to the incoming column signals, the connection values of the row, which is indicated by the left-hand learning-matrix. Repeating these steps for each code word, the transformation rule is

being established. Note, that with this procedure modifications of the connection values of any row do not influence the correct function of all other rows. Other adaptive translator networks often need repeated training sequences, if a part of the transformation rule has to be changed.

After having described the learning-matrix dipole we will give an example for a possible application of this network. In the following part of this paper a more complicated adaptive system will be described, in which a dipole may improve the efficiency of an adjustment unit.

First we shall define the problem of that adaptive system. Assume given a system 1 (see Fig. 3) which can be interpreted as a discrete, finite and deterministic automaton, with M parallel binary inputs and

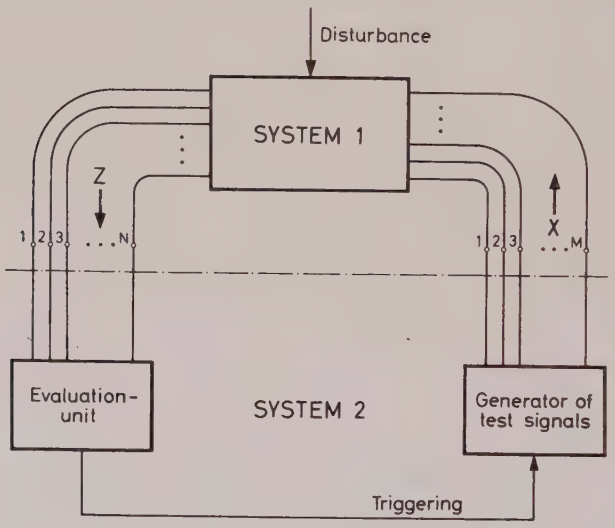


Fig. 3. Main Elements of the Adjustment System

N parallel binary outputs. The combinations of the output signals are one-to-one mappings of the states Z of the system. The application of an input-signal X_t at time t changes the present state Z_t and results in a new state Z_{t+1} at time $(t + 1)$. The transition into state Z_{t+1} is only a function of Z_t and X_t and of no other previous states or input signals. It should be mentioned, that a transition cannot be reversed by removing the input signals, but only by applying that input signal X, which causes the backward-transition.

A second system is in contact with this first system, possessing as many inputs as system 1 possesses outputs and as many outputs as system 1 shows inputs. The problem is, that system 1 should be brought by system 2 in a predetermined but unknown state Z_T. For this purpose, system 2 is given an adjustment criterion, which enables it to evaluate each present state Z_t and to decide, whether it is equal Z_T or not, and if a transition has been valuable. Because we do not handle that problem in this paper, we suppose that a given "evaluation unit" can do that work. This unit delivers a signal, in case the state Z_T the system 2 looks for has not been reached.

Because system 2 has at first now knowledge about system 1, system 2 cannot compute the chain of necessary transitions to reach Z_T. More or less efficient random test signals may perhaps solve the problem. After Z_T being once adjusted, system 2 ceases to produce new test signals.

In addition to the already mentioned properties of system 1, this system will not always remain in state Z_T. It changes randomly its state without changing its structure. Therefore system 2 has to adjust system 1 anew. For simplicity's sake it may be assumed that such random changes do not occur during the adjustment procedure.

With this assumption for such adjustment systems the typical diagram results (shown in Fig. 3). It consists of two main units, the evaluation unit and a generator to produce test signals. This generator will be triggered by the evolution-unit as long as Z_T has not been reached. Both units are essential for the adjustment procedure. If random changes of the state of system 1 occur, they have to begin testing it anew, but experience of earlier trials is not available, and therefore the efficiency of system 2 in general is not very high. A first improvement of the system performance naturally will consist in adding an appropriate storage unit for the knowledge gained during the test periods. There are several ways of organizing such a storage unit. The course of the adjustment procedure itself refers to a favourable structure. For each state Z_t an input-signal X_t occurs, causing a transition of system 1 into another state Z_{t+1}. The evaluation-unit evaluates this new Z_{t+1} and decides whether this transition was useful with regard to the problem of finding Z_T or not. If this X_t made progress in direction towards Z_T, it is sufficient to store Z_t and X_t together. If the test signal fails, the X_t will be omitted, immediately or later on, if a better test signal has been found.

The way from a given state to the target-state Z_T therefore may be described by a sequence of $Z_t - X_t$ sets.

The storage of such couples of binary coded signals can be easily done with the earlier described learning-matrix dipole with regard to its properties as a translator.

The left-hand matrix stores the states Z_t and the right-hand matrix takes over the test-signals X_t, using the same row as the other learning-matrix. The sequence of the various states Z_t in time and the sequence of the rows must not be identical, because the classification properties of the learning-matrix during the operating mode 2 define the appropriate row and not the more or less random time sequence. Between the inputs of system 2 and the left-hand learning-matrix a small storage has to be inserted which delays the input signals for one time unit.

The structure of system 2 after inserting the learning-matrix dipole is shown in Figure 4.

The adjustment procedure then will be modified in the following way: If a state Z_t is given differing from Z_T, a sequence of test signals X_t has to be produced. First, the right-hand learning-matrix determines whether a row corresponding to state Z_t (if generally already present) has stored a valuable X_t or not. If such an X_t is found, it will be applied to system 1, in the other case the generator of test signals will be stimulated

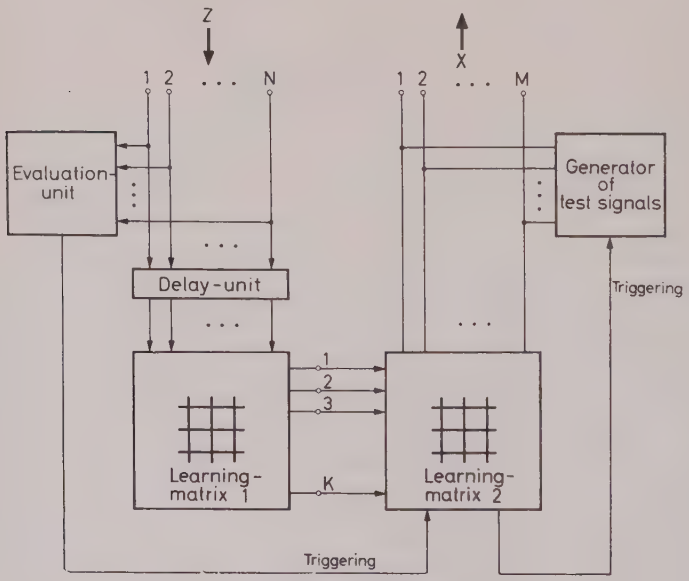


Fig. 4. Modified System 2

to produce a new one. If this has been proved valuable, that X_t will be stored together with Z_t in the dipole. This rule will be applied until Z_T has been found.

In addition to the described function of the dipole as a translator, we can use similarity decoding for further improvement of the system's performance if the following condition holds: States of system 1 differing only little with respect to the adjustment criterion are also similar with respect to the binary code used; that means the mutual Hamming distances don't exceed a given threshold value. It may be possible to use either a fixed or a variable threshold value to adapt the classification to various circumstances. Therefore the input signals from system 1 can be classified using similarity decoding for building up a smaller number of classes.

The threshold value determines whether an input signal Z_t belongs to an already stored class of states or if it is part of a new class, which has to be stored in a new separate row of the learning-matrix dipole. Because of controlling the storage procedure of the dipole by a parameter of the system itself, this technique is sometimes called "automatic classification".[3]

Considering the application of the learning-matrix dipole in system 2, we note that the right-hand learning-matrix only is used to store binary signals and to act as a common translator, while the decoding properties remain unused. Further specifications of the input signals of system 1 are then necessary to extend the abilities of the right-hand matrix to improve system performance. Above we made plausible restrictions about the coding of the states Z_t of system 1. Consequently we make similar assumptions on the input signals of system 1, and we define that test signals X_t with small mutual Hamming distances are also similar with respect to their success measured by the adjustment criterion. This quality may improve the production of new test signals X_t in the following way: Given a certain $Z_t - X_t$ (already stored in the dipole), the evaluation unit controls the right-hand matrix to determine whether a valuable X_t has been stored or not. If this inquiry fails, the generator of test signals delivers a new X_t, that hitherto would be applied to system 1 without any further control of its value. With the mentioned restrictions on the test signal presentation, new X_t signals are first offered to the right-hand matrix for similarity test with the already stored, useless X_t. Only if no or small similarity is found, the new X_t will become efficient to

system 1 and also will be stored in the dipole in the appropriate row. Repeated worthless test signals therefore will be omitted and the performance of the generator of test signals is improved.

The intention of this paper was to show in principle a possible application of the learning-matrix dipole as a part of a more complex adaptive system.

REFERENCES

1. Steinbuch, K. Die Lernmatrix, Kybernetik **1,** 36–45 (1961).
2. Steinbuch, K., Schmitt, E. Adaptive Systems Using Learning Matrices. This volume pp. 751–768.
3. Görke, W., Wettstein, H. Neuere Untersuchungen an Lernmatrizen. In: Neuere Ergebnisse der Kybernetik. Munich: Oldenbourg 1964.
4. Steinbuch, K. Adaptive Systems Using Learning Matrices, Kybernetik **2,** 148–152 (1965).
5. Steinbuch, K. Automat und Mensch, 3rd edit., Berlin, Springer 1965.
6. Steinbuch, K., Widrow, B. A Critical Comparison of Two Kinds of Adaptive Classification Networks. IEEE, Transact. on Electronic Computers vol. **EC-14,** 737–740 (1965).

EDWIN R. LEWIS

Librascope Group

General Precision, Inc.

Glendale, California

Synchronization in Small Groups of Neurons: A Study with Electronic Models

INTRODUCTION

If a group of muscle fibers is coordinated, one would generally expect coordination among the neurons innervating those fibers. Perfect synchrony in motoneuron firing would not generally be desirable, however, since it would lead to jerky motion. Examples of imperfect synchrony yet strong coordination have been found in flight systems of flies and locusts.[14,15] Motoneurons in these systems fire with certain phase relationships, independent of firing frequency. Since synaptic coupling between cells would introduce fixed latencies rather than fixed phase relations, it has been suggested that synchronization is brought about by electrotonic (syncytial) coupling which allows interactions of the subthreshold potentials. In another group of highly coordinated neurons, the crustacean cardiac ganglion, electrotonic coupling has been experimentally demonstrated.[4,11,12] Isolated cells of these ganglia tend to fire spontaneously at moderately high frequencies. In intact ganglia, however, this spontaneity is subdued; the cells produce periodic bursts of much lower frequency. Other examples of electrotonic coupling between neurons have been found in the sea hare—*Aplysia*,[1] the leech—*Hirudo*,[5] earthworms—*Lumbricus* and *Eisenia*,[13] puffer fish,[2,3] and hydra.[10]

Electronic models based on the squid-axon data have been found to faithfully simulate subthreshold neuronal activity as well as spikes.[8,9] These models are being used in a study of synchronization through electrotonic coupling. This paper describes results of experiments with a pair of models coupled through a resistance. This pair exhibits stable alternation of firing which seems to be akin to the fixed-phase firing observed in locusts and flies.

THE MODEL

The model used in this study was based on the squid-axon data of Hodgkin and Huxley.[6] Individual electronic circuits in the model simulate the voltage and time dependencies of potassium and sodium ion fluxes across a single patch of neural membrane. The dispensation of these circuits is shown schematically in Figure 1. The nonlinear, active filters indicated in the figure were designed with transfer functions to match the functional relationships between specific ionic conductances and membrane potential. Detailed discussions of these filters appear elsewhere.[7,8]

In the design of this model emphasis was placed on subthreshold phenomena, particularly those affecting the time of occurrence of a spike. The range of subthreshold membrane potentials is generally close to the equilibrium potential for potassium ion flux across the membrane, so changes in net potassium ion driving force had to be represented in the model. The net driving force on sodium ions in the subthreshold range, on the other hand, is varied by only a small fraction of its total magnitude; so it is represented by a constant in the model.

Several parameters in the model are variable. During the experiments described in this paper, all but one of these parameters were fixed, however. The more important parameter values are listed in Table I.With these settings, the model simulates the action of 0.25 sq cm of squid axon membrane with the voltages and currents increased by a factor of 100.

Hodgkin and Huxley computed extremely low values for the sodium conductance across the resting squid-axon membrane. When sufficiently

Table I: Fixed Parameter Values

V_L	—5.0 volt
V_K	—10.0 volt
V_{Na}	0.0 volt (irrelevant under the assumption of a constant driving force)
C_m	0.25 microfarad
G_L	0.06 millimho
G_K (at equilibrium)	0.06 millimho
G_K (maximum)	2.0 millimho
G_{Na} (at equilibrium)	0.01 millimho (variable)
G_{Na} (maximum)	2.0 millimho

larger values were employed in the model, spontaneous periodic spikes appeared. This "spontaneous" mode was used extensively in our experiments.

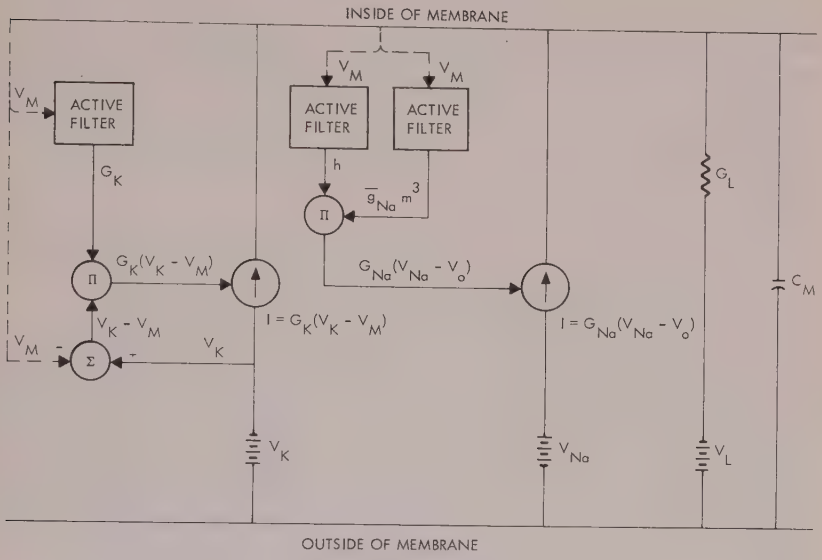


Fig. 1. Block diagram of the electronic model.

Our preliminary studies of electronic coupling have been conducted with the simplest configuration we could imagine. We assumed the extracellular resistance between two patches of spontaneously spiking membrane to be zero and the intracellular coupling to be purely resistive. Distributed capacitance in the electrotonic bridges was ignored. Two units of the type shown in Figure 1 were connected according to the plan of Figure 2. The sole parameter of coupling was the resistance, *R*.

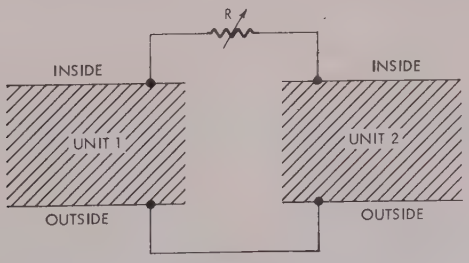


Fig. 2. Test configuration of models.

Extraneous coupling through a common power supply was avoided by an inelegant but expedient method; separate power supplies were employed. Other means of extraneous coupling, such as test equipment connected to both units, were present and unavoidable, however. To test for coupling through these other means, we connected every piece of test equipment we planned to use; we disconnected R; and we set both units to essentially identical spontaneous spike frequencies. The resulting cross-interval histogram showed no discernable correlation between the spike trains.

TWO MODES OF STABLE INTERACTION

With R disconnected, each unit was adjusted to approximately the same frequency. When R was subsequently connected, one of two things happened. The units either became synchronized, or they fired in stable alternation. The prevailing mode was determined by the relative phase of the two spike trains at the moment R was connected. Both modes shown in Plate I were obtained with all conditions identical except the timing of connection.

Believing that the mode of stable alternation might be related to the phase locking observed by Wilson and Wyman in the locust[14] and the fly,[15] we decided to examine it in detail. The remainder of this paper is essentially a discussion of the results of that examination, in which we asked the following questions: 1. Does the mode of stable alternation depend on precise adjustments of parameters in the model, or will it endure the fluctuations one expects in and about a patch of membrane? 2. Can the mechanisms responsible for this mode be rationalized in terms of the model's biological substrate, or are they indigenous to the electronic circuits?

PARAMETRIC RANGE OF STABLE ALTERNATION

With both units set to produce spontaneous spikes at mean intervals of 100 ms, the coupling resistor was connected, disconnected, then connected again until stable alternation was obtained. The simulated equi-

librium sodium current of Unit 1 was then increased, resulting in decreased spike interval for both units. The interval at which stable alternation disappeared was recorded for several values of coupling resistance. The process was repeated for decreased sodium current and increased intervals and for stable synchronization as well as stable alternation.

The data are displayed in Figure 3, which is essentially a map showing the ranges of the two stable modes with Unit 2 tending to produce spikes at 100-ms intervals but being perturbed by the action of Unit 1 through

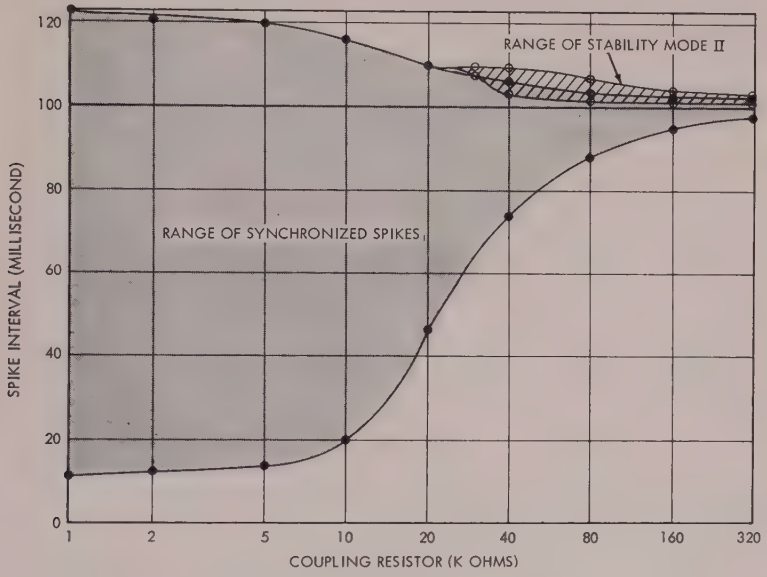


Fig. 3. Ranges of stable synchronization and stable alternation for resistively coupled models.

the coupling resistor R. The range of stable synchronization extends to both sides of the uncoupled spike interval of Unit 2, while the range of stable alternation (mode II in the figure) is entirely above that interval. In addition, stable alternation can exist only for intermediate values of R. At very high values all stable interaction disappears; at very low values stable synchronization prevails, stable alternation is untenable. The three preceding sentences are valid generalizations for all cross sections which we examined in the parameter space.

SUSCEPTIBILITY TO NOISE

With both units set to produce spontaneous spikes at 100-ms intervals, a 10^5-ohm coupling resistor was connected and stable alternation was obtained. Noise, limited by filters to a band from 20 to 50 cps, was added to the simulated sodium current of Unit 1. The time interval between each spike in Unit 2 and the succeeding spike in Unit 1 was measured automatically, and a histogram of these intervals was computed by a pulse-height analyzer. The resulting analyzer outputs, which we shall call cross-interval histograms, are shown in a series of photographs beginning with Plate 2.

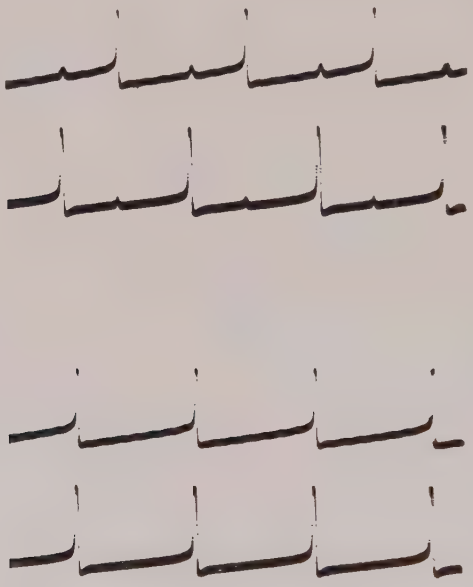


Plate 1. Top: Stable alternation of spikes in two resistively coupled models. Bottom: Stable synchronization in the same models.

Plate 2 shows histograms taken with the coupling resistor disconnected and without the applied noise voltage. The top trace is a spike-interval histogram for Unit 2; the middle trace is a spike-interval histogram for Unit 1; and the bottom trace is the cross-interval histogram, which shows no discernable correlation between spikes in Unit 1 and those in Unit 2. The same series of histograms is shown in Plate 3, but in this case the 10^5-ohm coupling resistor was connected. The cross-interval histo-

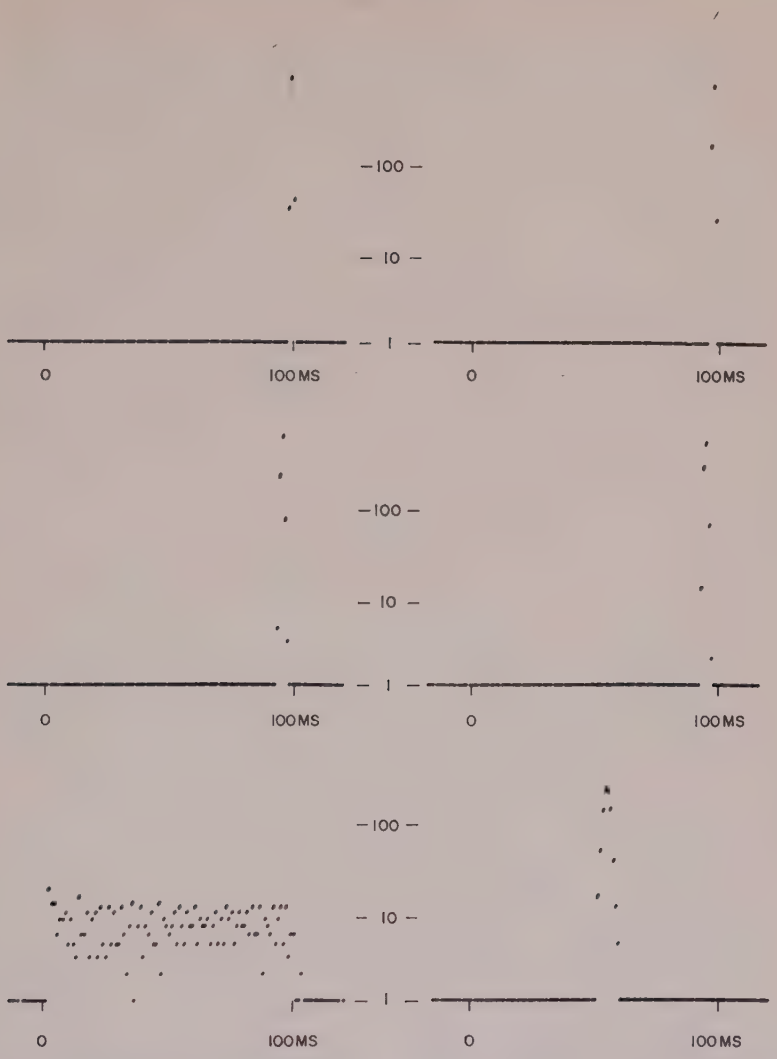


Plate 2. Histograms of Units 1 and 2 uncoupled.

Plate 3. Histograms of Units 1 and 2 in stable alternation.

Top: Spike interval histogram of Unit 2
Middle: Spike interval histogram of Unit 1
Bottom: Cross interval histogram, intervals from each spike in Unit 2 to the next spike in Unit 2 to the next spike in Unit 1.

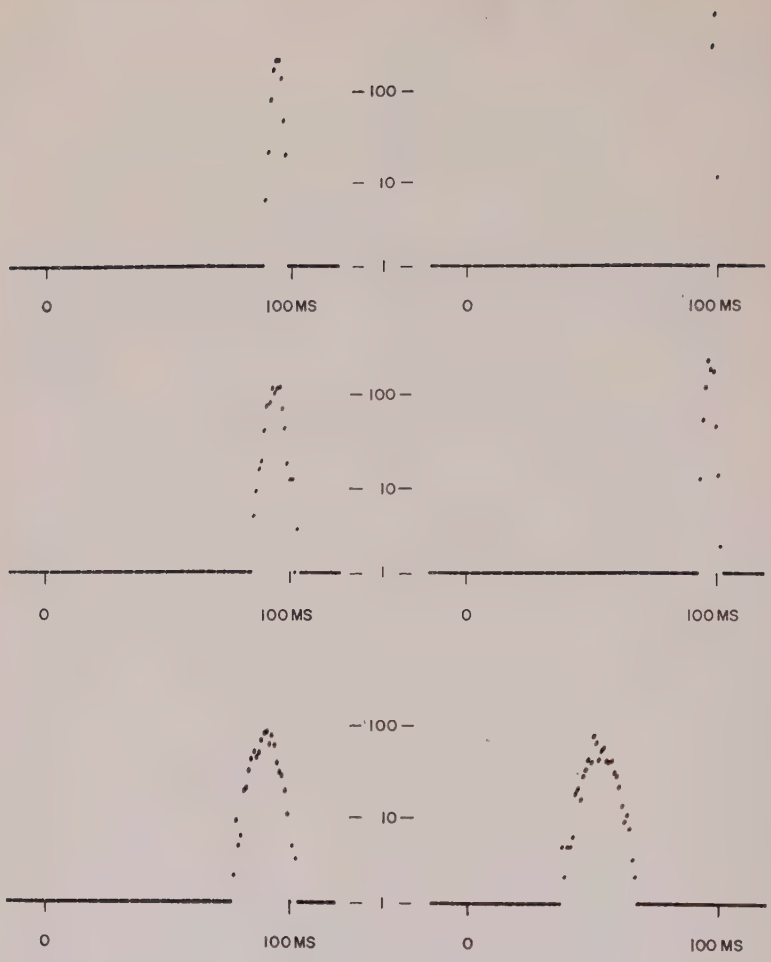


Plate 4. Interval histograms of Unit 1 after applications of noise, but before coupling.

Plate 5. Histograms of Units 1 and 2 coupled, with noise applied to Unit 1.
Top: Spike interval histogram of Unit 2
Middle: Spike interval histogram of Unit 1
Bottom: Cross interval histogram, intervals from each spike in Unit 2 to the next spike in Unit 1.

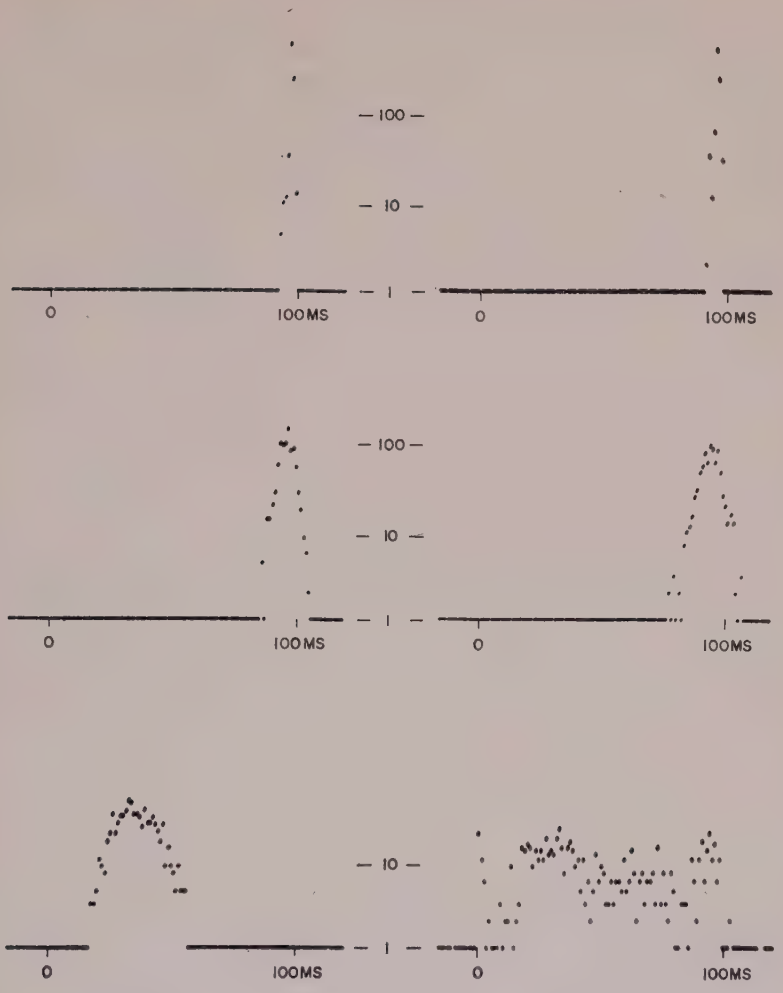


Plate 6. Histograms of Units 1 and 2 coupled, with noise applied to Unit 1.

Plate 7. Histograms of Units 1 and 2 coupled, with noise applied to Unit 1.

Top: Spike interval histogram of Unit 2
Middle: Spike interval histogram of Unit 1
Bottom: Cross interval histogram, intervals from each spike in Unit 2 to the next spike in Unit 1.

gram shows stable alternation, with all spikes from Unit 1 occurring in a narrow spike-phase range of Unit 2.

Plates 5, 6, and 7 show the same histograms with noise applied to Unit 1. The histograms in Plate 5 were taken with a low noise level. Those in Plate 6 were taken at approximately the maximum noise level for which stable alternation occurred. Stable alternation has disappeared in Plate 7; the cross-interval histogram shows only remnant phase preferences between the two units, and the most pronounced preference is spike synchrony, not alternation. Plate 4 shows the effects on uncoupled Unit 1 of the noise levels responsible for the histograms in Plates 5, 6, and 7.

MECHANISMS FOR STABLE ALTERNATION

With Unit 2 set to produce spontaneous spikes at 100-ms intervals and Unit 1 set to produce spikes at a mean interval of 1 sec but with considerable variation (induced by noise), a 10^5-ohm coupling resistor was connected. Spikes of Unit 1 occurred at various phases of the spike train from Unit 2. The effects of the phase of a spike from Unit 1 on the length

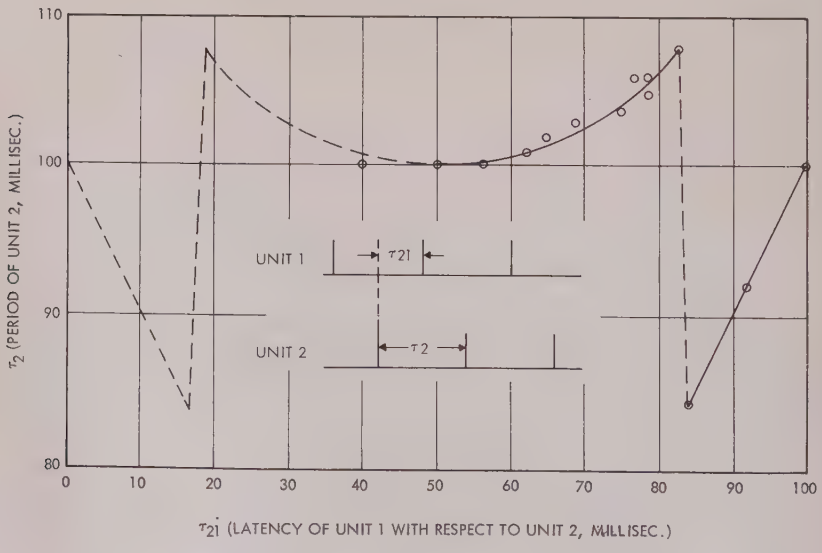


Fig. 4. Effects of spikes in Unit 1 on spike interval of Unit 2.

of the concurrent spike interval of Unit 2 were observed. The data are displayed in Figure 4.

If a spike from Unit 1 occurred early in the spike interval of Unit 2, it had little effect on that interval. If the Unit 1 spike occurred later, it prolonged the interval. If it occurred near the end of the interval, it instigated a spike in Unit 2, shortening the interval. All of these phenomena are apparent in the solid line of Figure 4.

The dashed line is a left-to-right reversal of the solid line. It shows the length of spike interval in Unit 2 as a function of the expected latency between a spike in Unit 1 and the next spike in Unit 2. If the latency is expected to be very short (based on the 100 ms period of Unit 2), it will in fact be zero, and the Unit 2 spike interval will be shortened. A somewhat longer expected latency will cause a prolonged spike interval in Unit 2. A very long expected latency will have no effect.

When each unit is uncoupled and set to produce spontaneous spikes at 100-ms intervals, the two units are essentially identical. The curves in Figure 4 are valid, therefore, when the units are interchanged so that the phase of occasional spikes from Unit 2 determines interval length in Unit 1. The superposition of the dashed and solid curves yields an approximation to the interactions when both units are set to produce spikes at 100-ms intervals and are coupled by a 10^5-ohm resistance. Let the abscissa be the time one would expect a spike from Unit 1 to occur after a spike from Unit 2 if the two units were not coupled. The dashed portion of the curve gives the perturbation of that time brought about by coupling; the solid portion gives the perturbation of firing time of the other unit, Unit 2.

From the composite curve of Figure 4 it is apparent that one of two things will happen when Units 1 and 2 are both firing at 100-ms intervals and are suddenly coupled by a 10^5-ohm resistor. If the expected value of τ_{21} is between 0 and 16 ms or between 84 and 100 ms, the two spike trains will become synchronized immediately. If, on the other hand, the expected value of τ_{21} is between 18 and 50 ms, the spike intervals of Unit 1 will be lengthened while those of Unit 2 remain unchanged. If τ_{21} is between 50 and 82 ms, the spike intervals of Unit 2 will be lengthened while those of Unit 1 remain unchanged. In both cases τ_{21} will approach 50 ms and stable alternation will occur. Stable alternation is implicit, therefore, in the solid curve of Figure 4 (i.e. in the function $\tau_2(\tau_{21})$).

Some aspects of the function $\tau_2(\tau_{21})$ can be explained quite simply by means of the photographs in Plate 1. Following a spike, the unit that fired undergoes hyperpolarization and subsequent slow depolarization until threshold is reached and another spike occurs. If a spike occurs in Unit 2 during the slow depolarizing phase of Unit 1, it produces a biphasic response. Unit 1 undergoes brief, rapid depolarization followed by residual hyperpolarization. If the brief depolarization does not exceed threshold, the hyperpolarization prevails, and the net effect is inhibitory. Thus if the spike in Unit 2 occurs sufficiently late in the slow depolarizing phase of Unit 1, Unit 1 will fire; if the spike occurs earlier, the spontaneous firing of Unit 1 will be delayed. The electrotonically coupled spike is thus inhibitory or excitatory, depending on its time of occurrence.

The magnitude of the hyperpolarizing current in Unit 1 following a spike in Unit 2 depends on the difference in the potentials of the two units. This difference increases as Unit 1 becomes more depolarized; so the electronically coupled spike is more inhibitory if it occurs later in the slow depolarizing phase of Unit 1. This accounts for the positive slope of $\tau_2(\tau_{21})$ between 50 and 82 ms.

DISCUSSION

We have answered our two questions about stable alternation. It is not particularly sensitive to random fluctuations of spike interval, nor is it dependent on precise adjustments of parameters; and it can be explained quite easily in terms of known neural properties.

A question which we have not answered is "will stable alternation endure large perturbations of spike frequency induced by synaptic inputs applied simultaneously to both units?" This question must be answered before we can relate our results to the findings in the locust or the fly.

ACKNOWLEDGMENTS

Research sponsored by Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Contract No. AF 49 (638)-1232 and by the Aerospace Medical Research Laboratories, Aero-

space Medical Division, Air Force Systems Command, United States Air Force, Contract No. AF 33 (615)-2464. Further reproduction is authorized to satisfy needs of the U. S. Government.

REFERENCES

1. Arvanitaki, A., and Chalanozitis, N. Interactions electriques entre le soma geant *A* et les somata immediatement contigus (Ganglion pleurobranchial d'Aplysia). *Bull. Inst. Oceanog.* No. 1143, 1–30 (1959).
2. Bennett, M. V. L. Electrical connections between supramedullary neurons. *Fed. Proc.* **19**, 282 (1960).
3. Bennett, M. V. L. Comparative physiology of supramedullary neurons. *Biol. Bull. Woods Hole* **119**, 303 (1960).
4. Hagiwara, S., Watanabe, A., and Saito, N. Potential changes in syncytial neurons of lobster cardiac ganglion. *J. Neurophysiol.* **22**, 554–572 (1959).
5. Hagiwara, S., and Morita, *H.* Electrotonic transmission between two nerve cells in leech ganglion. *J. Neurophysiol.* **25**, 721–731 (1962).
6. Hodgkin, A. L., and Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerves. *J. Physiol.* **117**, 500–544 (1952).
7. Lewis, E. R. Neural analog studies III.: Synaptic transfer function. Semi-annual Report No. 7, Laboratory for Automata Research. Glendale: Librascope Group, General Precision, Inc., 1964, p. 1.
8. Lewis, E. R. An electronic model of the neuron based on the dynamics of potassium and sodium ion fluxes. In: *Neural Theory and Modeling*, edited by R. F. Reiss Stanford: Stanford University Press, 1964, p. 154.
9. Lewis, E. R. Neuroelectric potentials derived from an extended version of the Hodgkin-Huxley model. *J. Theoret. Biol.* **10**, 125–158 (1966).
10. Spangenberg, D. B., and Ham, R. G. The epidermal nerve net of Hydra. *J. Exp. Zool.* **143**, 195–201 (1960).
11. Watanabe, A., The interaction of electrical activity among neurons of lobster cardiac ganglion. *Jap. J. Physiol.* **8**, 305–318 (1958).
12. Watanabe, A., and Bullock, T. H. Modulation of activity of one neuron by subthreshold slow potentials in another lobster cardiac ganglion. *J. Gen. Physiol.* **43**, 1031–1045 (1960).
13. Wilson, D. M. The connections between the lateral giant fibers of earthworms. *Comp. Biochem. Physiol.* **3**, 274–284 (1961).
14. Wilson, D. M., and Wyman, R. J. Motor output patterns during random and rhythmic stimulation of locust thoracic ganglia. *Biophys. J.* **5**, 121–143 (1965).
15. Wyman, R. Probabilistic characterization of simultaneous nerve impulse sequences controlling dipteran flight. *Biophys. J.* **5**, 447–471 (1965).

R. G. RUNGE

M. UEMURA

and S. S. VIGLIONE

Pattern Recognition Research Department, Astropower Laboratory

Missile and Space Systems Division, Douglas Aircraft Co., Inc.

Newport Beach, California

Electronic Synthesis of the Neural Networks in the Pigeon Retina

INTRODUCTION

The objectives in modeling retinal neural networks are twofold. One purpose is to augment current capabilities for processing non-numeric visual information by borrowing ideas from deterministic retinas such as that of the pigeon. In parallel it is hoped that by studying and simulating these networks new insight into the actual mechanism by which such networks function may be obtained. It is felt that to accomplish these objectives it is necessary that any postulated models be based on neurophysiology and neuroanatomy.

There are approximately 988,000 optic fibers in the pigeon retina.[1] The majority of these fibers represent ganglia neurons which output to the brain center of the pigeon. These ganglion cells have been divided into six classes in accordance with the input light stimulus which produces a maximum response. These six classes, recently described by H. R. Maturana,[2,3] are designated as the following type of detectors: (1) Luminosity, (2) General Edge, (3) Convex Edge, (4) Horizontal Edge, (5) Vertical Edge, and (6) Directional Edge. These detectors have been observed by recording the activity of single retinal ganglion cells from the cut and intact optic nerve fibers of curarized pigeons monitored by means of metal filled micropipettes.

This paper describes a postulated model of one of these detectors, the directional edge detector. The postulated model has been constructed using electronic analog circuitry to simulate sensory and data processing neurons. The resultant model is found to produce the bulk of the stimulus/response functions described for the directional detectors in the literature.

SPECIFICATIONS FOR THE DIRECTIONAL MODEL

Figure 1 illustrates the stimulus/response performance reported by Maturana for the directional edge ganglions which form approximately 30 per cent of the accessible cell population. This physiological data establishes the minimal functional test specifications for the model.

The responsive receptive field (RRF) of the directional detectors is defined as the region from which a response is elicited when an edge is moved in the exclusive direction. The RRF determined in this fashion is from 55 to 110 microns and is indicated by the dash lines in Figure 1.

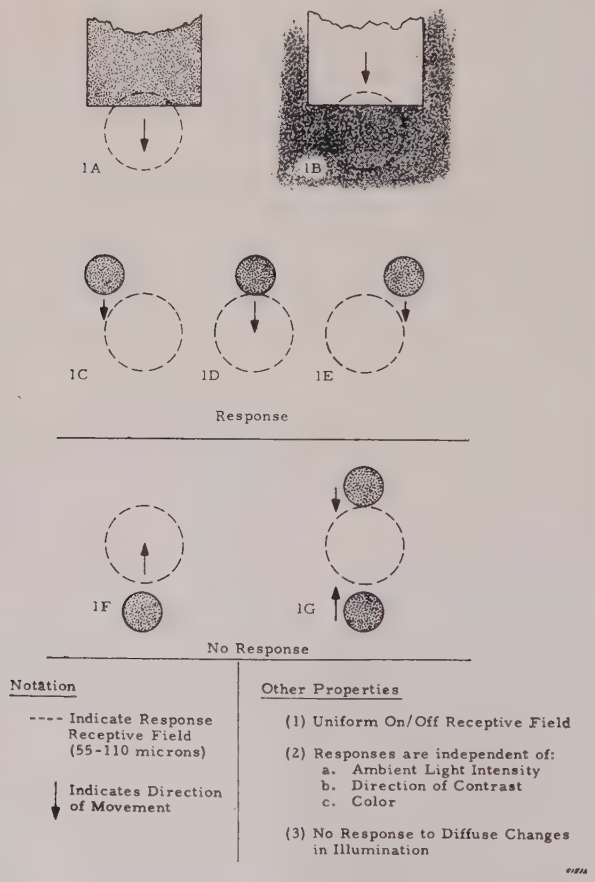


Fig. 1. Physiological response of the direction detector ganglions

1A and 1B of Figure 1 indicate that the cell responds to an edge in the exclusive direction and is independent of the direction of contrast. This is also true for a moving spot. In addition it is found that a response occurs immediately upon entry of the RRF for a light edge. The response for a dark edge occurs after appreciable entry into the RRF. The response of a dark spot moving through the center of the RRF in the exclusive direction (1 D) is significantly larger than that for the spot moving through the peripheral regions (1 C and 1 E). This implies that a reduction of ganglion excitation occurs. If a dark or light spot is moved in opposition to the exclusive direction (1 F) the ganglion has either no response or a very minimal response. This fact definitely implies that the direction opposing the exclusive direction provides at least temporary inhibition. From the simultaneous spot data (1 G) it is apparent that the inhibition from the nonresponsive direction is in excess of the excitation from the exclusive direction assuming a zero threshold ganglion. The directional response of this class of detector may be increased or decreased by variations of ambient intensity, direction of contrast and color of the spot or edge but the general mode of response remains unaltered. Internal to the RRF a uniform ON/OFF response to a spot of light is obtained. However, if the ambient light level is drastically increased or decreased no response is obtained.

ANATOMY OF THE PIGEON RETINA

Reduced to its most essential elements the retina of all vertebrates may be considered as consisting of three superimposed layers of neurons which form synaptic connections in two plexiform layers (see Fig. 2 and 3).

1. Outer Nuclear Layer

This layer contains the cell bodies of rods and cones, the light receptors of the retina. These cells typically lie in two regular layers. The bodies of these cells vary in size from 2.2 μ at the fovea to 4.4 μ at 3 mm from the fovea. The output of the receptors form synaptic junctions in the outer plexiform layer mainly with bipolars of the inner nuclear layer. Birds such as the duck, chicken, pigeon and canary all possess retinas containing both rods and cones, with cones dominating. In general the output spread of rod sphericals is larger than that of cones.

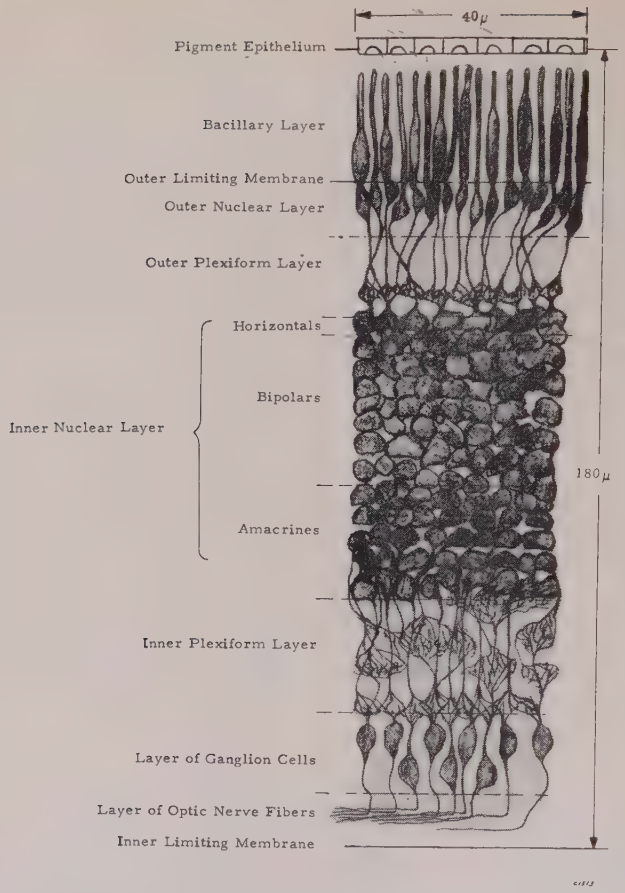


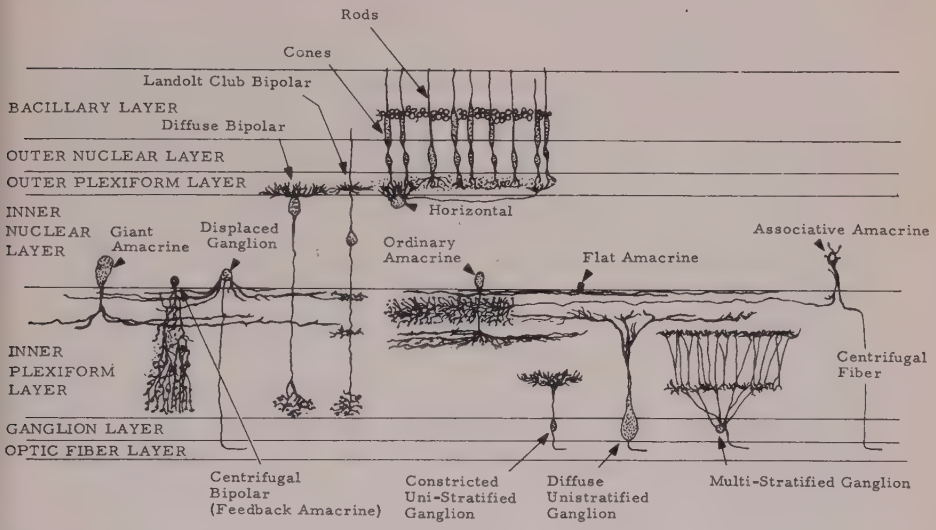
Fig. 2. Cross section of pigeon retina (3 mm from fovea).

2. Inner Nuclear Layer

This region is further subdivided into the horizontal layer, bipolar layer and amacrine layer. Adjacent to the outer plexiform layer, except near the fovea, lies one very regular layer of horizontal cells. The output axonal spread (teledendrons) of these cells synapse on the end feed of receptors. Two types are found in birds, those of short and long axon length. The long axon type are found infrequently.

At 3 mm from the fovea, seven layers of bipolars are found primarily between the horizontal cell layer and the amacrine cell layer. Cajal[4]

divides the bipolars of birds into two classes, those with and without Landolt's club (see Fig. 3). The Landolt club type of bipolar is very abundant in the pigeon. In general, the more diffuse type of bipolar (without Landolt club) is thought to be primarily associated with rods. The bipolar outputs to the inner plexiform layers. The Landolt club type generally emits several axonal outputs stratifying at various layers of the inner plexiform.The diffuse type bipolar emits a single axonal spread near or on the ganglion cell bodies.



(Courtesy Ramon y Cajal)
Histologue Du Systeme Nerveux
Paris; Maloine 1909-1911

Fig. 3. Pigeon neural anatomy.

Between the outer limit of the inner plexiform and the bipolar layers is the amacrine layer. These cells have not been shown to have axonal outputs with the exception of the associative amacrine. In birds the quantity of amacrine cells is large relative to that of monkeys and man. At 3 mm from the fovea they occur in five layers. Polyak[5] has shown that in monkeys amacrines have projections to the outer plexiform terminating on receptors. Five types of amacrines are found. They include ordinary or parasol amacrines, flat amacrines, giant amacrines, feedback amacrines

and associative amacrines. (In addition displaced ganglia are found in this layer). The associative amacrine is unique in that it has both a dendritic input and an axonal output in the inner plexiform layer. The cell bodies of this layer are from 6 to 7 μ in diameter with the exception of displaced ganglions and giant amacrines which have cell bodies varying from 9 to 15 μ. The ordinary amacrines have a stratification in the inner plexiform which spreads from 100 to 200 μ.

3. Ganglion Layer

This layer of cells projects back to the optic tectum and lateral geniculate of the pigeon. Their dendrites are located in the inner plexiform layer. The majority of these cells may be broken into three classes. Some of the cells are single stratified and constricted, others are single stratified and diffuse and still others are multistratified. Constricted and diffuse refer to the size of the dendritic input spread. The bodies of these cells range in size from 7 to 20 μ. All of the fibers in the pigeon are myelinated. The dendritic spread of a particular type of ganglion is reasonably constant, however if all types are included, spreads from 30 to 600 μ can be found. The following is an estimate of the number and typical density of each type of cell. These estimates are based on Chievitz's[6] cell counts, the total number of ganglion cells* in the retina and an estimate of the total retinal area (118 mm^2).†

Cell Type	Total Population	Typical Area Per Cell (micron)^2/cell
Receptors	1,920,000	61.4
Horizontals	1,020,000	116.0
Bipolars	6,100,000	19.3
Amacrines	3,680,000	32.1
Ganglions	979,000	120.6

* This number is taken to be 979,000 from Bruesch and Arey,[1] in conjunction with Cowan and Powell.[7]

† The retinal area is estimated from Chievitz's measurement using a hemispherical approximation.

POSTULATED NEURAL NETWORK OF THE DIRECTIONAL GANGLION

Figures 4 and 5 illustrate the connectivity and excitatory/inhibitory arrangement of the directional model. Figure 4 shows the inputs to a bipolar. The receptors in the immediate area are considered excitatory. The horizontals in the surrounding region are considered inhibitory, when their input receptors are illuminated. The receptor's outputs are implemented by means of analog voltages proportional to the logarithm of the light intensity. Horizontals are of opposite polarity to receptor outputs and receive inputs exclusively from cones. The input sum to the bipolar, neglecting this cell's temporal integrating capabilities, is given by

$$B = \frac{\sum_{i=1}^{m} R_i - \sum_{j=1}^{n} H_j}{m + n} \tag{1}$$

where

m = number of receptor inputs
n = number of horizontal inputs
B = static bipolar input sum
R_i = i-th central receptor
H_j = j-th surrounding horizontal

The bipolar is assumed to be excited if $B > 0$. Each bipolar is arranged so that $m = n$, that is, the number of horizontal inputs equal the number of cone inputs. Notice that if a spot is turned "on" over the bipolar's center the cell will respond due to the increase in cone excitation. If a dark spot is placed in the surround while an ambient intensity is maintained on the central receptors this will also cause the bipolar to respond due to the decrease in horizontal inhibition. An increase or decrease in diffuse light will cause the input sum to cancel and $B = 0$. The resulting bipolar's field is shown in Figure 4. It should also be pointed out that the bipolar will respond to an edge independent of the direction of contrast.

The operation of the directional ganglion is easily understood with reference to Figure 5. Assume a spot (light or dark) originates from the left and moves to the right. As it does so the associate amacrine AA is excited by its corresponding bipolar. The output of this cell inhibits the ordinary amacrines in the central region of the ganglion field. This in turn prevents these cells from inhibiting the directional ganglion when the

central bipolars are excited. As the spot moves to the next group of bipolars their corresponding ordinary amacrine activity begins to inhibit the ganglion. However, this inhibition is rapidly followed by a large volley of excitation from the central bipolars. This causes the ganglion's input sum to exceed its zero threshold and ganglion response results. If, however, the spot enters from any direction other than that which excites the associative amacrine the ganglion will not respond.

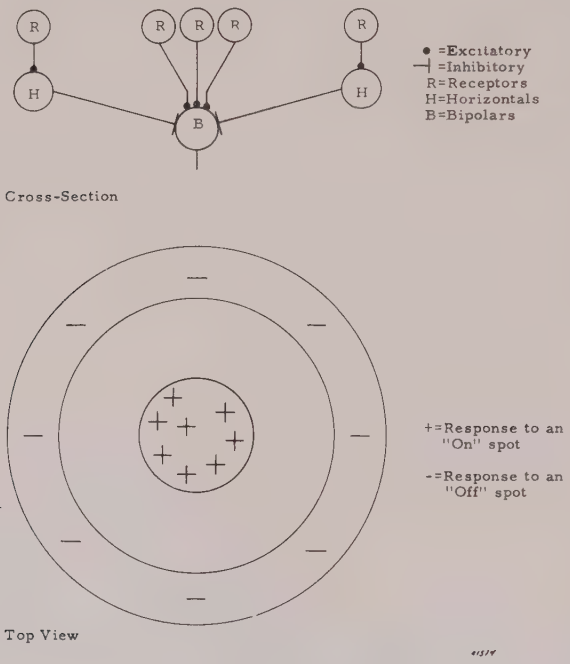


Fig. 4. Receptive Field of bipolars.

Now let a spot of light be shown "on/off" in the central region of the ganglion's field. The "on/off" bipolars will become excited. This will produce an increase in the ganglion's input sum causing the ganglion to "fire." The response will not be sustained since soon after the bipolar's excitatory volley the inhibitory central amacrines will respond to the bipolar volley. Since all bipolars are unresponsive to diffuse illumination no response will occur at the ganglion level for changes in ambient light intensity.

It should be pointed out that the angle over which the cell will respond decreases as the axon length of the associative amacrine increases.

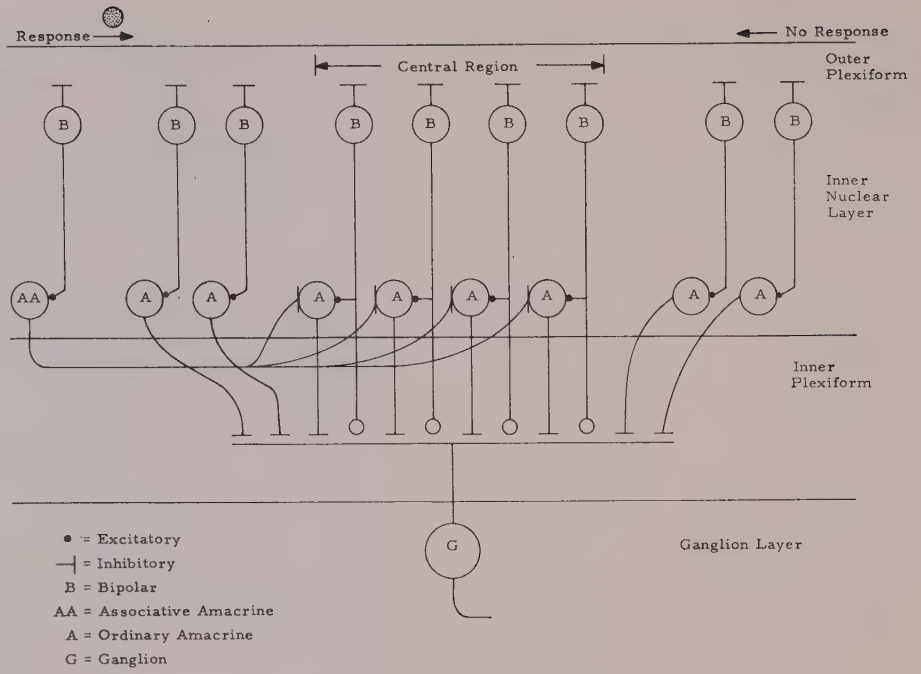


Fig. 5. Directional Ganglion Cross-Section

Further, the ganglion has a minimum and maximum velocity for which the moving spot or edge will be effective. If the spot is moved too slowly the inhibiting effect of the associative amacrine on ordinary amacrines may become ineffective. Since the central excitation is preceded by surrounding inhibition this will be particularly true for an edge, where maximum inhibition precedes excitation. If an edge is moved too rapidly, the surrounding inhibition may not have time to decay, hence the central excitation may be inadequate. If further investigation of these cells shows that they are independent of the speed of motion of the object in the sensitive direction it will be necessary to delete surrounding inhibition in the responsive direction.

CONSTRUCTION OF THE MODEL

Figure 6 is an illustration of the resulting directional motion detector model. The model contains 76 cone cells, 76 horizontals, 25 bipolars, 24 ordinary amacrines, 1 associative amacrine and one ganglion. The model is constructed on 18 vector cards. Each card has 43 finger contacts

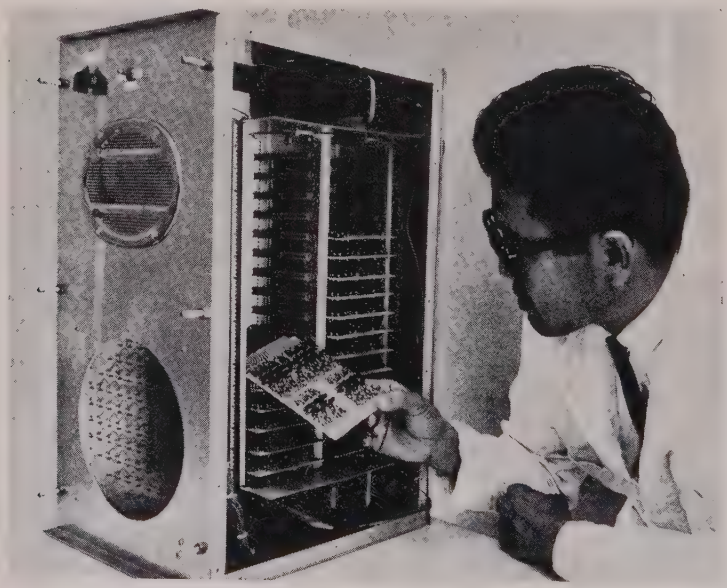


Fig. 6. Model of directional motion detector.

which mount into a taper pin connector. Taper pin connectors are employed to allow maximum flexibility for interconnection of neurons. The housing for the model allows for 32 circuit boards in order that additional ganglion classes may be modeled utilizing the same sensory array and a portion of the other neuron analogs.

A schematic of the cone cell analog is shown in Figure 7. The output of the cone utilized by this model is essentially digital in nature. This is in contrast to the postulated analog sensor. Digital cones were employed due to the lack of uniformity of the H-11 silicon photodiode light sensor. The output of the cone is essentially a +12 volts when the light intensity is greater than 1500 foot candles and zero when the illumination is less than 50 foot candles. The output of the horizontal cell is the logical inverse of that of the cone. The horizontal output is not implemented by means of opposite polarity as postulated. The reason for this discrepancy stems from the lack of uniformity and predictability of the light sensors. This hardware simplification has caused the resulting directional model to elicit a response to the "on and off"of diffuse light and to have a dependence upon the ambient background when a spot or edge is moved in the exclusive direction.

Each bipolar analog (Fig. 8) is stimulated by eight horizontals (inhibitory) and one cone (excitatory) with the weight of the central receptor

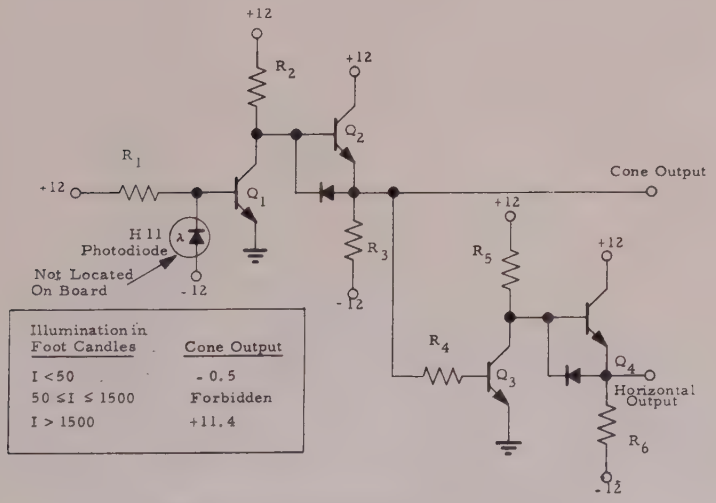


Fig. 7. Cone and horizontal cell analogs.

being eight times greater than that of the surrounding horizontals. The postulated model with $m = n$ is essentially satisfied by this weighting technique. This results in an appreciable savings in the number of receptors required to implement the model. The bipolar performs a weighted summation of these nine inputs and if the rate of change of the input sum is positive, a $+12$ volt pulse, 40 msec in duration is generated (the 40 msec pulse is used to simulate a number of bipolars being stimulated simultaneously and to reduce the number of analogs required).

The essential circuitry of the bipolar is a complementary collector monostable multivibrator with an input summing network and triggering circuit. One drawback of the bipolar circuit is that it does not employ a time integral on the input sum prior to activating the bipolar. This cir-

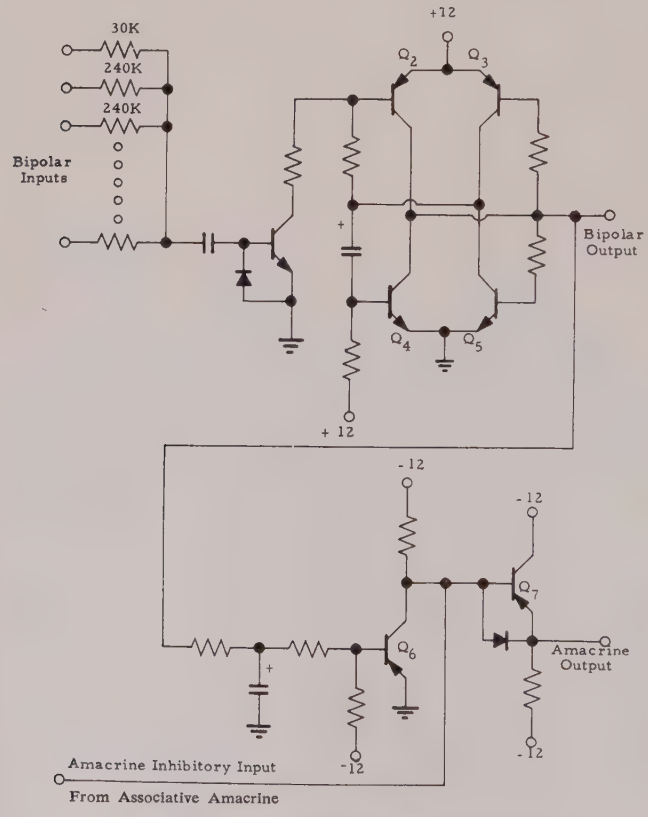


Fig. 8. Bipolar and Regular Amacrine Analogs.

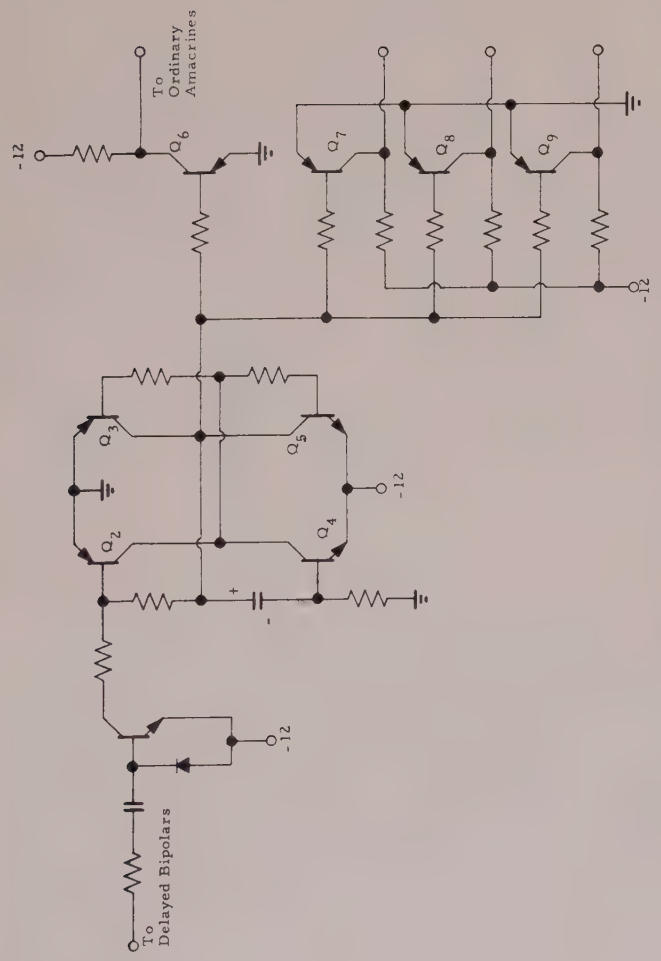


Fig. 9. Associative amacrine analog.

cuit is termed a dynamic bipolar. Anatomically it could bear some relation to the Landolt club bipolars.

The ordinary amacrines receive their inputs directly from bipolars and generate a negative 12 volt pulse of 40 msec duration delayed by one msec with respect to the positive going output of the bipolar. It is mechanized by means of a symmetrical RC delay network followed by an inverter. In addition to receiving an input from the bipolar, the ordinary amarine may have its activity inhibited by the associative amacrine. In the electronic model the associative amacrine (Fig. 9) receives its inputs from two bipolars. The output of this cell normally rests at -12 volts and acts as a parallel load of high resistance to the collectors of the ordinary amacrines. When the associative amacrine is stimulated it clamps the

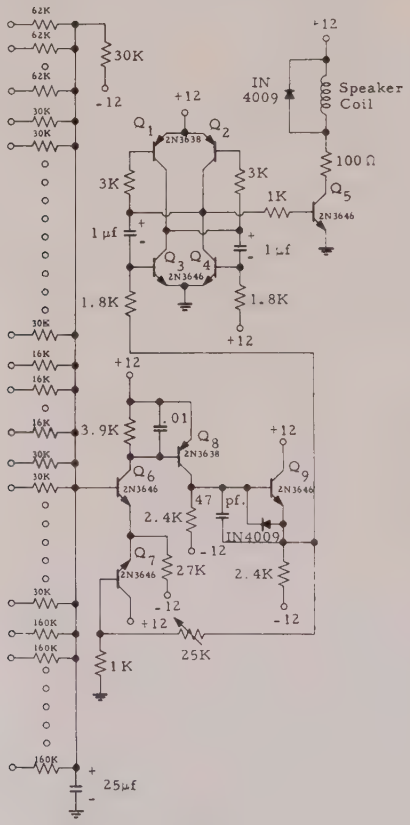


Fig. 10. Ganglion cell analog.

output of ordinary amacrines to zero volts for a period of 200 msec thereby rendering the ordinary amacrines ineffective. However this inhibition of amacrine activity is not put into effect until 80 msec after the bipolars stimulating the associative amacrine are excited. This is done to account for the delay associated with the long axons of the associative amacrine. The associative amacrine inhibits the activity of four ordinary amacrines at the center of the ganglion's receptive field.

The ganglion cell (Fig. 10) performs a weighted summation of its inputs from bipolars and ordinary amacrines and if the Σ excitation (bipolars) − inhibition (amacrines) >0 it generates a train of pulses of 1 msec duration. The pulse repetition rate increases in proportion to the amount by which the input sum exceeds zero. The largest weights are used to simulate cells in the central region of the ganglion receptive field. The weights were established by determining the amount of overlap between axonal spread (bipolars and amacrines) and ganglion dendritic spread. The ganglion neuron is preceded by an amplifier to allow for attenuation loss of the large input summing network. The output from the ganglion cell is applied to a speaker so that activity of the cell can be monitored.

DISCUSSION

The constructed model has successfully demonstrated the ganglion response to the directional motions of a spot or edge, both with a dark edge (spot) on an illuminated field and a light edge (spot) on a dark field and is being modified to incorporate additional functions reported in the literature.[2,3,8]

Additional modeling work is in progress to devise a model using a consistent set of interconnections and excitatory/inhibitory neuronal functions to implement all of the six ganglion classes. The model should be completed and tested within the year.

ACKNOWLEDGEMENTS

The assistance of C. C. Kesler and A. G. Mucci of the Astropower Laboratory in deriving the conceptual form of the directional motion detector and determining the probability of connections between lateral

and transverse neurons is greatly appreciated. In addition many discussions held with Dr. R. L. Binggeli of the School of Medicine at the University of Southern California proved invaluable in defining neuron functional responses and in analyzing and interpreting the pertinent literature.

REFERENCES

1. Bruesch, S. R., and Arey, L. B. The Number of Myelinated and Unmyelinated Fibers in the Optic Nerve of Vertebrates, *J. Comp. Neurol.*, **77**, 631–665 (1942).
2. Maturana, H. R. Directional Movement and Horizontal Edge Detectors in the Pigeon Retina, *Science* **142,** 977–979 (15 Nov. 1963).
3. Maturana, H. R. Functional Organization of the Pigeon Retina, "*Symposium on Information Processing in the Nervous System*, **22***nd Internat. Congress of Physiological Science*," III, 170–180 (1962).
4. Ramon y Cajal, S. Histologie du Systems Nerveux de l'Homme et des Vertebrate, Tome II, Talleres Graficos Montana, Madrid, Spain, 1955.
5. Polyak, Stephen L. "The Retina," University of Chicago Press, Chicago, Illinois, 1941.
6. Chievitz, J. H. Untersuchungen ueber die Apea Centralis Retinae, *Arch. Anat. Entwicklungsgesch.* Suppl. **139** (1889) 139–196.
7. Cowan, W. M., and Powell, T. P. S. Centrifugal Fibers in the Avian Visual System, *Royal Society Proceedings*, Ser. B., **158**, 1963
8. Wylie, R. M. Responses of Neurons in the Optic Tectum of the Pigeon, Ph. D. Thesis, Harvard University, Cambridge, Mass., April 1962.
9. Maturana, H. R. Synaptic Connections of the Centrifugal Fibers in the Pigeon Retina, *Science* **150,** 15 Oct. 1965.
10. Van Buren, J. M. "The Retinal Ganglion Cell Layers," Charles C. Thomas Publisher, Springfield, Illinois, 1963.

APPENDIX

PREDICTING THE NUMBER OF DIRECTIONAL DETECTORS

This model has one associative amacrine from the surrounds that inhibits activity of the normally inhibitory ordinary amacrines in the central area of a ganglion's receptive field. As the central area of the ganglion is approached the central bipolars cause the directional ganglion's input sum to exceed its firing threshold since the normally inhibitory effect of the ordinary amacrines in this region has been inhibited. In essence the lack of an associative amacrine inputting to ordinary ama-

crines in a ganglion's central area implies that such a ganglion cell possesses no directional properties and would yield ganglions similar to the general edge detectors. If one associative amacrine cell influences a ganglion by means of these central ordinary amacrines it is assumed that it is a one-directional detector. If two associative amacrines influence a ganglion in this manner the ganglion may have more than one direction of selectivity. This may be the mechanism for vertical and horizontal edge detectors.

As previously stated, recent physiological data indicate that directional movement detectors form about 30 per cent of the accessible ganglion cell population. If the assumption that the associative amacrine is the cause of directional ganglions is correct then it should be possible to predict that about thirty per cent of the accessible cell population has exactly one associative amacrine influencing its response. The following is a calculation of the probability that exactly k associative amacrines terminate on or near a ganglion's dendritic input. These calculations show good agreement between physiological data and the theory of the directional model.

Figure 11 is a drawing illustrating the axonal spread of the associative amacrine just tangent to the ganglion's dendritic spread. The dimensions

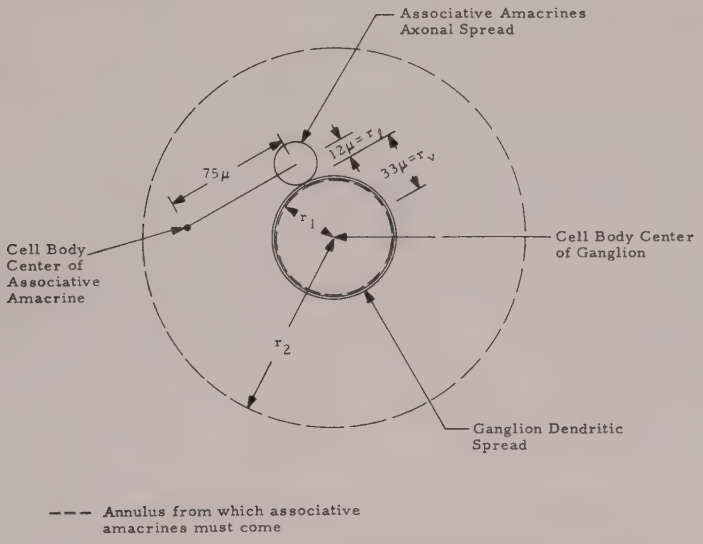


Fig. 11. Spatial relation of associative amacrines to a ganglion

of the drawing are estimated by scaling the drawings of Cajal.[4] The associative amacrine is assumed to inhibit the ordinary amacrines if its output spread is tangent to ganglion's spread. Circular distributions in axonal and dendritic spread are assumed.

P_c = probability that axonal spread of a laterally projecting neuron (associative amacrine, horizontal, etc.) terminates on or in the dendritic spread of a vertically projecting neuron (ganglion, bipolar, etc.).

$$P_c = \frac{\Theta}{\pi} = \frac{1}{\pi} \cos^{-1} \frac{a^2 + R^2 - b^2}{2aR} \tag{2}$$

where

a = axon length of laterally projecting cell (between cell body and axonal spread centers).

b = distance from the vertical cell body center to the axonal spread where the axonal spread is tangent to the dendritic spread.

$b = r_v + r_l$

r_v = input dendritic spread radius of the vertically projecting neuron.

r_l = output axonal spread radius of the laterally projecting neuron.

$R = \left(\frac{r_1^2 + r_2^2}{2}\right)^{1/2}$ = equal area radius as measured from the center of the vertical neuron.

r_1 = minimum radius to allow contact between vertical and lateral neuron (from the center of the vertical neuron).

$= a - r_v - r_l$

$r_2 = a + r_v + r_l$ = maximum radius to allow contact between vertical and lateral neuron (from the center of the vertical neuron).

In the case of the pigeon retina these dimensions have been estimated from Cajal's drawings of the bird's retina. They are as follows (see Fig. 11).

$a = 75\,\mu \quad r_v = 33\,\mu \quad r_l = 12$

From these dimensions it follows that

$r_1 = 30\,\mu \quad r_2 = 120\,\mu$ so that $R = 87.5\,\mu$

This yields a value for Θ of .538 radian (30.8°), so that

$$P_c = .171$$

A = area of the annulus between r_1 and $r_2 = 4.24 \times 10^4\ \mu^2$

A_r = total area of the pigeon retina $= 1.18 \times 10^8\ \mu^2$

$$P_A = \frac{A}{A_r} = \text{probability of being in the annulus between } r_1 \text{ and } r_2$$
$$= 2.78 \times 10^{-4} \tag{3}$$

The total area of the retina has been estimated from Chievitz's data[6] by assuming the pigeon retina as a hemisphere whose radius is 4.32 mm. Using the same hemispherical approximation concerning the human retina would yield a hemispherical radius of 12.2 mm and a retinal area of 935 mm^2. This calculated retinal area is very close to that measured by Van Buren[10] which is 943 mm^2 on the average using a midpapilla to midfovea distance of 3.614 mm as corresponding to 17° of arc on the human retina.

The probability that m associative amacrines lie in the annulus and may influence the ganglion, given that n are distributed uniformly across the retina, can be calculated using (4).

$$P_n(m) = \frac{n!}{(n-m)!\,m!} P_A^m (1 - P_A)^{n-m} \tag{4}$$

Here n is taken as 9250 which is approximately the mean of the data given by Cowan[7] as to the number of centrifugal fibers in the retina. It is well known from Cajal's work that centrifugal fibers go to associative amacrines. Recent work by Maturana[9] indicates many centrifugal fibers terminate on a single amacrine cell, giving rise to the assumption that the number of associative amacrines is near in quantity to the number of centrifugal fibers. Substituting $n = 9250$ in the above formula yields the following table:

m	$P_n(m)$	m	$P_n(m)$
0	.076	5	.072
1	.197	6	.031
2	.253	7	.011
3	.218	8	.003
4	.139		

Then

$P(k)$ = probability that exactly k associative amacrines terminate on a ganglion's dendritic spread.

$$P(k) = \sum_{m=0}^{\infty} \frac{m!}{(m-k)!\,k!} P_c^k (1 - P_c)^{m-k} P_n(m) \tag{5}$$

where $m \geqq k$

Substituting in the value of $P_c = .171$ and $P_n(m)$ from the above table gives:

$$P(0) = .645$$

$$P(1) = .283$$

$$P(2) = .062$$

$$P(>2) = .010$$

Thus it is seen that the model would predict that 28.3 per cent of the ganglions in the pigeon retina are directional detectors. This is in close agreement with "about 30 per cent" given in the literature.

LOUIS L. SUTRO

Instrumentation Laboratory

Massachusetts Institute of Technology

Cambridge, Massachusetts

*Proposed Electronics to Represent the Properties of a Frog's Eye**

THE PROBLEM

The problem considered here is how to represent in electronics the very large numbers of neurons and interconnections that have been found in a frog's retina. As shown schematically in Figure 1 this retina comprises approximately a million photoreceptor cells, a million bipolar cells, and a half-million ganglion cells. Table I shows the division of ganglion cells into the principal groups identified by Lettvin *et al.*[1,2,3]

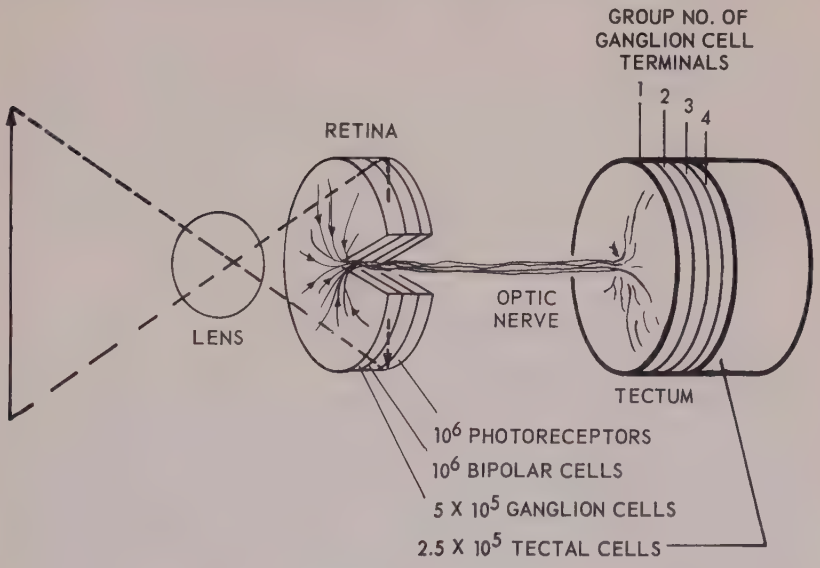


Fig. 1. Schematic of frog visual system.

* Prepared under the auspices of DSR Project 52-203, sponsored by the Space Systems Division, U. S. Air Force Systems Command through Contract AF 04(695)-641.

Table I: Division of Ganglion Cells in the Frog's Retina and Function of Each

Group No.	Estimated % of Total	Function
1	48.5	Edge detector
2	48.5	Detector of a bug (movement-gated dark convex centripetally-moving edge)
3	1.5	Event (moving or changing contrast) detector
4	1.5	Dimming detector

Four conceptual models of a frog's retina have been devised at our laboratory. The electronics described here is intended for the Model 3 completed in the spring of 1965.[4] Dr. Moreno-Diaz' model,[5] which he is describing at this symposium, has been completed since this electronics was devised.

APPROACHES

The electronics to be described are employed by designers of television cameras and computers who achieve very large numbers of operations in short spans of time by time-sharing equipment. One design employing both analog and digital computation we call the Alpha System. A second design employing only analog computation we call the Beta System.

PHOTORECEPTOR LAYER

For both the Alpha and Beta System a device that can represent the million photoreceptor cells and also be small and light is a vidicon tube. In a hypothesis advanced by Lettvin[6] a photoreceptor is conceived as forming a current in response to the onset of illumination and forming an impedance in response to continued illumination. Since impedance may be measured by a current, both onset of illumination and continued illumination may be represented by currents. In Figure 2 the photoreceptor layer is represented by two vidicons, one delivering a current proportional to the illumination received since the last frame, the other

a current proportional to a time integral of illumination received over several frames, weighted so that the most recent frame counts more than earlier ones. The arrangement of vidicons and mirrors is that employed in a color television camera.[7]

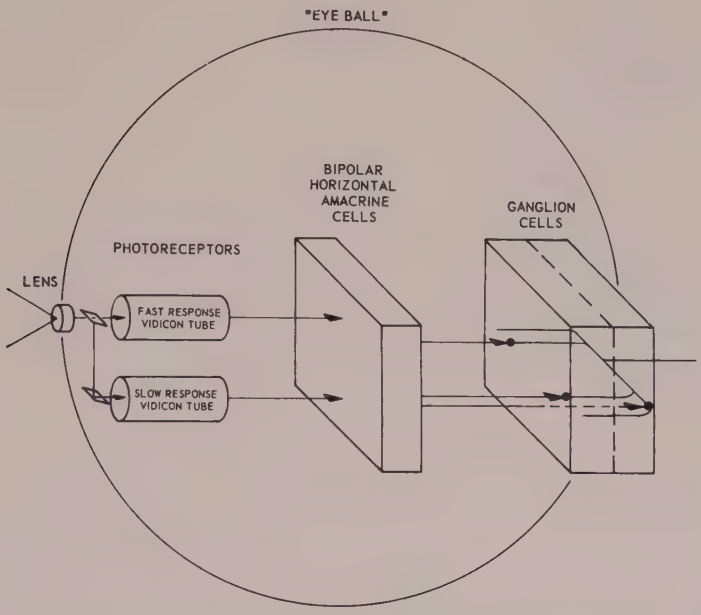


Fig. 2. Gross block diagram of the Alpha System.

In the Alpha System, the raster on the face of each vidicon is rectangular and comprises 510 lines with 510 positions per line. A thousand lines with 1000 positions per line is possible but less convenient to test. Odd-numbered lines are offset one-half position with respect to even-numbered lines. The number of lines and positions per line was chosen because it satisfies the conditions that such a number 1) provides a total of positions approximating the number of photoreceptors in a frog's retina, 2) is nearly a power of 2, for convenience in counting, and 3) is divisible by 3, a condition advantageous in the design of the models of the Group 3 and 4 ganglion cells. The Beta System is less precisely defined.

MODELS OF BIPOLAR-CELL AND GANGLION-CELL LAYERS IN THE ALPHA SYSTEM

Figure 2 is a gross block diagram of the Alpha System, showing the two vidicons and suggesting that the bipolar and ganglion-cell layers can be represented by sheet computers. The *E*-shaped structure of a Group 2 ganglion cell can be seen in the sheet computer representing the ganglion cells. To the left of the dotted line is the upper synaptic level; to the right, the lower synaptic level. All three layers of the model are shown within a sphere to suggest that these are the electronic equivalent of an eye.

Table II shows in columns 1 and 2 a condensation of data presented in Rfs. 3, 4 and 5. The data in columns 3 and 4 are derived from column 1 by the methods given in the notes in the table. Columns 5 and 6 present simplified specifications employed in the structure of the model. Thus Groups 1 and 2 ganglion cells are represented in the model as having a responsive retinal field (RRF) of 3°, and Groups 3 and 4 as having RRF's of 9°.

Table II: Characteristics of the RRF's of Ganglion cells

1	2	3	4	5	6
		Living Retina		Model Retina	
Group	RRF	No. of Photoreceptors in Dia. of RRF*	No. of Photoreceptors in Circular RRF†	Assumed RRF	No. of Photoreceptors in Dia. of RRF
1	1° to 3°	12 to 36	114 to 1020	3°	36
2	3° to 5°	36 to 60	1020 to 2840	3°	36
3	7° to 120	84 to 144	5550 to 16,400	9°	108
4	up to 15°	up to 180	up to 25,000	9°	108

Notes

* Photoreceptors average 4 μ center to center.[8]
48 μ on the retina subtends 1° of visual angle (Ref. 2, p. 140). Therefore, one photoreceptor subtends 1/12° of visual angle.

† This is $\pi d^2/4$ where d is the number in column 3.

By using time-shared electronics, only one electronic model of each bipolar and ganglion cell is needed. In Figure 3 each model bipolar and ganglion cell is represented by a cylinder on its side, behind an array of shift registers.

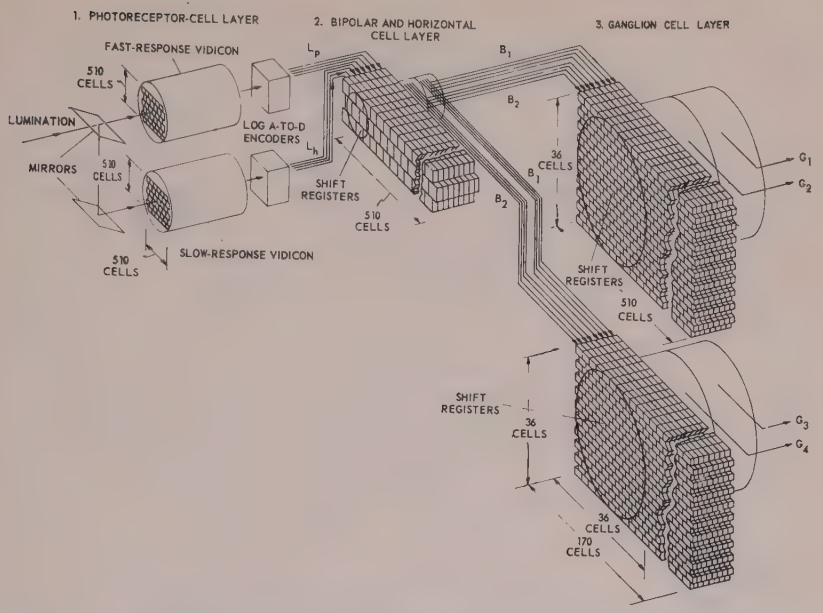


Fig. 3. Detailed block diagram of the Alpha System.

OPERATION OF THE ALPHA SYSTEM

The output of each camera tube in Figure 3 is encoded as one four-bit digital word for each of the quarter-million positions on the face of the tube. Each digital word is fed into six parallel shift registers at the top of the bank of shift registers called "bipolar and horizontal cell layer." These data inputs are labelled L_p, which is the digitalized logarithm of present illumination, and L_h, which is the digitalized logarithm of the history of illumination.

Data are advanced from left to right in the shift registers at the same rate that the electron beams in the two vidicons advance. As data reach the right end of the top row, they are transferred to the left end of the next lower row of shift registers. When the electron beams of the two camera tubes have swept across three lines of the raster, the shift registers in the bank are full. From then on, for each new position of the electron beam, computations take place in the cylinders behind the shift registers and one digital word will be discarded from the right end

of the third row of shift registers. For economy, those parts of the shift registers to which no computing elements are attached can be delay lines.

The input to the computation in the cylinder is from the seven words stored in the part of the bank of shift registers within the circle in Figure 3. As shown in Figure 4, the words within the circle consist of a

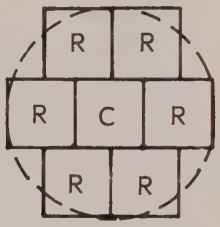


Fig. 4. Words enclosed in the circle on the bipolar

central word (C) surrounded by six ring words (R). The computation carried out in the front half of the cylinder is a comparison of the central word with each of the ring words determining if there is contrast (an edge) between them. The computation carried out in the back half is a comparison of L_p with L_h over the area determining if there has been a brightening or dimming.

The result of the computation in each of the two model bipolar cells is a six-bit word (B_1 or B_2). This word is fed to two banks of shift registers, one representing the first two groups of ganglion cells (G_1 and G_2), the other the second two groups (G_3 and G_4).

Following the assumptions made in Table II, the RRF of the Groups 1 and 2 ganglion cells are represented by a circular area 36 cells in diameter. The hardware for this can be a 36-row shift register backed up by analog computing elements to perform the operations of each kind of ganglion cell. The E structure, shown at the right of Figure 2, can be built in the cylinder labelled G_2.

Models of the Groups 3 and 4 ganglion cells are also 36 cells in diameter. The RRF of these cells is three times as great, however, so the required scaling is accomplished by having each row of the shift register represent every third line of the raster, and each position of the shift register every third position of the raster. For a 510 by 510 raster in the camera tube, the number of positions in each shift register will be 170.

In summary, time-shared electronics have been applied to develop a parallel-sequential-parallel-sequential mode of operation with respect to time. Information presented in parallel (i.e., simultaneously) to the vidicons is encoded in sequence. It is then passed to shift registers representing the input to the bipolar layer where, for each position of the electron beams, information is taken out in parallel from seven elements to models of the bipolar cells. From these, information proceeds sequentially to a second bank of shift registers. From there it passes in parallel to models of the ganglion cells. Information leaves these elements sequentially.

The general Alpha scheme was presented at the 1963 Bionics Symposium, using a magnetic tape as buffer storage.[9] Larry Baxter devised the shift registers which make it possible to build an electronic model that will respond as fast as a frog's eye.

THE BETA SYSTEM

The electronic design presented in the previous sections was directed toward modelling as accurately as possible the structure and functioning of an animal retina. Now we turn to a design by Larry Baxter which models only the essential features of an animal retina, to assure being able to carry the model into space.

Essential features of the photoreceptor and bipolar layers of an animal retina appear to be:

(a) Ability to detect a difference between present and past illumination from the same point in space (temporal difference).

(b) Ability to detect difference in illumination from two adjoining points in space (spatial difference).

These features are incorporated in the system shown in Figure 5.

Light entering the two-vidicon camera is split equally between a vidicon with characteristics normally used in television (low lag) and one capable of storing the effects of many scans (high lag). Comparing the output of the two tubes in one summer (Σ) yields an indication of dimming, in another an indication of brightening. The output of the lowlag vidicon is fed also to a differentiator which detects changes in the grightness of the scene. A sharp change is interpreted as an edge (Group 1 banglion cell). A combination of edge and dimming information is fed to the

bug detecting logic (Group 2). Dimming alone represents the Group 3. The combination of brightening and dimming in an OR relation represents the Group 4.

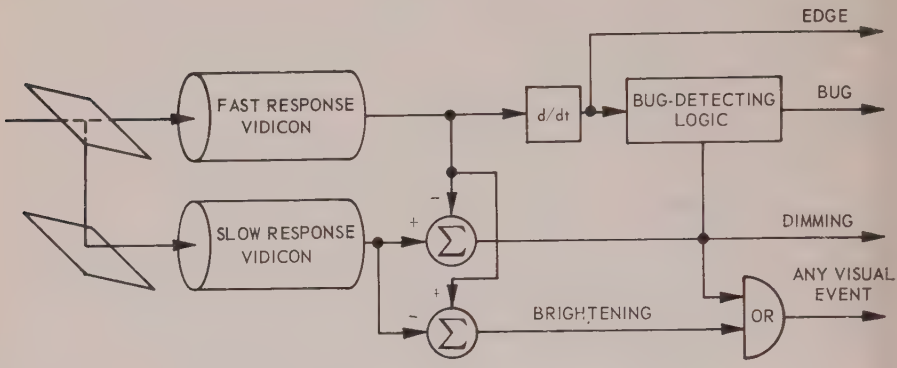


Fig. 5. Beta System.

CONCLUSION

Designing the Alpha and Beta Systems has served two useful functions. The first has been to assist in the creation of the conceptual models of a frog's retina of which the latest and by far the best is that by Dr. Moreno-Diaz.[5] The second function has been to suggest useful sensors for spacecraft. One such sensor is that being devised for the first soft landing on Mars. At present it does not appear possible to send entire pictures back from the landing craft. Instead it may be useful to send back the kind of information that "a frog's eye tells a frog's brain."[1]

REFERENCES

1. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. What the Frog's Eye Tells the Frog's Brain, *Proceedings of the IRE*, **47,** 1940–1959, (1959).
2. Maturana, H. R., Lettvin, J. Y., McCulloch, W. S., and Pitts, W. H. Anatomy and Physiology of Vision in the Frog (Rana Pipiens), *Journal of General Physiology*, **43,** 129–175, July 1960.
3. Lettvin, J. Y., Maturana, H. R., Pitts, W. H., and McCulloch, W. S. *Two Remarks on the Visual System of the Frog*, in *Sensory Communication*, W. Rosenblith, ed., pp. 757–776, John Wiley, New York 1961.

4. Sutro, L. L., editor. **1964** *to September* **1965** *Advanced Sensor and Control System Studies*, R-519, Instrumentation Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, January 1966.
5. Moreno-Diaz, R., An Analytical Model of the "Bug Detector" Ganglion Cell in the Frog's Retina, "Bionics Symposium" 1966. This volume pp. 481–491.
6. Lettvin, J. Y. *Research Laboratory of Electronics Quarterly Progress Report No.* **70,** Massachusetts Institute of Technology, Cambridge, Massachusetts, 15 July 1963, pp. 327–337.
7. Fink, D. G., editor, "Television Engineering Handbook," McGraw-Hill, New York, 1957.
8. Maturana, H. R., personal communication to L. Sutro, April 8, 1963.
9. Sutro, L. L., Plan for the Simulation of the Photoreceptor and Bipolar Layers of a Frog's Retina, "1963 Bionics Symposium" Papers Preprints, Aeronautical Systems Division, Air Force Systems Command, United States Air Force, Wright-Patterson Air Force Base, Ohio.

W. C. LIN
Case Institute of Technology†
Cleveland Ohio
K. S. FU
School of Electrical Engineering, Purdue University
Lafayette, Indiana

*An Adaptive Pattern Recognition System Using Neuron-Like Elements**

ABSTRACT

An adaptive pattern recognition system is proposed in this paper. The system contains two adaptive loops, one for feature transformation and the other for categorization. An algorithm has been developed for the feature transformation, and its convergence through a learning process has been proved. Concepts of correlation, hyperplane and hypersphere classification techniques have been employed for the categorization process. A transistorized neuronlike threshold device with bi-polar outputs has been developed and the device used as the basic element to realize the proposed system. A machine based on the proposed system which consists of three essential layers—namely, transformation, correlation and decision layers—has been built in the laboratory. The adaptation is performed by adjusting thresholds of the neuron-like elements only. Experimental results for geometric figures, hand-printed and handwritten characters are presented.

INTRODUCTION

A complete pattern recognition system contains, in general, two essential parts—namely, feature extraction and categorization. In addition, it is felt that a desirable system has the following characteristics:

(1) The system should be trainable for recognizing more general or complex patterns.

† Formerly with School of Electrical Engineering, Purdue University, Lafayette' Indiana.

* This work was partially supported by the National Science Foundation Grant GP-2183.

(2) That it can be trained to achieve good performance with relatively small samples per class.

(3) That it can mostly carry out the feature extraction process "on-line" with little outside supervision.

(4) That it can recognize patterns with a reasonable amount of noise, distortion, displacement and orientation variation.

(5) That it should be able to improve its performance even during the recognition phase.

(6) That a machine can really be built with a reasonable cost according to the proposed system.

In this paper, an adaptive pattern recognition system having the above characteristics is proposed; a prototype machine which has been designed and built according to the proposed system is described and the experimental results are presented.

THE PROPOSED SYSTEM

A block diagram of the proposed system is shown in Figure 1. The receptor portion consists of transducer and vector space transformation. The categorizer portion consists of class references storage, correlation and decision. The transducer is actually a converter that converts physical patterns into electrical signals. For example, in a speech recognition system, it may be a microphone; in a visual pattern recognition system, it may be an array of photo-sensitive devices. If there are k different classes of patterns, and each class has h members, then the ith member of the jth class pattern can be represented by a vector V_{ji} in the signal space of n-dimension, providing that the signal has n-components. If each component is to be quantized into two levels, say, "0" and "1," then V_{ji} is an n-dimensional binary vector in the signal space, represented by

$$V_{ji} = (v_1, v_2, \ldots, v_n)_{ji}$$

where

$(v_1, v_2, \ldots, v_n)$ = a row vector of n random variables which may assume the value of "1" or "0".

As shown in Figure 1, V_{ji} is the output of the transducer and the input of the transformation layer. The output of the transformation layer

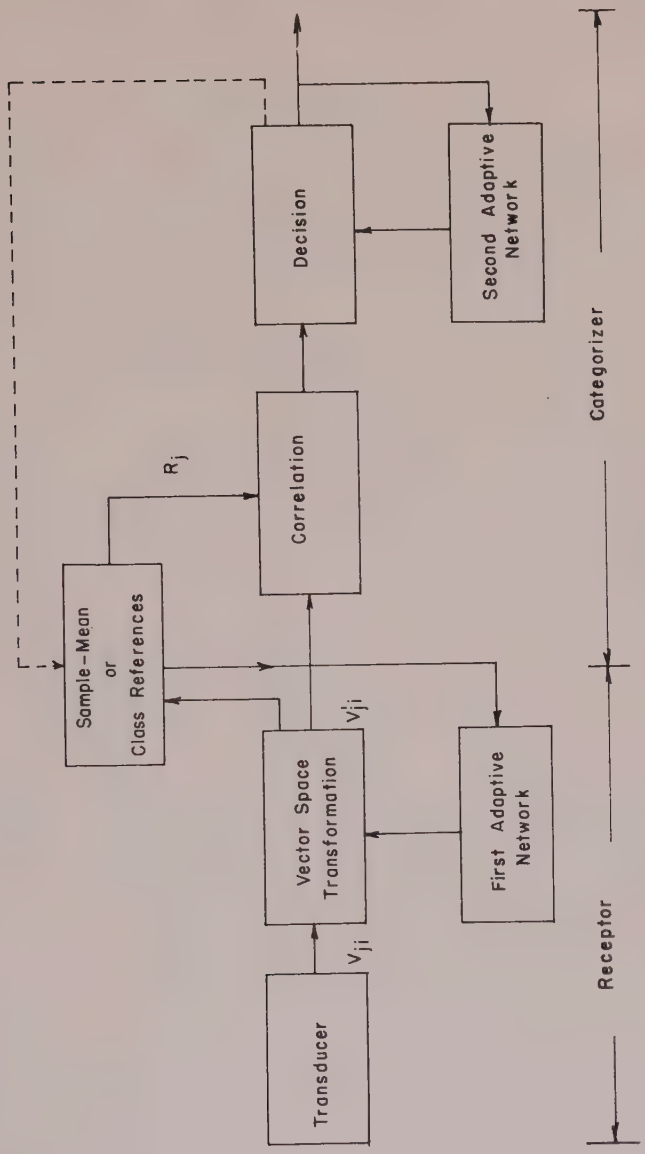


Fig. 1. System Block Diagram

is again a vector, V'_{ji}, corresponding to V_{ji}, but is now in a new m-dimensional space which may be called the "Feature Space." Then the ith member of the jth class pattern in this new space is

$$V'_{ji} = (v'_1, v'_2, \ldots, v'_m)_{ji}$$

and the value of m is desired to be much less than n. The purpose of this transformation is two-fold: First, the members of the same class are clustered together, and yet the distance among different classes remains proportionally in the new space as in the old. Second, the dimensionality of the feature space is reduced considerably for the categorization process. An algorithm for the transformation process has been developed, and its convergence has been proved.[1] The actual implementation of the transformation layer will be described in the next section. The basic principle of the transformation is to have an adaptive network (the first adaptive network in Fig. 1) being adjusted such that the probability of the occurrence of each binary bit in the feature space will be approximately equally probable.

A set of reference vectors (or sample means)

$$R = [R_1, R_2, \ldots, R_j, \ldots, R_k]$$

with

$$R_j = \frac{1}{h} \sum_{i=1}^{h} V'_{ji}$$

has been stored. The correlation layer is a network that determines a set of correlation coefficients $\{E_j\}_{ji}$ between a sample V_{ji} and each of the references, or

$$\{E\}_{ji} = [E_1, E_2, \ldots, E_j, \ldots, E_k]_{ji}$$

where

$$E_j = V'_{ji} \cdot R_j, \quad j = 1, 2, \ldots, k.$$

Since vectors to be considered here are all binary vectors, the correlation coefficients are simply the scalar products of two vectors.

The decision layer decides in which class the unknown pattern should be classified. The following decision rules are used. In order to avoid

confusion between unknown patterns and training samples, it is necessary to define Xs as the unknown and Vs as the training samples. Then

$$X \in j\text{th class if and only if}$$

$$X' \cdot R_j > X' \cdot R_e, \quad e \neq j \tag{1}$$

and

$$|X' - R_j| \leqq d_j \tag{2}$$

where X' = unknown pattern vector in the Feature Space

$e = 1, 2, \ldots, k$

d_j = a positive scalar

A combination of a hyperplane and hypersphere is used as the decision boundary.

The operation of the system is divided into two phases: (1) learning phase and (2) recognition phase. During the learning phase, several training samples of each pattern class are applied to the transducer. By "training samples," one means patterns whose memberships are already known. The machine is trained through the first adaptive network. The mean of the samples of each class is determined accordingly and stored as a reference. Then the same set of samples is again applied to the input and correlated with the samples-mean just obtained. The decision layer is then trained through the second adaptive network (Fig. 1) such that the machine can classify all of the samples correctly. The recognition phase begins by applying unclassified patterns to the input of the transducer; its correlation coefficients are determined, and the decision is then made; its corresponding "class mean" is modified in order to update the reference. This operation has the advantage that the machine performance is also improved, on the average, even during the recognition phase. Although the machine may occasionally make wrong decisions, and the reference will be misadjusted, the machine is expected to make reasonably correct decisions most of the time. It is believed that the overall performance will be improved, and this will be verified by the experimental results.

A simple example is given in the following to clarify the above description.

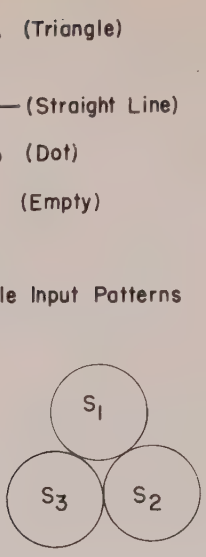


Fig. 2. Classification of Simple Visual Patterns.

As shown in Figure 2, assume that classes of patterns are available, i.e., triangle (V_1), straight line (V_2), dot (V_3) and empty (V_4), and the transducer is made of three photocells. By using the vector notation described, one may have:

$$\text{(triangle)} = V_{11} = 1\,s_1 + 1\,s_2 + 1\,s_3$$

$$\text{(straight line)} = V_{21} = 0\,s_1 + 1\,s_2 + 1\,s_3$$

$$\text{(dot)} = V_{31} = 1\,s_1 + 0\,s_2 + 0\,s_3$$

$$\text{(empty)} = V_{41} = 0\,s_1 + 0\,s_2 + 0\,s_3$$

where s_1, s_2, s_3 are unit vectors in the signal space. Now, a transformation vector, $T = (t_1, t_2, t_3)$ may be introduced. If the vector T is designed such that

$$t_1 = t_2 = t_3 = \text{One unit},$$

then $$T = t_1 s_1 + t_2 s_2 + t_3 s_3 = 1\,s_1 + 1\,s_2 + 1\,s_3.$$

The transformation process is described as follows:
(i) the scalar product is performed, i.e.,

$$V_{11} \cdot T = 1 + 1 + 1 = 3$$

$$V_{21} \cdot T = 0 + 1 + 1 = 2$$

$$V_{31} \cdot T = 1 + 0 + 0 = 1$$

$$V_{41} \cdot T = 0 + 0 + 0 = 0.$$

Notice that the products are a set of scalars. Let c be the scalar.

(ii) by referring to its product, the vector is assigned to a feature space with lower dimension, in such a way that the occurrence of all points in the feature space is to be equally probable. In this example, the dimension of the signal space is three, and that of the new space (feature space) is to be two with f_{11}, f_{12} as the unit vectors. To carry this out, a set of thresholds $\{\theta\}$ is determined. The vectors are assigned according to which range of the thresholds their scalar products have fallen into. Let the thresholds be $\theta_1 = 0.5$, $\theta_2 = 1.5$, $\theta_3 = 2.5$, then vectors in the feature space can be assigned as follows:

$$\text{(triangle)} \quad V'_{11} = 1f_{11} + 1f_{12} \text{ since } \theta_3 < V_{11} \cdot T$$

$$\text{(straight line)} \quad V'_{21} = 1f_{11} + 0f_{12} \text{ since } \theta_2 < V_{21} \cdot T < \theta_3$$

$$\text{(dot)} \quad V'_{31} = 0f_{11} + 1f_{12} \text{ since } \theta_1 < V_{31} \cdot T < \theta_2$$

$$\text{(empty)} \quad V'_{41} = 0f_{11} + 0f_{12} \text{ since } 0 < V_{41} \cdot T < \theta_1.$$

Now, assume that two unknowns are applied, and that they happen to be a "straight line," $X_1(1,0,1)$ and a "dot," $X_2(0,1,0)$, then

$$X_1 \cdot T = 2 \text{ which falls in } \theta_2 \sim \theta_3$$

$$X_2 \cdot T = 1 \text{ which falls in } \theta_1 \sim \theta_2$$

and

$$X'_1 = 1f_{11} + 0f_{12}$$

$$X'_2 = 0f_{11} + 1f_{12}.$$

Decisions are:

$$X'_1 \in V_2 \text{ (straight line)},$$

$$X'_2 \in V_3 \text{ (dot)},$$

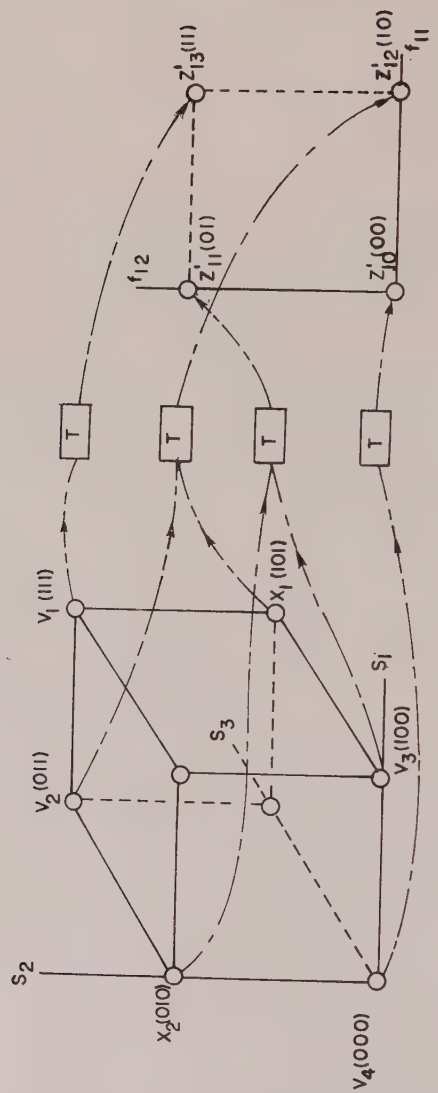


Fig. 3. An Example for Illustrating Space Transformation Process.

since both $X_1 \cdot V_2$ and $X_2 \cdot V_3$ yield maximum correlation coefficients. The geometrical presentation of this transformation process is shown in Figure 3. Notice that the Euclidean distance between X_1 and V_2; X_2 and V_3 in the signal space is

$$D_{12} = [(0 - 1)^2 + (1 - 0)^2 + (1 - 1)^2]^{1/2} = \sqrt{2}$$

and

$$D_{23} = [(1 - 0)^2 + (0 - 1)^2 + (0 - 0)^2]^{1/2} = \sqrt{2}$$

respectively, while their distance in the feature space is zero. As for the inter-class distance, it still remains in proportion. The clustering effect of intra-class sample points due to the transformation is now clearly shown.

The use of this simple example is not intended to prove the feasibility of the proposed system; instead, the example is merely used to clarify the operational process of the system.

THE ADAPTIVE RECOGNITION MACHINE

In this section, the prototype machine implementing the proposed system is described. First, the basic neuron-like element which has been extensively used to realize the system is described. The descriptions of the three essential layers—i.e., transformation, correlation, and decision—and the machine operation are followed.

1. Neuron-like Element (Bi-polar Threshold Logic Element—BTLE)

An electronic neuron-like element using two transistors and six resistors has been developed and proved to be useful in this adaptive pattern recognition system. This element may be considered as a realization of the "Formal Neuron" defined by McCulloch and Pitts.[2] In this research, it is not the intention to design an electronic model neuron which has all the properties of a biological neuron; rather, a model having some properties which appear to be very useful for the system realization has been developed. The simplicity of the model is highly concerned, and it has the following properties: (a) Spatial summation, (b) Adjustable threshold, (c) Excitatory and/or inhibitory outputs. As the conventional threshold element, this model has a summing network at the input. When the sum of the inputs exceeds a threshold, the device changes its state

at the outputs. However, this model has two outputs, namely, inhibitory and excitatory. Thus, it may be called "Bipolar Threshold Logic Element." The switching variables may be expressed in pairs, i.e.,

x_i^+ as inhibitory, and

x_i^- as excitatory.

They assume two logical states:

$$x_i^+ = \begin{cases} +1 \\ 0 \end{cases}, \quad x_i^- = \begin{cases} -1 \\ 0 \end{cases}$$

The element can be used to realize the following:

A switching function

$$F(x_1^{\pm}, x_2^{\pm}, \ldots, x_n^{\pm}) = 0$$

iff

$$\sum_{i=1}^{n} x_i^- w_i^- + \sum_{i=1}^{n} x_i^+ w_i^+ < \theta$$

and

$$F(x_1^{\pm}, x_2^{\pm}, \ldots, x_n^{\pm}) = \pm 1$$

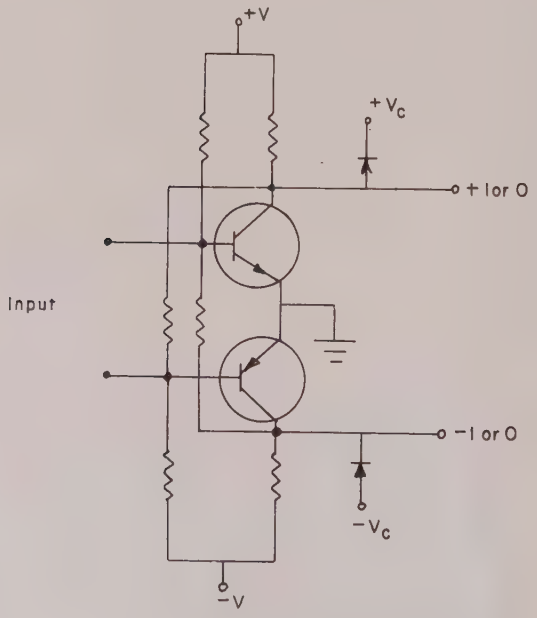


Fig. 4. Bipolar Threshold Logic Element.

iff

$$\sum_{i=1}^{n} x_i^- w_i^- + \sum_{i=1}^{n} x_i^+ w_i^+ \geq \theta$$

where w_i denotes the weight of the corresponding switching variable x_i. θ denotes the threshold.

The circuit diagram of a Bipolar Threshold Logic Element is shown in Figure 4.

2. Implementation of the Three Essential Layers Using the Basic Elements

Transformation Layer

The transducer is simply a 12 × 12 photo cell grid followed by a set of quantizers which yield a binary vector in the signal space. The transformation layer is employed to transform the signal vector from a 144-dimensional space into a 24-dimensional feature space. First, the 144 cells are divided into 12 zones and each zone contains 12 cells. Each zone is then transformed independently; hence, it is responsible for generating two binary bits in the feature space. It will not lose generality if the transformation process in only one zone is discussed. The transformation process may be described as follows: Two bipolar threshold logic elements are used to form the transformation pair. Let $V(v_1, v_2, \ldots, v_{12})$ denote a zone vector, and $T(t_1, t_2, \ldots t_{12})$,which is the weighting vector of the threshold element, denote the transforming vector, then

$$V \cdot T = \sum_{i=1}^{12} v_i t_i = C$$

where C is a scalar.

If $t_1 = t_2 = \ldots = t_{12} = 1$ unit, then the range of C is from 0 to 12. When C is equal to or greater than the threshold, the element yields ± 1 at its outputs. As shown in Figure 5, the inhibitory output $(+1)$ of the second element is fed back to the input of the first element. The outputs yield a two-bits binary vector V'. Thus, the process transforms a 12-bit vector (V) into a 2-bit vector (V'). Notice that there are three adjustable thresholds, θ_1, θ_2, and θ_3, which yields the adaptive network.

Let the thresholds of the BTLE pair satisfy the following inequalities,

$$\theta_1 < \theta_2 < \theta_3$$

It can be shown that

$$\begin{aligned}
&\text{if} \quad 0 < C \leq \theta_1, \quad \text{then} \quad V' = [0, 0] \\
&\text{if} \quad \theta_1 < C \leq \theta_2, \quad \text{then} \quad V' = [-1, 0] \\
&\text{if} \quad \theta_2 < C \leq \theta_3, \quad \text{then} \quad V' = [0, -1] \\
&\text{if} \quad \theta_3 < C, \quad \text{then} \quad V' = [-1, -1]
\end{aligned}$$

An adaptive rule is now followed. It is desirable that the threshold set $\{\theta_1, \theta_2, \theta_3\}$, be adjusted such that the occurrence of the four possible points of V' be equally probable.

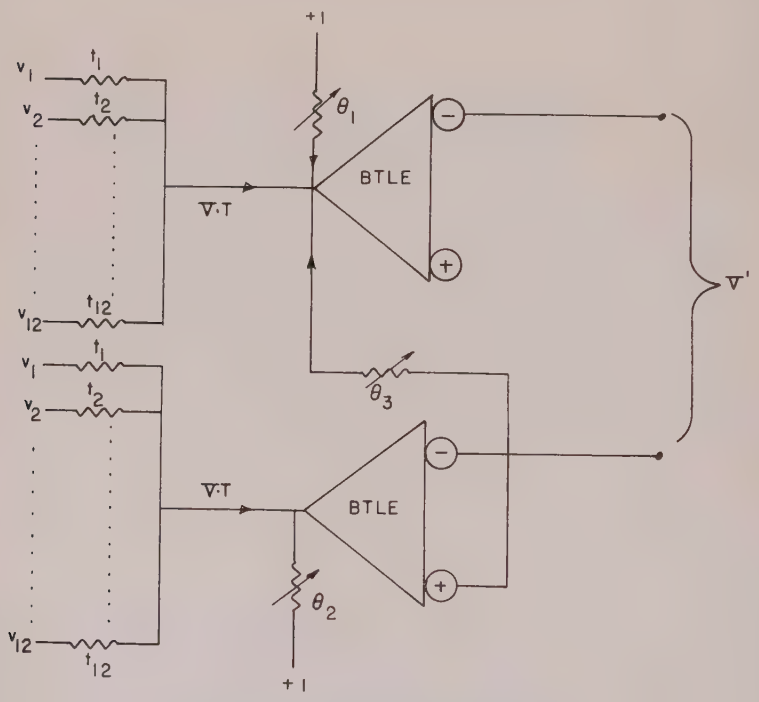


Fig. 5. Transformation BTLE PAIR.
(BTLE: Bipolar Threshold Logic Element).

Consider K classes of patterns. One sample is taken from each class, then there are k vectors of V, which in turn yield k values of C. A frequency function of C, $P(c)$, can be determined. If the thresholds θ_1, θ_2,

and θ_3 are adjusted such that

$$\int_0^{\theta_1} P(c)\,dc \simeq \int_{\theta_1}^{\theta_2} P(c)\,dc \simeq \int_{\theta_2}^{\theta_3} P(c)\,dc \simeq \int_{\theta_3}^{c_{\max}} P(c)\,dc$$

then

$$P(0, 0) \simeq P(-1, 0) \simeq P(0, -1) \simeq P(-1, -1),$$

where

$$P(0, 0),\ P(-1, 0),\ P(0, -1) \quad \text{and} \quad P(-1, -1)$$

denote the frequency of occurrence of point (0, 0), ... etc. in the feature space.

An adaptive algorithm to achieve this is described in a later section.

Correlation Layer

This layer contains a memory unit for storing class-mean or class references and the networks determining the correlation coefficient. First, a sample mean or a most representative sample of each class is to be determined through the given samples during the training phase, and then is stored as a reference. BTLE's are used in this layer as memory units and used to supply information to the correlation networks. During the recognition phase, an unknown pattern vector X is correlated with all the references, and the correlation coefficients are determined. The coefficients indicate the degree of similarity between the unknown and the references. The correlation network is a realization of the following switching function

$$f_i(X, R) = x_i' r_i' + x_i r_i$$

where x_i is the ith component of the unknown vector X,

r_i is the ith component of the reference vector R,

$f_i(X, R)$ is the ith component of the correlation coefficient between X and R.

$$X \cdot R = \sum_{i=1}^{24} f_i.$$

Notice that the switching function is an exclusive "OR," or a non-linearly separable function. However, because of the availabilities of

both inhibitory and excitatory outputs, the correlation network is quite simple. It is shown in Figure 6.

Decison Layer

The outputs of the correlation layers are fed to the decision layer. The decision layer is also a layer of BTLE's with adjustable thresholds which, in this case, are equivalent to the radii of a set of hypersphere decision boundaries. It can be shown that the radius can be expressed in terms of the correlation coefficients.[3] If an unknown falls inside the

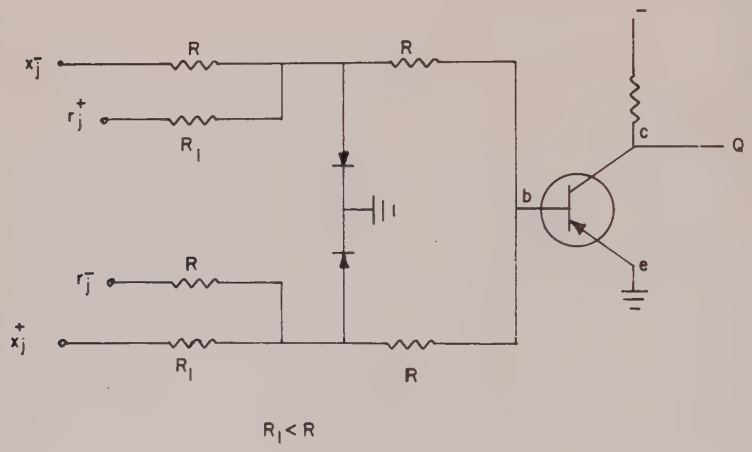


Fig. 6. Correlation Network.

sphere, it is classified as belonging to that class. Hyperplane boundaries are used to determine the unknown which falls into the overlapping region of hyperspheres. The decision rules can now be formulated in terms of the correlation coefficient as follows:

$$X \in R_j$$

if and only if

(i) $4(R_j \cdot R_j - X \cdot R_j) \leq g^2$

and

(ii) $X \cdot R_j > X \cdot R_e \ (j \neq e)$

where

$j = 1, 2, \ldots, k$

$e = 1, 2, \ldots, k$

R_j = the jth class reference

X = the unknown pattern vector

g = the radius of the hypersphere of j^{th} class.

3. Machine Operation

The operation of the machine can be subdivided into two phases, namely, (a) learning or training phase and (b) recognition phase.

(a) Learning phase.

As shown in Figure 1, there are two adaptive loops, one for determining the threshold settings such that the occurrence of the binary bits in the feature space would be equally probable, the other for determining the radii of the hyperspheres such that the error of misclassifying the samples be kept to a minimum. For the second loop, since there is only one variable per class, the algorithm is quite simple. However, the algorithm for the first loop is not obvious, and it is described in the following. The proof of its convergence can be found in the references.[1,3] Since the algorithm is identical in each zone, that in only one zone is described:

1. Construct the frequency diagram $P(c)$ through the available samples.
2. Start with an arbitrary set of threshold θ_j where $j = 1, 2, 3$.
3. From the frequency diagram $P(c)$, select a set of the most representative samples as the initial references, one for each class.
4. Construct a new frequency diagram from the selected samples, and label it with the initial (arbitrary) set of thresholds.
5. Adjust the jth threshold θ_j to a new value θ'_j such that

$$\theta'_j = \theta_j + \Phi(w_j)$$

with the contraint $\theta_j < \theta_{j+1}$

and

$$\Phi(w_j) = \frac{w_j^-}{w_j^+} \Big/ f(c = \theta_j)$$

with

$$f(c = \theta_j) = 1 \quad \text{when} \quad 1 > f(c = \theta_j) \geq 0$$

$$f(c = \theta_j) = f(c = \theta_j) \quad \text{when} \quad f(c = \theta_j) \geq 1.$$

where

$$w_j^- = \int_{\theta_j}^{\theta_{j+1}} P(c)\,\delta(c - c_i)\,dc - \int_{\theta_{j-1}}^{\theta_j} P(c)\,\delta(c - c_i)\,dc$$

$$w_j^+ = \int_{\theta_j}^{\theta_{j+1}} P(c)\,\delta(c - c_i)\,dc + \int_{\theta_{j-1}}^{\theta_j} P(c)\,\delta(c - c_i)\,dc$$

6. The process is terminated as

$$\tau = \sum_{j=1}^{3} |w_j^-| = \text{minimum}$$

A simple example can be used to clarify the algorithm.

Assume that there are eight classes of patterns, and each class has four samples which can be denoted as follows:

1st class $V_1 = V_{11}, V_{12}, V_{13}, V_{14} =$ 

2nd class $V_2 = V_{21}, V_{22}, V_{23}, V_{24} =$ 

3rd class $V_3 = V_{31}, V_{32}, V_{33}, V_{34} =$ 

4th class $V_4 = V_{41}, V_{42}, V_{43}, V_{44} =$ 

5th class $V_5 = V_{51}, V_{52}, V_{53}, V_{54} =$ 

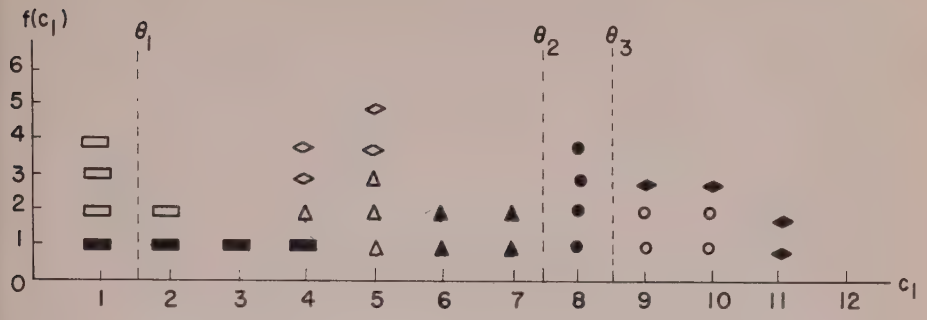
6th class $V_6 = V_{61}, V_{62}, V_{63}, V_{64} =$ 

7th class $V_7 = V_{71}, V_{72}, V_{73}, V_{74} =$ 

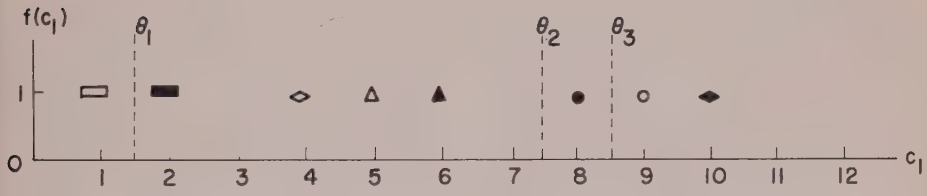
8th class $V_8 = V_{81}, V_{82}, V_{83}, V_{84} =$ 

All the samples are "shown" to the machine.

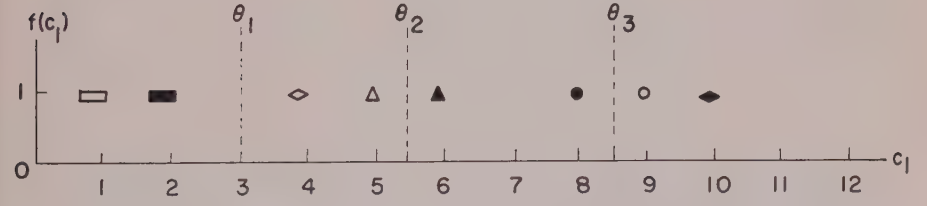
Now let us consider the process in the first zone only. Assume that the frequency diagram of the 32 samples is as shown in Figure 7(a),



(a) Frequency Diagram of All Samples with Arbitrary Setting of Thresholds.



(b) Frequency Diagram of Representative Samples with Arbitrary Setting of Thresholds.



(c) Frequency Diagram of Representative Samples with Adapted Setting of Thresholds.

Fig. 7. Example for Adaptive Rule of First Adaptive Loop in First Feature Subspace.

and that the most representative samples are shown in Figure 7(b) with an arbitrary setting of the thresholds,

$$\theta_1 = 1.5$$

$$\theta_2 = 7.5$$

$$\theta_3 = 8.5$$

Accordingly

5th class is assigned to the point (0, 0);

1st, 2nd, 3rd and 7th, to point (0,1);

8th, to point (1,0) and

4th and 6th to point (1, 1), in the feature space.

With this setting, one obtains

$$\omega_1^- = 4 - 1 = 3 \qquad \omega_1^+ = 4 + 1 = 5$$

$$\omega_2^- = 1 - 4 = -3 \qquad \omega_2^+ = 1 + 4 = 5$$

$$\omega_3^- = 2 - 1 = 1 \qquad \omega_3^+ = 2 + 1 = 3$$

$$\tau = 7$$

The adaptive process begins according to the algorithm described previously. The detailed results are shown in Table I and the final frequency diagram is shown in Figure 7(c). Notice that the probability of each point in the feature space is now equal to each other, or

$$P(0, 0) = P(0, 1) = P(1, 0) = P(1,1) = 1/4$$

The adaptive process for this subspace is completed, and the adaptive processes of the other subspaces can be carried out in the same manner. The class references are determined. A question may arise in Figure 7(c). It shows that classes 1 and 5 are assigned to the same point (0, 0) in the feature space. It appears that the machine tends to classify classes 1 and 5 as the same class. This is actually not the case, since it merely indicates that in that subspace, they look the same. The probability of

Table I

Training Cycle	θ_1	ω_1^-	ω_1^+	$\Phi(\omega_1)$	θ_1'	θ_2	ω_2^-	ω_2^+	$\Phi(\omega_2)$	θ_2'	θ_3	ω_3^-	ω_3^+	$\Phi(\omega_3)$	θ_3'	τ
1	1.5	3	5	+0.6	2.1	7.5	—3	5	—0.6	6.9	8.5	1	3	+0.3	8.8	7
2	2.1	1	5	+0.2	2.3	6.9	—2	4	—0.5	6.4	8.8	1	3	+0.3	0.1	4
2	2.3	1	5	+0.2	2.5	6.4	—1	5	—0.2	6.2	9.1	—1	3	0.3	8.8	3
4	2.5	1	5	+0.2	2.7	6.2	—1	5	—0.2	6.0	8.8	0	4	0	8.8	2
5	2.7	0	4	0	2.7	6	—1	3	—0.3	5.7	8.8	0	4	0	8.8	1
6	2.7	0	4	0	2.7	5.7	0	4	0	5.7	8.8	0	4	0	8.8	0

classes 1 and 5 being assigned to the same point in all 12 subspaces is low unless they really belong to the same class.

Recognition Phase

After the learning phase operation is completed, the machine is ready to recognize the unknown pattern. For this prototype machine, there are three classes of references; thus the machine can decide whether the unknown pattern belongs to one or none of the three classes. If it is known that the unknown is always one of the three, then the radii of the hyperspheres can be set easily since no rejection is necessary. The performance of the machine may be improved further if the class references are also modified or updated during the recognition phase.

EXPERIMENTAL RESULTS

Experimental results are presented in two different forms, one for demonstrating the relationship between "Uniformity Index" and the mean correlation coefficient among classes, the other for demonstrating the recognition of patterns through the proposed algorithm. The uniformity index is defined as the sum of the probabilities of all points deviated from the uniform probability in the feature space.

Five sets of patterns different in nature have been used for testing the machine. They are: (1) Chinese characters, (2) traffic signs, (3) Arabic numerals (4) handwritten English characters and (5) hand-printed English characters. As shown in Figure 8, with uniformity index equal to 58, the mean correlation coefficient among the four Chinese characters is about 32 (the maximum is 36), which means that the machine cannot tell any difference among the four classes. By adjusting the thresholds of each zone, the similarity among classes does indeed decrease. Overall, 374 samples were used for the experiment; the recognition results are shown in Table II. Notice that there is a difference in recognition rate if the class reference is selected differently. For instance, when a sample of handprinted "A" was selected arbitrarily as the class-reference from the 100 samples, only 81 out of 100 were recognized correctly. When the most representative one out of 12 samples was selected as the class-reference, the recognition results were improved to 95 out of 100 samples.

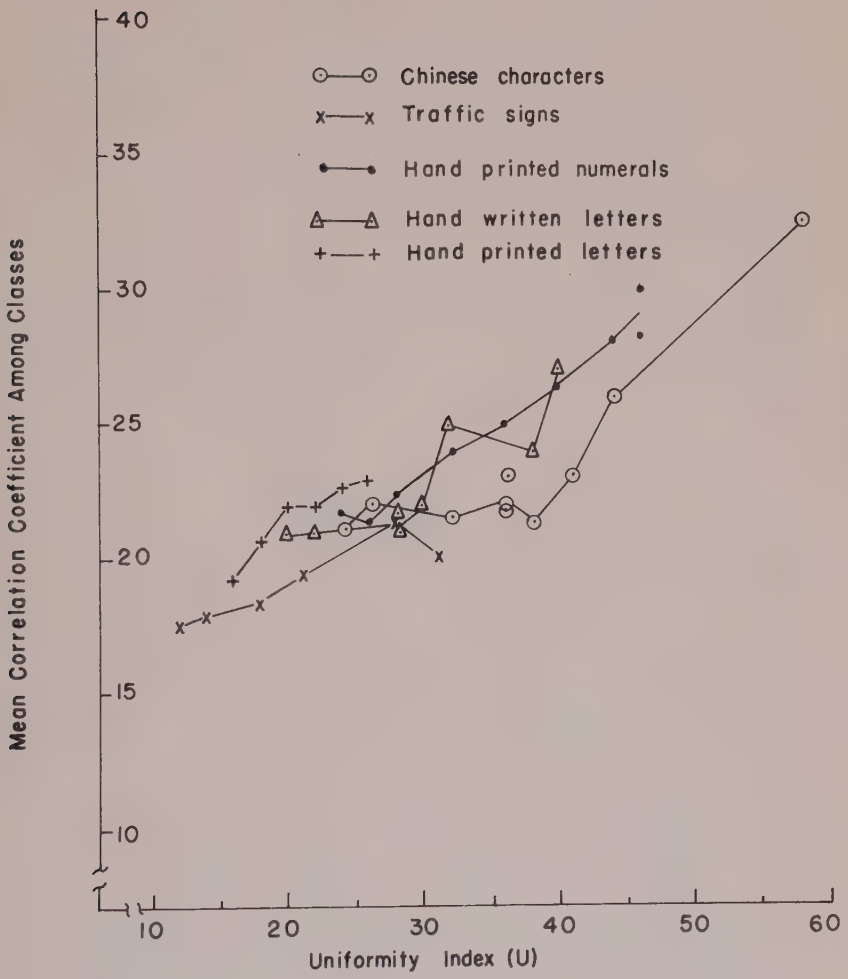


Fig. 8. The Training Curves.

Table II: Recognition Results after Training

Pattern Classes		Training Samples One per Class	Mean Correlation Coeff.	References One per Class	No. of Testing Samples per Class	No. of Correct Recognition per Class
Hand Printed Numerals	2	arbitrarily	21.5	the training samples	10	10
	3	selected			10	7
	5	samples			10	8
	2	arbitrarily	21.5	mean of the ten samples	10	10
	3	selected			10	10
	5	samples			10	10
Hand Written Letters	a	arbitrarily	21	the training samples	12	12
	b	selected			12	12
	c	samples			12	12
Hand Printed Letters	A	arbitrarily	14	the training samples	12	12
	B	selected			12	10
	R	samples			12	6
	A	arbitrarily selected			100	81
					100	95
		representative of 12 samples			100	93
	R	representative of 25 samples			100	97
Chinese Characters	普	itself	21	itself	12 different times	12
	遍	itself				12
	大	itself				12
	學	itself				12
Traffic Signs	⬢	itself	17.5	itself	1	1
	▮	itself		itself	1	1
	▼	itself		itself	1	1
	◆	itself		itself	1	1

CONCLUSION

A pattern recognition machine using BTLE's as basic elements was constructed in the laboratory. The input weights of the elements were fixed, and only the thresholds were made adjustable. No precision component was used, and the room temperature ranged from 65° to 95°F during the test period. Once the thresholds are set, no significant change of the machine performance due to the component imprecision has been observed. The machine has also shown its adaptability to recognizing more general patterns. Preliminary experiments on the recognition of English and Chinese characters, numerals, and traffic signs indicate rather satisfactory results. It is also interesting to note that with a proper set of thresholds, the transformation pair of BTLE's becomes a full adder. It is believed that by taking advantage of the properties of bipolar outputs, more applications of this element may be found.

REFERENCES

1. Lin, W. C., and Fu, K. S. An Adaptive Pattern Recognition Machine Using Bipolar Threshold Logic Devices, *Proc. of the National Electronic Conference* **21** (1965) 535–540.
2. McCulloch, W. S., and Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bull. Math. Biophysics* **5** (1943) 115.
3. Fu, K. S., and Lin, W. C. An Adaptive Pattern Recognition Machine Using Neuron-like Elements, Research Publications of School of Electrical Engineering, TR-EE 65-14, Purdue University, Lafayette, Indiana.
4. Sebestyen, G. S., "Decision-making Process in Pattern Recognition." Macmillan, 1962.
5. Highleyman, W. H., Linear Decision Functions with Application to Pattern Recognition, *Proc. IRE* **50,** (June 1962) 1501–1514.
6. Widrow, B., and Hoff, M. E. Adaptive Switching Circuits, 1960 Wescon Convention Record, Part IV, August 23, 1960.
7. Kautz, W. H. The Realization of Symmetric Switching Functions with Linear-Input Logical Elements, *IRE Trans. on Electronic Computers*, EC-**10** (September 1961) 371–378.

E. M. CONNELLY

Melpar, Inc.

Falls Church, Virgina

Vehicle Control Experiments with Large Artificial Nerve Network (LANNET)

INTRODUCTION

The experimental work described herein is a study of the use of self-organizing trainable logical networks (TLN) as stable controllers in a multi-axis problem. The basic concepts of a TLN are reviewed including the results of theoretical work on the theory of training. The organizational requirements are specified, and a method of achieving rapid organization using a potential function concept is given. The results of an experimental program are presented.

In order to provide some background for readers not familiar with this equipment, a brief description is given. Detailed descriptions are given in references.[1,2]

LANNET[1] is a programmable trainable logical network (TLN), i.e., a logical device that can be trained (organized) to provide any logical function of its input variables. A TLN employs gates termed "statistical switches" which have an adjustable probability of being in an open/closed state. Since each statistical switch is constructed with a finite number of states, each corresponding to a given closure probability, the total device can be considered as a finite state device. Some of these states are statistical while others are deterministic or logical states. The organizational process consists of a sequentially biased trial and error procedure where a training network or goal circuit evaluates results and directs the adjustment of statistical switch states via reward/punish signals. This process tends to transfer the network from undesired states to desired states. It has been shown[2,3] that the training process is a finite Markov process, and that the application of Markov techniques lead to a rather complete theoretical description of the organizational characteristics.

The LANNET accessory equipment[4] recently constructed provides the input/output analog and digital equipment required to interconnect LANNET with the "outside world" and to simulate dynamic problems. This device includes:

(a) Analog-to-digital and digital-to-analog converters for communication with external equipment.
(b) Analog (operational amplifiers) equipment for dynamic simulations.
(c) Paper tape punch and reader for LANNET memory load and dump.
(d) Programmable logic for housekeeping operations.
(e) Abstract property filters for signal separation and classification experiments.

THE CONTROL PROBLEM

A block diagram of the experimental simulation is given in Figure 1. The plant is taken as a space vehicle in the terminal phase of a rendezvous maneuver. A redundant jet configuration for the vehicle is assumed, as shown in Figure 2. While such a configuration may not be practical for an actual vehicle, it is a useful experimental configuration. Inspection of the jet structure shows, for example, if positive translation (only) along the x axis is required, then jet combinations x_1^+, x_2^+, x_3^+, x_4^+; x_1^+, x_2^+; x_3^+, x_4^+; x_3^+, x_2^+; x_1^+, x_4^+ are suitable. If jet x_1^+ fails so that it cannot be energized, then jet combinations x_3^+, x_4^+; x_3^+, x_2^+ can be used. A simple coordinate system, as shown in Figure 3, is selected for the problem.

The vehicle equations of motion are taken as follows:

$$\frac{dx'}{dt} = \frac{1}{M_0}\int_0^t (F_x - F_y \sin \Psi + F_z \sin \Phi)\, dt + \dot{x}'_0$$

$$\frac{dy'}{dt} = \frac{1}{M_0}\int_0^t (F_x \sin \Psi + F_y - F_z \sin \theta)\, dt + \dot{Y}'_0$$

$$\frac{dz'}{dt} = \frac{1}{M_0}\int_0^t (-F_x \sin \Phi + F_y \sin \theta + F_z)\, dt + \dot{z}'_0$$

$$\frac{d\Phi}{dt} = \frac{1}{I_\Phi}\int T_\Phi\, dt + \dot{\Phi}_0$$

$$\frac{d\theta}{dt} = \frac{1}{I_\theta}\int T_\theta\, dt + \dot{\theta}_0$$

$$\frac{d\Psi}{dt} = \frac{1}{I_\Psi}\int T_\Psi\, dt + \dot{\Psi}_0$$

The control simulation is designed to use LANNET outputs to control each jet, and a coded feedback "error" signal is presented to LANNET inputs. However, the total number of jets exceeds the output capability of LANNET, and a three-phase system is employed. During Phase I the LANNET outputs are directed to the X axis jets and, in a similar manner in Phases II and III, the outputs are directed to the Y and Z axis jets, respectively. Thus, the channels subject to control for each phase are given by

Phase	Jets Activated	Control Channels Activated
I	X	X, Ψ
II	Y	Y, Ψ, θ
III	Z	Z, Φ, θ

The three-phase system has the effect of decoupling the vehicle equations given above. Thus, in Phase I, for example, the forces and torques produced by the Y and Z jets are taken as zero. This simplifies the design of the training network.

The control problem is to direct the simulated vehicle from any initial state to the terminal state. Thus, the LANNET must be organized as a controller to provide the required convergence property. As shown in Figure 1, the goal circuit extracts information from the plant and generates reward/punish signals to properly direct LANNET organization. In order to show how this is accomplished, some results of the theory of TLN organization are required.

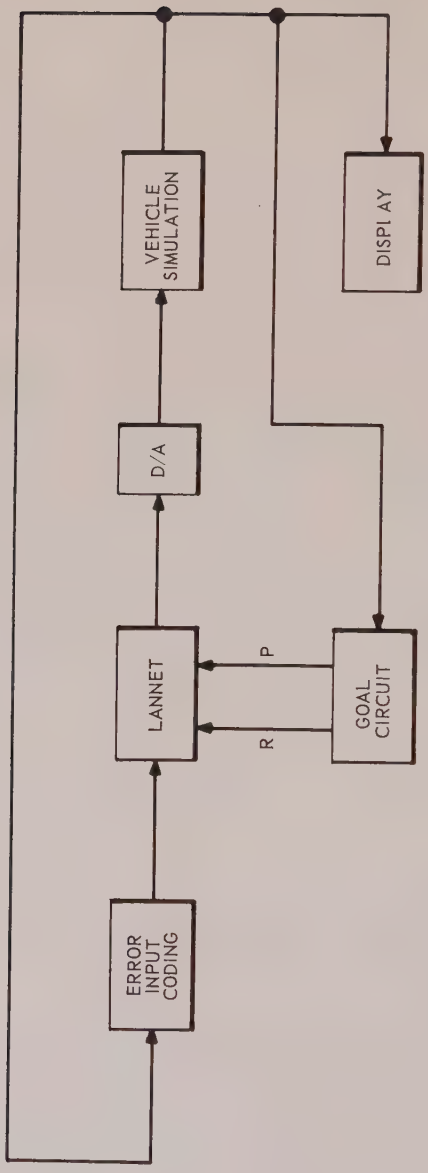


Fig. 1. Control Simulation Block Diagram.

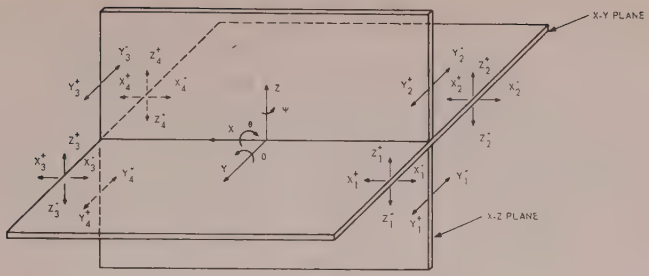


Fig. 2. Jet Configuration.

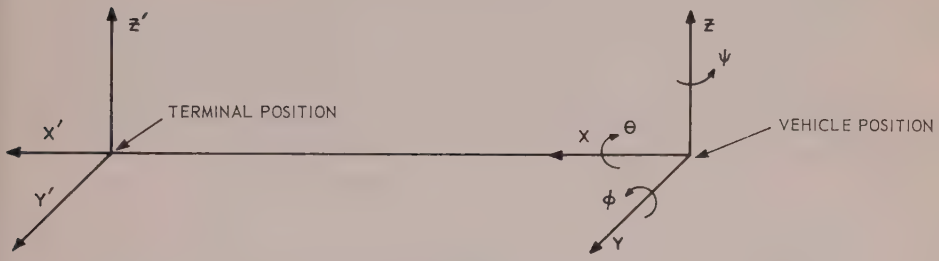


Fig. 3. Space Coordinate System.

1. Development of Goal Functions for Self-Organizing Control Systems

Properties of the Organized TLN

The development of a mathematical representation of the TLN as a finite state device and the organization as a Markov process indicate the role of the goal circuit during training, as well as in establishing certain properties of the organized TLN. This information can be used to develop goals for self-organizing control systems. It can be shown[3] that the critical goal (GC) function is its ability to detect correct circuit operation. Consider the following notations.

Let:

(a) P_{RR} be the probability of generating a reward, given that a reward should be generated.

(b) P_{pp} be the probability of generating a punish signal, given that a punish should be generated.

If the GC is such that P_{RR} equals 1 and, with proper TLN construction, the desired state is an absorbing state. Thus, the system will train to, and remain at, the desired logical controller function. The value of P_{pp} affects the speed of the training process, but the value of P_{RR} characterizes it. To show that all desired states are absorbing states, it is necessary to show that, once the system is in that state, it will never leave (except for retraining due to plant changes or internal TLN failures). This is easily visualized, since a TLN in the correct state will present the correct output, which leads to a reward training signal from the GC. Thus, the system is maintained in the desired state.

If P_{RR} is not equal to 1, the desired state cannot be absorbing, and the organized solution is of a cyclic nature. It is possible to force the cyclic solution to approach the stationary solution by increasing the memory of the statistical switches. However, the important concept is to design the GC such that P_{RR} is maximized.

Training of a Controller

The importance of the goal function to the organization process is easily appreciated, since it governs the nature of the process during training as well as the final organized state of the system. The training of an element in a dynamic closed-loop system forces consideration of not only the system characteristics of the final trained state but also the characteristics (for example, stability) of the system during the training process. These properties are of increasing importance and concern if the organization times approach that of the dynamic system time constants.

An interesting and useful approach is to formulate subgoals wherein certain properties of the solution are specified at each instant of time. If this can be done, then the controller can be trained (and retrained after internal failures) rather rapidly. Additional long-term modification of the controller can also be incorporated to satisfy an overall objective; however, this technique, which is essentially a solution identification, leads to a rather simple search process during training. One aspect of such a search process is that bold (large) changes in parameters can be employed to minimize the effects of noise.

The emphasis of this study is to develop subgoals that lead to the organization of a stable controller. Such a system can be used as an inner loop of an adaptive optimal system.

Lyapunov Functions

The Lyapunov second method[4] has been employed to establish the stability of control systems. The stability characteristic of interest here is the asymptotic stability of a free dynamic system with respect to a point of equilibrium. Consider a set of differential equations

$$\dot{x}_j = f_j(x_1, \ldots, x_n), \quad j = 1, 2, \ldots, n \tag{1}$$

where x_j are state variables. The system is stable in the sense of an asymptotic convergence to the origin if a positive definite function (V) of the state variables can be found where its time derivative ($\dot{V}$) along the trajectory of the system differential equations is negative-definite.

The approach taken here involves the use of this concept from a slightly different point of view. Consider a system of differential equations which represents the system to be controlled

$$\dot{x}_i = f_i(x_1, \ldots, x_n, u_1, \ldots, u_j), \quad i = 1, 2, \ldots, n \tag{2}$$

where x_i is a state variable, and u_k is a control variable. For each controller function (with the system inputs equal to zero), the controller variables can be written:

$$u_k = F_k(x_1, \ldots, x_n), \quad k = 1, 2, \ldots, j \tag{3}$$

Thus, for a given controller function, the system is a free dynamic system, i.e.,

$$\dot{x}_i = f_i(x_1, \ldots, x_n, F_1, \ldots, F_j), \quad i = 1, 2, \ldots, n \tag{4}$$

Now, the approach is to form a Lyapunov function and test the polarity of its derivative. If the derivative ($\dot{V}$) is negative, a relatively simple measurement of this can be used to generate training signals.

Assume that the system with the present controller is asymptotically stable, and the Lyapunov function is such that $\dot{V}$ is negative-definite. The goal circuit employing this technique will have a large P_{RR}, i.e., a probability of 1 (neglecting noise factors) of providing a reward signal, given that a reward is proper. On the other hand, consider a system that is unstable with the present controller and a V that is sign-indefinite, say negative at some regions of the phase space and positive on others. Now, when $\dot{V}$ is negative, the system is rewarded incorrectly, i.e., P_{RP} is not

zero, however, the training theory shows that this property, which perhaps should be minimized, is not critical. Thus, the basic consideration of interest is ensuring that at least one logical function (state of the TLN) provides a stable system. The existence of such a function is dependent (for a given plant) on the Lyapunov function selected and the input-output coding technique employed.

THE EXPERIMENTAL GOAL CIRCUIT

This control problem is an experiment in the use of the derivative of a potential function to obtain training signals for LANNET. Three experiments were run using the three control channels X, Y, Z independently. Several techniques were employed. These are developed and discussed in the following paragraphs.

The first test was to form the function

$$V = |X + k_1\dot{X}| + k_2\,|\dot{X}| + |\Psi + k_3\dot{\Psi}| + k_4\,|\dot{\Psi}|,\ k_1 > 0, k_2 > 0, k_3 > 0, k_4 > 0 \tag{5}$$

for the X axis. The derivative yields

$$\frac{dV}{dt} = \operatorname{sgn}(X + k_1\dot{X})\{\dot{X} + k_1\ddot{X}\} + k_2 \operatorname{sgn}(\dot{X})\{\ddot{X}\} + \operatorname{sgn}(\Psi + k_3\dot{\Psi})\{\dot{\Psi} + k_3\ddot{\Psi}\} + k_4 \operatorname{sgn}(\dot{\Psi})\{\ddot{\Psi}\} \tag{6}$$

the control jets govern the acceleration by the equations

$$\ddot{X} = F_1(X_1, \ldots, X_4^-) \qquad \ddot{\Psi} = F_2(X_1, \ldots, X_4^-). \tag{7}$$

A test was made using equation (5) to determine the training signals, i.e., if

$$\frac{dV}{dt} < 0 \rightarrow \text{Reward}$$

$$\frac{dV}{dt} \geqq 0 \rightarrow \text{Punish}$$

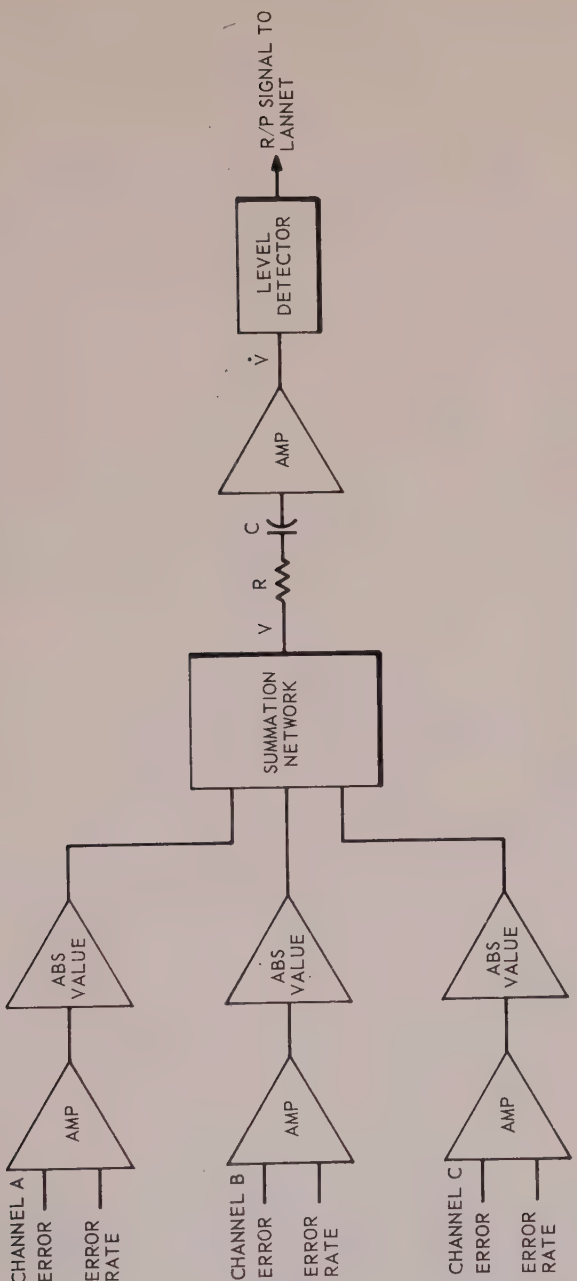


Fig. 4. Goal Circuit Diagram.

Some difficulty was observed in obtaining the polarity of the derivative due to the change in control sensitivity. This can be seen by taking

$$\frac{\partial \dfrac{dV}{dt}}{\partial \ddot{X}} = k_1 \operatorname{sgn}(X + k_1\dot{X}) + k_2 \operatorname{sgn}(\dot{X}) \tag{8}$$

If $k_1 \approx k_2$, the sensitivity is small in the region where

$$\operatorname{sgn}(X + k_1\dot{X}) = -\operatorname{sgn}(\dot{X})$$

and large in the region where

$$\operatorname{sgn}(X + k_1\dot{X}) = \operatorname{sgn}(\dot{X})$$

This problem is solved by setting $k_2 \leqq 1/2\, k_1$ or using the semi-definite form, i.e., $k_2 = 0$.

Correct LANNET organization for the semi-definite potential function formulation requires that the sign of the torque (force) be governed by a linear (switching) line. For example, proper organization of the x channel results in an acceleration equation

$$\operatorname{sgn}(\ddot{x}) = -\operatorname{sgn}(x + k_1\dot{x}). \tag{9}$$

Thus, the input coding for LANNET need only involve functions of the form sgn $(x + k_1\dot{x})$. Figure 4 is a diagram of the simulation using the goal circuit derived from the summation of semi-definite forms.

The sequence of operation was as follows:

(a) Command a LANNET decision
(b) Hold the LANNET output (as jet control commands) for 0.1 second
(c) At the end of the 0.1-second interval, sample the polarity of $\dot{V}$ and generate the reward/punish signal
(d) Return to step a.

The 0.1-second iteration period was selected as a reasonable compromise between a rapid control loop sampling and the filtering required to reduce the noise in computing $\dot{V}$.

EXPERIMENTAL RESULTS

The test results given here are a sample of the tests performed and are representative of the control response and speed of organization obtained. This series of tests was run with a single phase activated on each test.

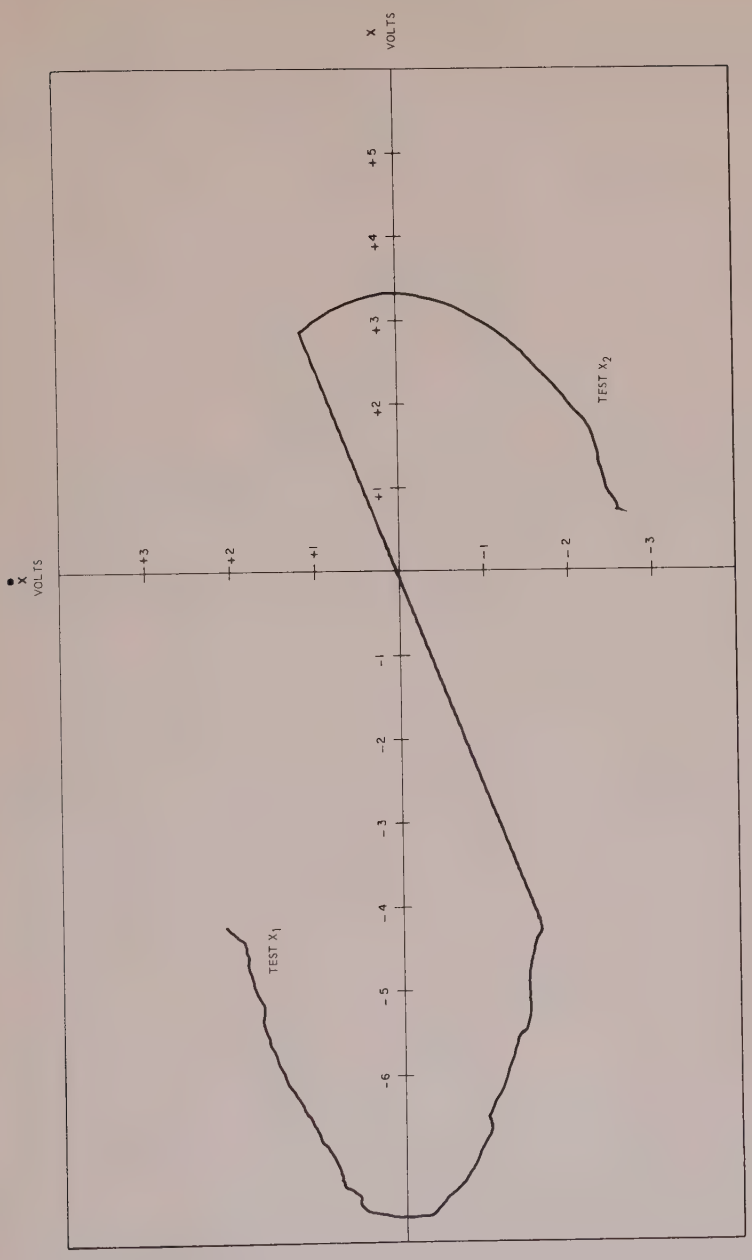


Fig. 5. *X* Channel, *X* Jets

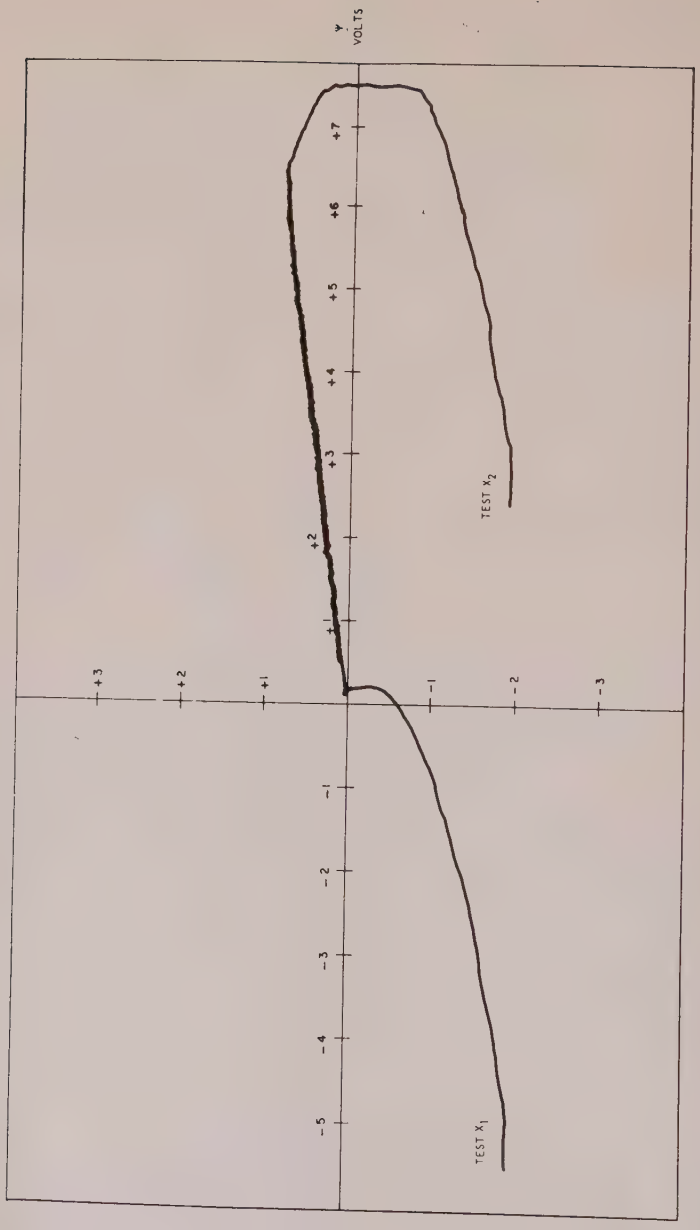


Fig. 6. Ψ Channel, X Jets

The phase plane trajectories shown in Figures 5 and 6 were obtained using the goal circuit shown in Figure 4. Two tests were run, indicated by the trajectories marked X_1, X_2. Prior to each test LANNET was set in the untrained state. The goal circuit equation was

$$V = |X + k_1\dot{X}| + |\Psi + k_3\dot{\Psi}|$$

and the trajectories show that the X channel had a tendency to override the Φ channel on both tests. This occurs whenever

$$\left|\frac{d|X + k_1\dot{X}|}{dt}\right| > \left|\frac{d|\Psi + k_3\dot{\Psi}|}{dt}\right|$$

Thus, as the Ψ trajectories show, the Ψ channel, while convergent, did not respond as rapidly as desired. Note, however, that LANNET was organized to provide stable control for both channels.

Another approach tested was to form the potential function based on the relative magnitude of each channel error function. Thus

$$V_I = |X + k_1\dot{X}| \quad \text{if} \quad |X + k_1\dot{X}| > |\Psi + k_3\dot{\Psi}|$$

$$V_I = |\Psi + k_3\dot{\Psi}| \quad \text{if} \quad |X + k_1\dot{X}| < |\Psi + k_3\dot{\Psi}|$$

A "most positive" diode switching network in place of the summation network shown in Figure 4 is used to provide this function. Thus, except at switching points, the goal circuit receives information from only one channel. This method is especially useful in the Y and Z systems where two angular channels are active in addition to the translation channel.

Let the notation for this "most positive" operation be

$$V_I = \text{MAX}\,\{|X + k_1\dot{X}|,\ |\Psi + k_3\dot{\Psi}|\}$$

The notation for systems II and III are

$$V_{II} = \text{MAX}\,\{|Y + k_5\dot{Y}|,\ |\Psi + k_6\dot{\Psi}|,\ |\theta + k_7\dot{\theta}|\}$$

$$V_{III} = \text{MAX}\,\{|Z + k_8\dot{Z}|,\ |\Phi + k_9\dot{\Phi}|,\ |\theta + k_{10}\dot{\theta}|\}$$

With this approach, LANNET organization is governed by the channel with the largest potential. This semi-definite form $x + k\dot{x}$ may be considered as a measure of distance from its switching line. Thus, in this sense

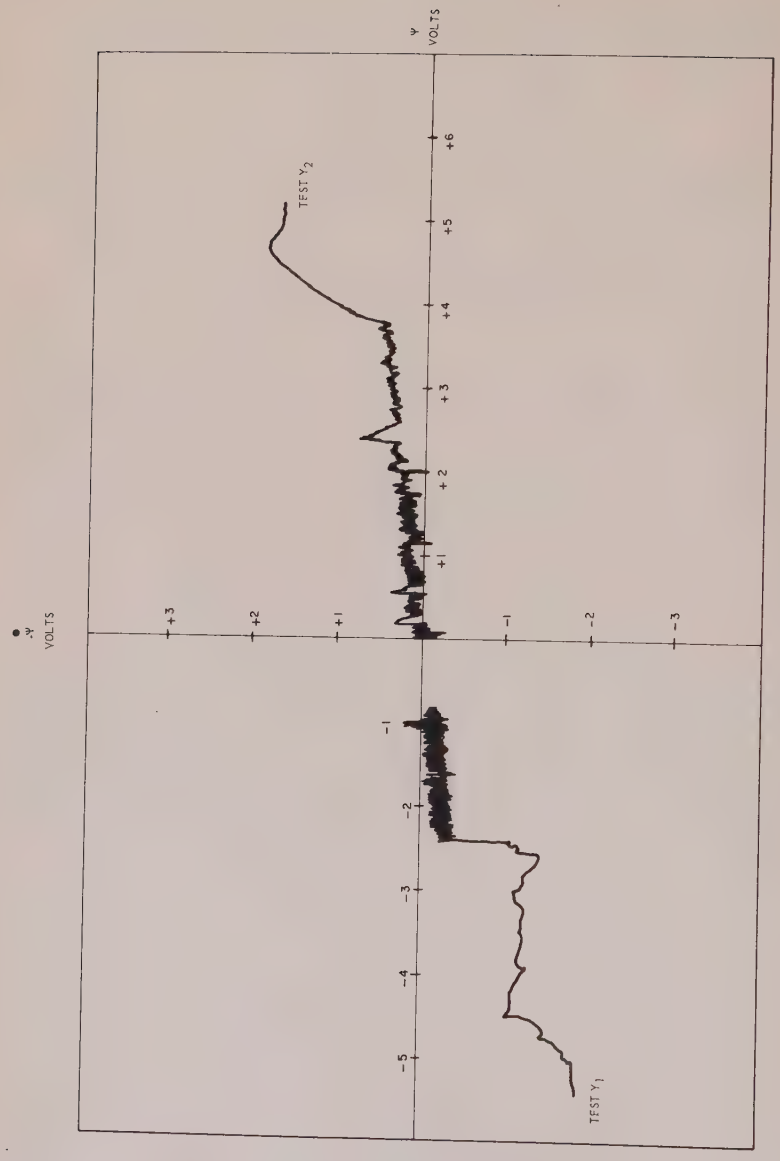


Fig. 7. Ψ Channel, Y Jets.

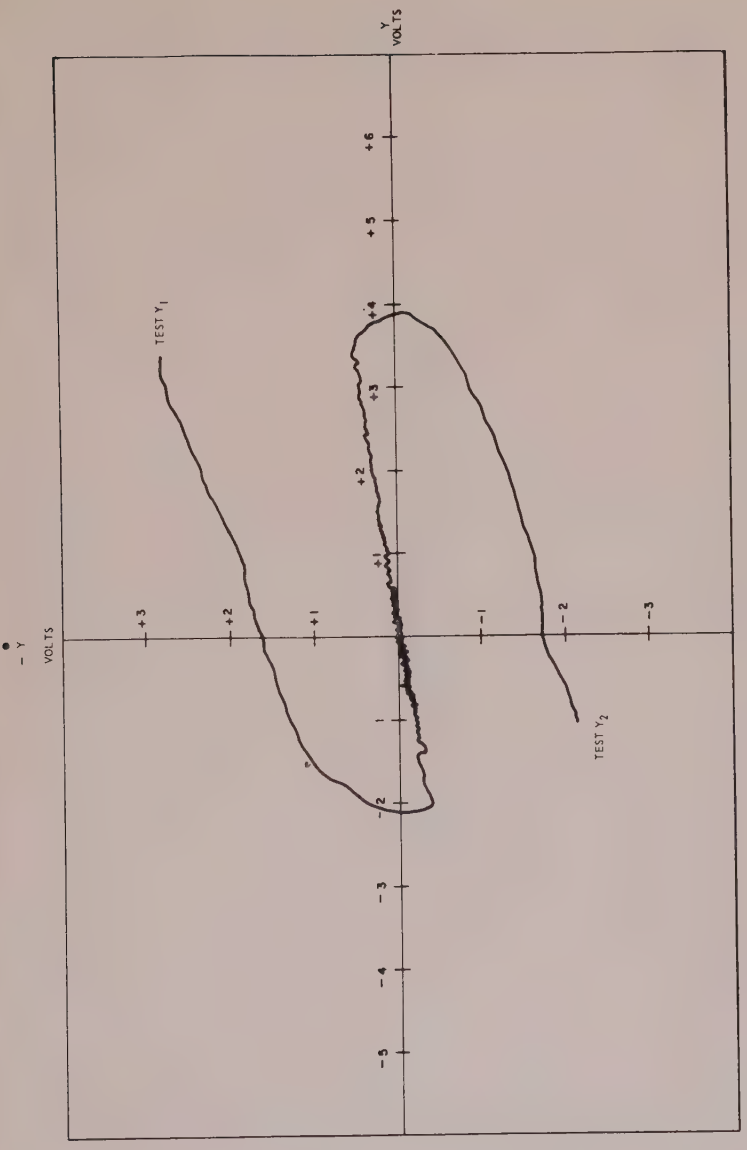


Fig. 8. *Y* Channel, *Y* Jets.

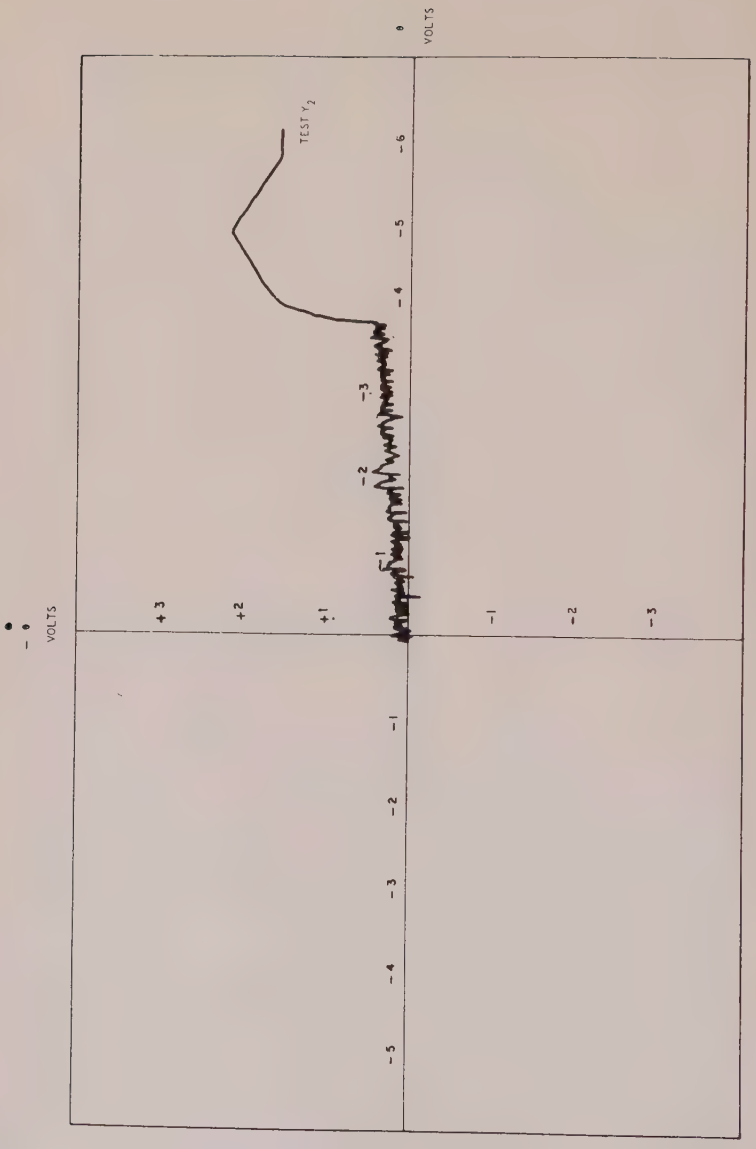


Fig. 9. θ Channel, Y Jets.

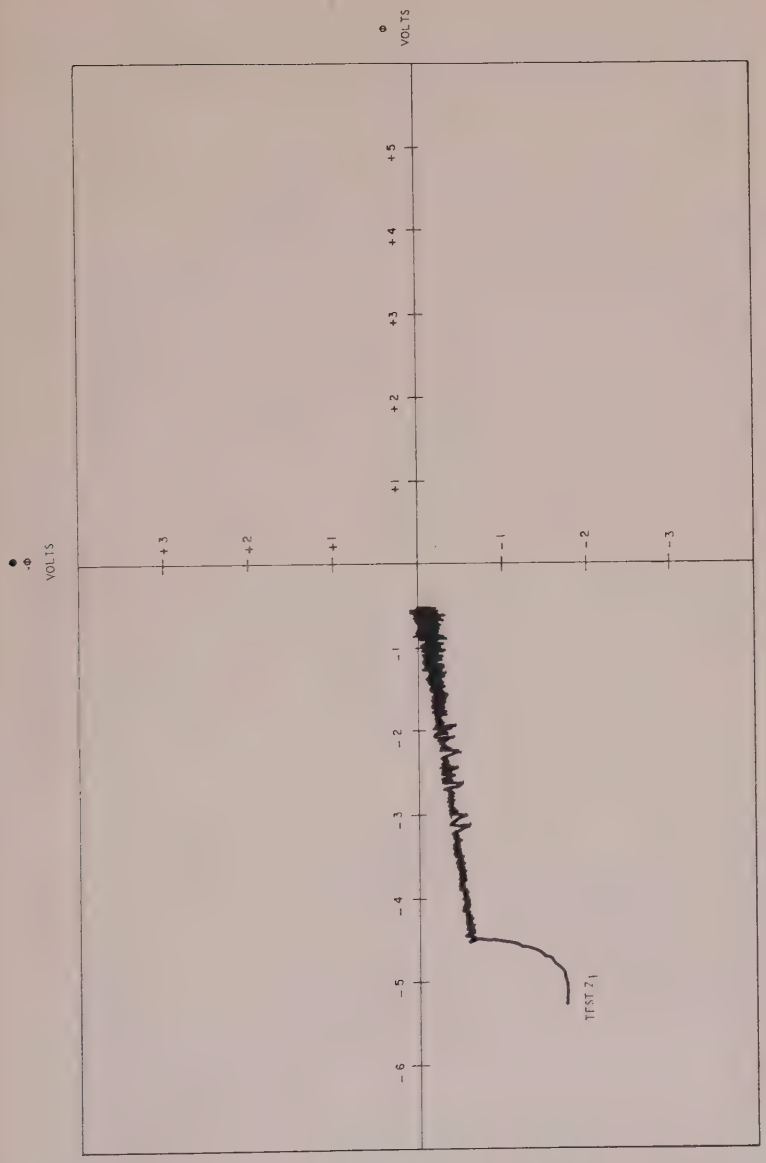


Fig. 10. Φ Channel, Z Jets.

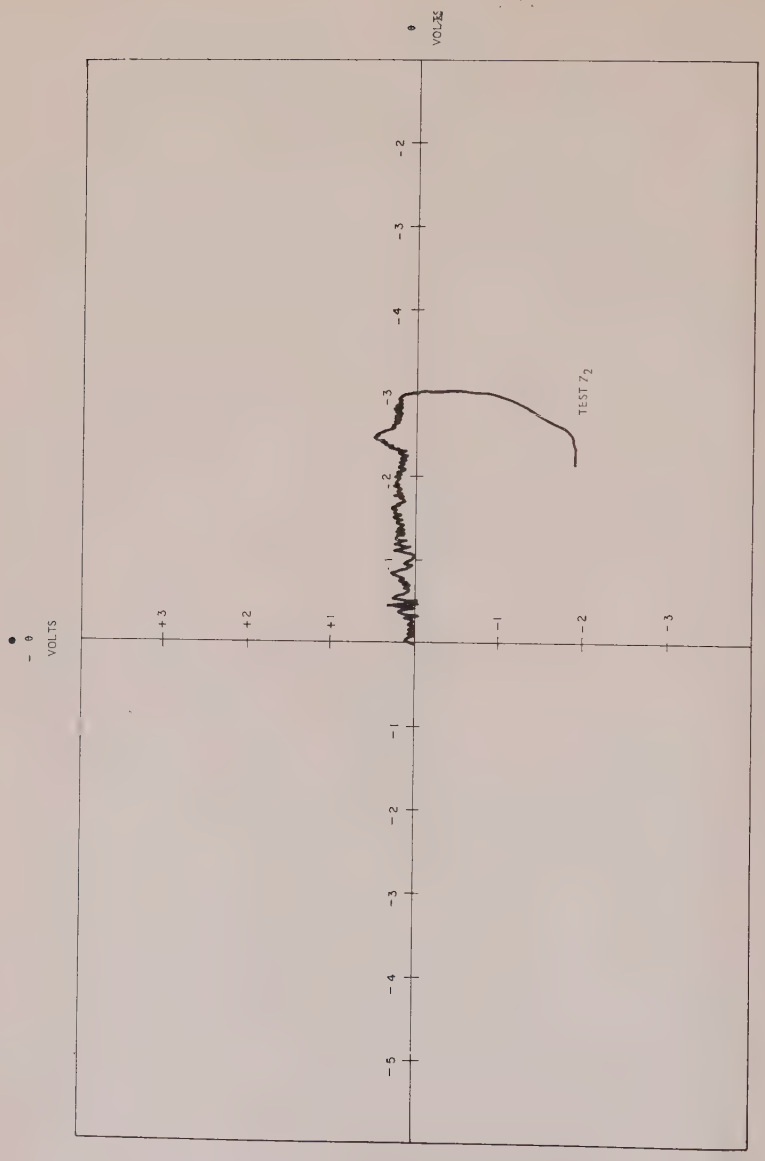


Fig. 11. θ Channel, Z Jets.

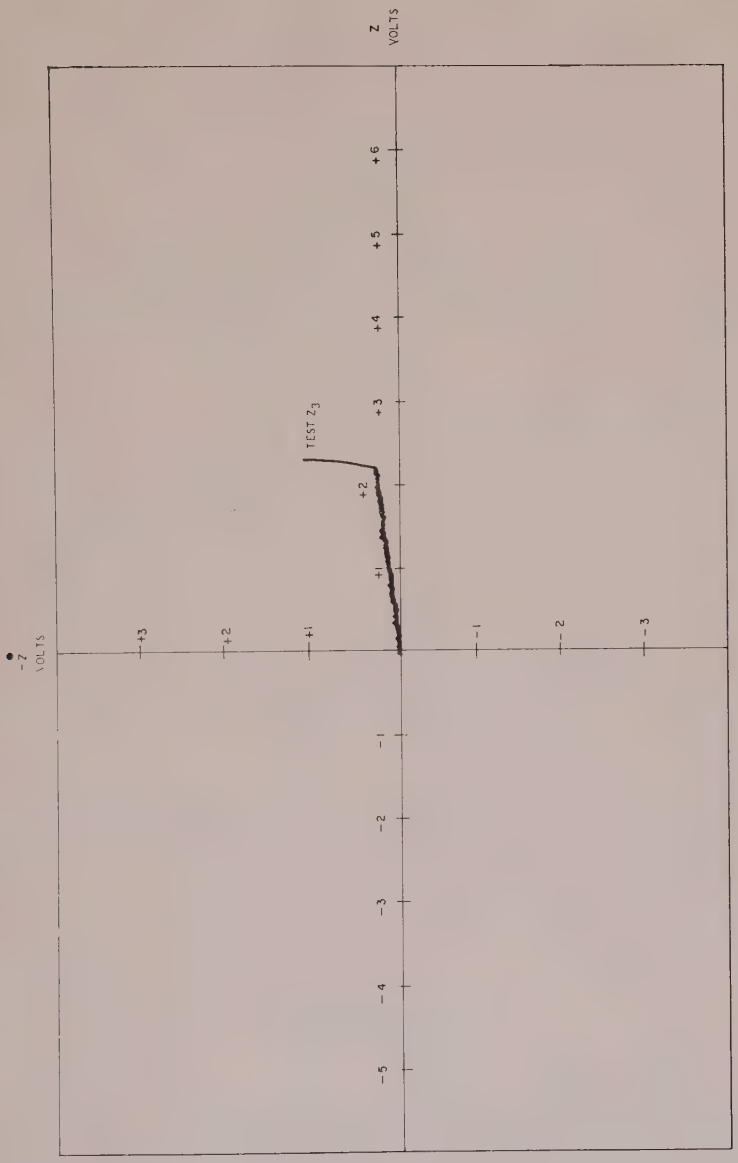


Fig. 12. *Z* Channel, *Z* Jets.

the most positive potential is associated with the channel at the largest "distance" from its switching line. This technique then operates such that LANNET is organized to cause convergence of the channel most distant from its switching line. Such organization may or may not cause other channels to be divergent and the divergence is masked by the most positive circuit until the associated channel potential becomes largest. The goal circuit, in effect, monitors the convergence of the system with the largest potential. If at least one LANNET configuration can provide convergence in all channels, then the LANNET learning associated with the stabilization of one channel influences or biases the search when operating with other channels. Thus, the organization process depends on present as well as past information, and the system can rapidly achieve a suitable state.

This technique was used in experiments Y_1 and Y_2 with phase plane trajectories shown in Figures 7 and 8. Both tests were run after placing LANNET in the untrained state. Test Y_1 was run using channels Ψ and Y controlled by the Y jets. Inspection of the trajectories indicates that the channel was initially convergent, but later drifts for a period of time and subsequently is reorganized to a convergent condition. This is due to the most positive goal circuit characteristic described above.

Test Y_2 (Fig. 7, 8, and 9) is a three-channel (θ, Ψ, Y) problem using the Y jets. Again, LANNET was initially set to the untrained state. Note that the θ and Ψ channels are divergent for a short time interval before LANNET is completely organized.

The total response of the system after LANNET organization, i.e., after the first run, is shown in Figures 10, 11, and 12. This is the second run of the Z, θ, Φ channels as controlled by the Z jets.

It is seen that once LANNET is organized the individual channel response is rapid.

CONCLUSIONS

The objective of the experimental program was to determine the properties of a self-organizing control system using the potential function approach. The results of the program indicate that LANNET can be properly organized on-line, and the organization time required is less than the plant time constants.

REFERENCES

1. Guinn, David F. Large Artificial Nerve Network (LANNET), *IEEE Transactions on Military Electronics*, Vol. Mil 7, Bionics Issue (L. M. Butsch and H. L. Oestreicher, eds.), pp 234-243 (1963).
2. A Study of Generalized Machine Learning, *Final Report AF* **33(616)-7682**, Fe*bruary* **1962**, ASD-TDR-62-166; and "A Study of Generalized Machine Learning," *Final Report AF* **33(616)-7682**, August 1963, ASD-TDR-63-714.
3. Feasibility Studies of Use of ARTRONS as Logic Elements in Flight Control Systems, *Final Report AF* **33(657-11026)**, February 1964, FDL-TDR-64-23.
4. Research of Specialized Equipment for Use With a Large Artificial Nerve Network, *Final Report AF* **33(615-1270)**, July 1965.
5. Hahn, Wolfgang."Theory and Application of Lyapunov's Direct Method," Prentice-Hall, Inc., Englewood Cliffs, N. J., 1963.

J. M. IDELSOHN

and R. M. CENTER

The Bendix Corporation

Research Laboratories Division

Southfield, Michigan

A. SPEAKE

Air Force Avionics Laboratory

Bionics Branch

Wright-Patterson Air Force Base, Ohio

Application of Bionics to Spacecraft Energy Allocation

INTRODUCTION

The energy allocation subsystem of a spacecraft distributes electrica energy from alternate sources to various loads through interconnecting switches. The control problem is to specify various interconnections which will make efficient use of available energy. Because of the unpredictable changes in environmental conditions and internal parameters, the energy allocation subsystem must be continually controlled.

This paper describes a study in which a conventional, programmed control approach was compared with a self-adaptive approach which uses PSV (Probability State Variable) devices. The study was sponsored by the Bionics Branch of the Air Force Avionics Laboratory as part of its work in evaluating bionic learning machine techniques. The objective of the study was to obtain comparative data to evaluate the self-adaptive approach for practical applications.

In the study, a bionic control network was designed using a group of Neurotron † PSV devices; this network was trained to optimize the interconnection of energy sources and electrical loads in accordance

† A description of the Neurotron is given on pages 873–5.

with a predetermined goal function. Also, a conventional programmed controller was designed on the basis of assumed a priori information. A digital computer was used to model the spacecraft energy system and the two alternative controllers, and also to obtain comparative performance data for a variety of simulated missions. In addition to the evaluation by means of computer simulation, a prototype Neurotron was designed, built, and tested.

The application considered in this study makes only partial use of the Neurotron concept; while Neurotron devices have both digital and analog capabilities, only the digital capabilities were used in the spacecraft energy allocation problem.

PROBLEM FORMULATION

The energy allocation subsystem uses solar cells, backed up by a secondary (chargeable) battery for dark periods. The components of this subsystem are a battery, solar panels, and a load divided into high-priority and low-priority sections.

In actual operation, the subsystem components are subject to certain variations. The solar panel output is a function of the amount of solar radiation impinging on the panels, the ambient temperature, the number of active solar cells, and the load on the panels. Simulation of these

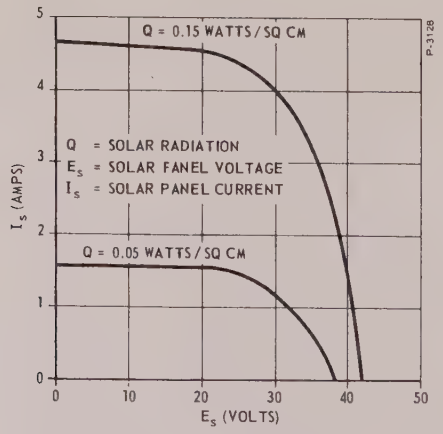


Fig. 1. Typical Solar Panel Characteristics.

effects was accomplished according to an exponential equation which served to closely approximate the characteristics of actual commercially available units.[1] Figure 1 illustrates the resulting volt-ampere characteristics of the simulated solar panels under two conditions of solar radiation. The battery output is a function of battery history, internal leakage, and battery load.[2,3] The discharge characteristics of simulated batteries having different ampere-hour capacities are illustrated in Figure 2. In addition to

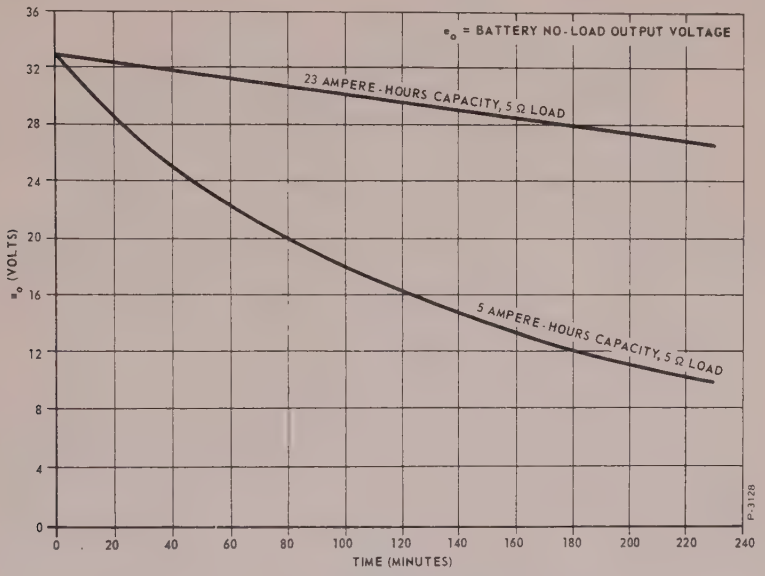


Fig. 2. Typical Battery Discharge Characteristics.

the solar panels and the battery, the high- and low-priority loads are time-varying in accord with the demands of the various electrical equipment, and are in part dependent upon interaction of the spacecraft with the environment. Figure 3 illustrates the various schedules of load and environmental parameters which were used in the simulation study.

The problem of electrical energy allocation is to determine the best interconnection of the elements as a function of time, given partly unpredictable component and environmental variations. An a priori knowledge of the subsystem component characteristics plus a general idea of the subsystem goal was used to arrive at a configuration for the energy allocation subsystem. This configuration may be realized by a

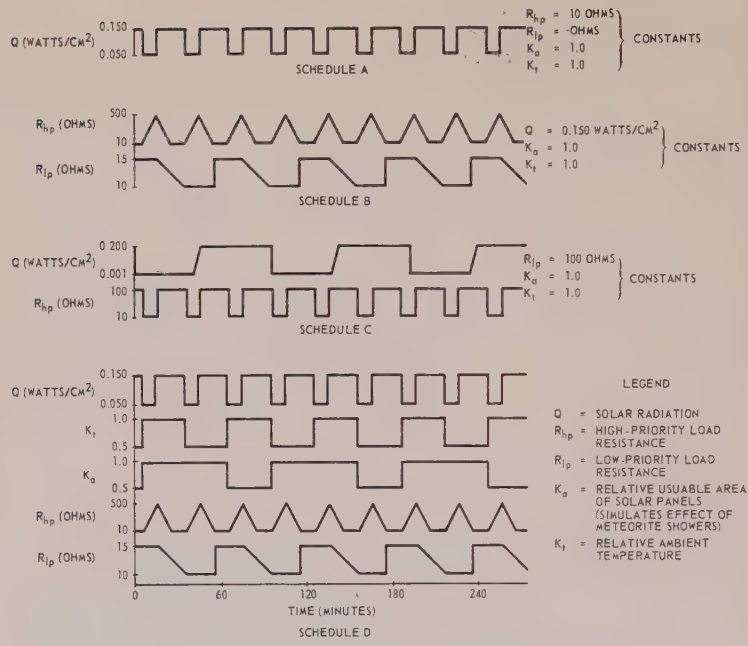


Fig. 3. Schedules of Load and Environmental Parameters

set of switch (or relay) contacts, as shown in Figure 4. In the simulation study, measures of battery output, solar panel output, and load requirement were provided by no-load battery voltage radiation impinging on the solar panels, and the resistance values of the loads, respectively.

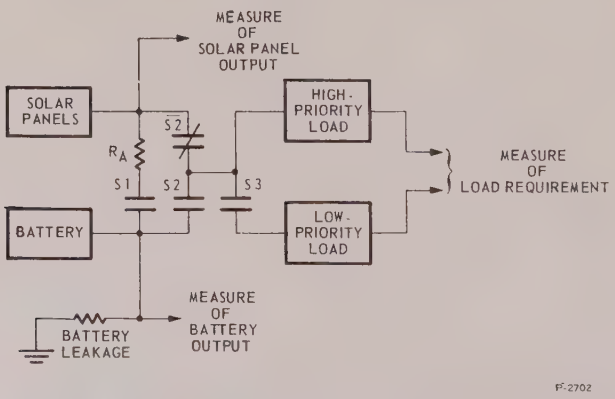


Fig. 4. Component Interconnections of Spacecraft Energy Allocation Subsystem

The goal of the energy allocation subsystem is to minimize the net cost to the system due to equipment failure as a result of inadequate load satisfaction, and due to lack of reliability as a result of maintaining less-than-normal battery charge. A cost function, H, has been defined as:

$$H = H_0 + i_b \cdot \left[\frac{E_0(\max) - e_0}{E_0(\max)}\right] \cdot H_b + P_{lp}H_{lp} + P_{hp}H_{hp} \qquad (1)$$

where

H_0 = constant = 1.0*
H_b = cost of maintaining less-than-normal battery voltage = 4.0
H_{lp} = cost of low-priority equipment failure = 2.0
H_{hp} = cost of high-priority equipment failure = 4.0
$E_0(\max)$ = maximum no-load battery voltage = 33 volts
e_0 = actual no-load battery voltage
i_b = battery current
P_{lp} = probability of low-priority equipment failure due to insufficient voltage
P_{hp} = probability of high-priority equipment failure due to insufficient voltage

The values of P_{lp} and P_{hp} were assumed to be directly related to the voltage applied to the loads according to the relationship plotted in Figure 5.

The function of the controller is to generate appropriate S1, S2, and S3 commands, given measures of battery output, solar panel output, and load requirements. The desirability of one set of switch commands relative to another can be determined by rating each possible set of commands in terms of the resultant degree of goal satisfaction. The specification of a subsystem goal is therefore implicit in the design of a controller, whether it be a programmed or a self-adaptive device. Knowledge of the goal was used to establish the control law in the programmed controller and was therefore incorporated into the design. In the self-adaptive controller, the goal was entered from an external goal circuit which computed the actual cost values from Eq. (1), and supplied REWARD and PUNISH signals to the controller. Figure 6 illustrates the

* The numerical values shown for the cost constants were those used in the simulation study and represent a particular choice of the relative importance of each term in Eq. (1).

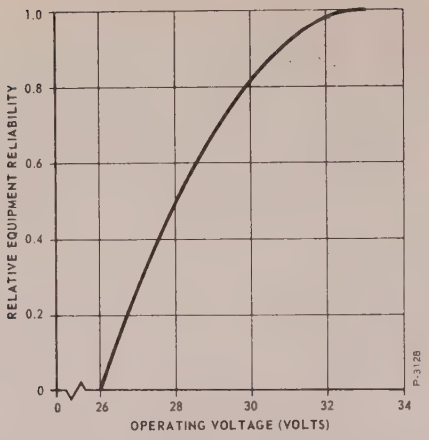


Fig. 5. Equipment Reliability Versus Operating Voltage

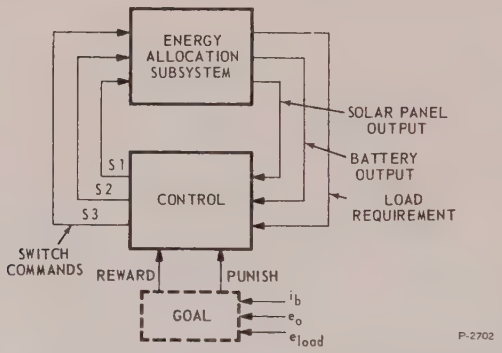


Fig. 6. Block Diagram for Control of Spacecraft Energy Allocation Subsystem.

relationship of the controller to the energy allocation subsystem. Four separate computer subroutines were used in the simulation study of the energy allocation subsystem and associated controller. These enabled simulation of the subsystem, the goal evaluation circuit, the programmed controller, and the self-adaptive controller, respectively.

PROGRAMMED CONTROLLER

A relatively straightforward control law was used in the programmed controller. This law specifies binary decisions based upon measures of battery and solar panel outputs. The battery output is classified as

sufficient if the actual no-load battery voltage, e_o, is greater than 30 volts. Similarly, the solar panel output is classified as sufficient if the radiation impinging on the panels exceeds 100 milliwatts per square centimeter (this corresponds to a no-load solar panel output of approximately 40 volts).

The programmed controller effects interconnections between subsystem components according to the flow diagram of Figure 7.

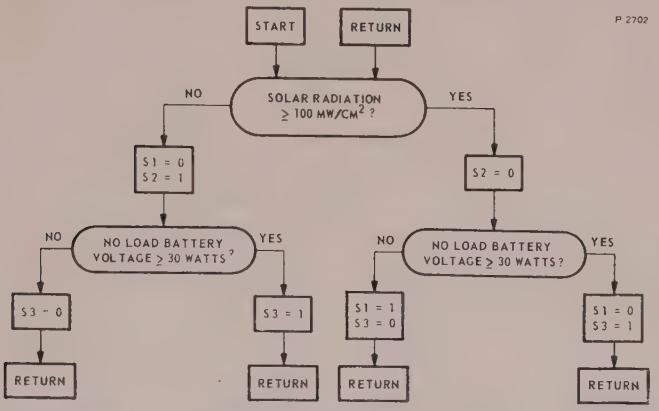


Fig. 7. Flow Diagram for Programmed Computer Control of Spacecraft Energy Allocation Subsystem.

SELF-ADAPTIVE CONTROLLER

The self-adaptive controller for the spacecraft energy allocation subsystem is made up of a network of Neurotron elements. In its most general form, the Neurotron is a two-input, one-output device which can form functions having both analog and digital properties. Only the digital properties are required for control of the energy allocation subsystem, however, and the description which follows is therefore limited to the digital section. A complete description of the Neurotron is contained in References 4 through 9.

1. General Description

The Neurotron device stores memory, or learning, as the probability of having particular transfer functions between input and output. The distinguishing characteristic of the device is that its operation is prob-

abilistic. That is, its state at any given time cannot be explicitly determined, but can only be specified in a probabilistic manner. For example, if $S_1, S_2, \ldots, S_M$ represent the possible states of the device, its state at any given time can only be represented by a corresponding set of probabilities, $P(S_1), P(S_2), \ldots, P(S_M)$, where, in general, $P(S_m)$ represents the probability of state S_m. Each state of the Neurotron digital section corresponds to a particular logic function, $Q_1(A, B), \ldots, Q_M(A, B)$, between inputs, A and B, and output, Q.

By adjusting the stored probabilities in accordance with desired input-output relationships, the Neurotron can be made to arrive at a desired transfer function. This is accomplished by the adjustment of certain stored probabilities through the application of REWARD or PUNISH signals. Repeated application of these signals can cause the device to converge upon a set of stored probability values which are optimum for a given situation. In a like manner, networks can be made to exhibit self-adaptive behavior, and they can be trained to solve a variety of complex problems. As applied to the energy allocation problem, the self-adaptive controller consists of a network of Neurotron devices, trained such that Neurotron states which lead to a low cost index tend to be preserved, and Neurotron states which lead to a high cost index tend to be destroyed. By utilizing a sufficiently large number of devices, a degree of redundancy and attendant high reliability can be achieved.

The digital section of the Neurotron is a two-input, one-output device, as shown in Figure 8. The inputs A and B are represented as pulse trains

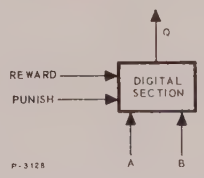


Fig. 8. Basic Neurotron Digital Section.

whose pulse density, or average number of pulses per unit time, is proportional to the amplitude of its respective variable. The pulse trains are unsynchronized, but have a uniform pulse amplitude and width. The output of the digital section, Q, is a similar pulse train, the density of which is a function of the inputs, A and B, and of the state of the digital section. The state of the digital section can be any one of the 16 possible

binary logic functions of two variables, and is controllable by REWARD and PUNISH signals. Only the digital section of the Neurotron is required when the problem does not involve dynamic considerations or the control of analog parameters, as was the case in the spacecraft energy allocation problem.

2. Application of the Self-Adaptive Controller to the Energy Allocation Problem

The performance of the self-adaptive controller was evaluated by conducting a number of simulated missions with three different Neurotron networks and a variety of training strategies. A Neurotron network can be classified as either monotype or genotype, depending upon the method used to formulate the network configuration. A monotype network configuration is formulated from a set of deterministic interconnec-

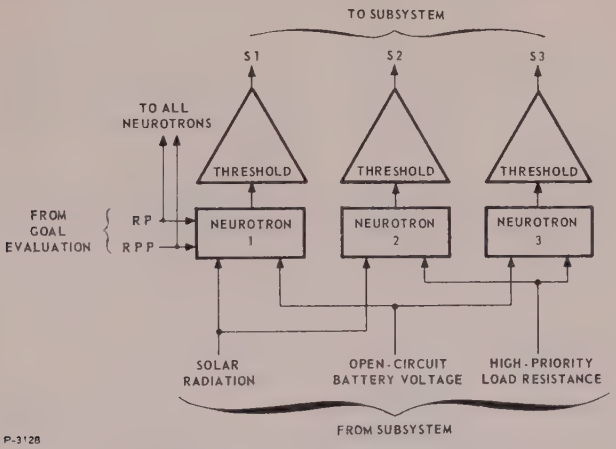


Fig. 9. Three-Neutron Monotype Network

tion rules, and a genotype network is formulated from a set of probabilistic interconnection rules. Two monotype networks and one genotype network were simulated during study of the spacecraft energy allocation problem. Figures 9 and 10 illustrate each type of network.

The inputs to all simulated networks included the solar radiation impinging on the panels, the open-circuit battery voltage, and the high-

priority load resistance. The outputs of the network, which normally are pulse densities, were converted into switch states, $S1$, $S2$, and $S3$ by comparison of the pulse density values with a threshold value.

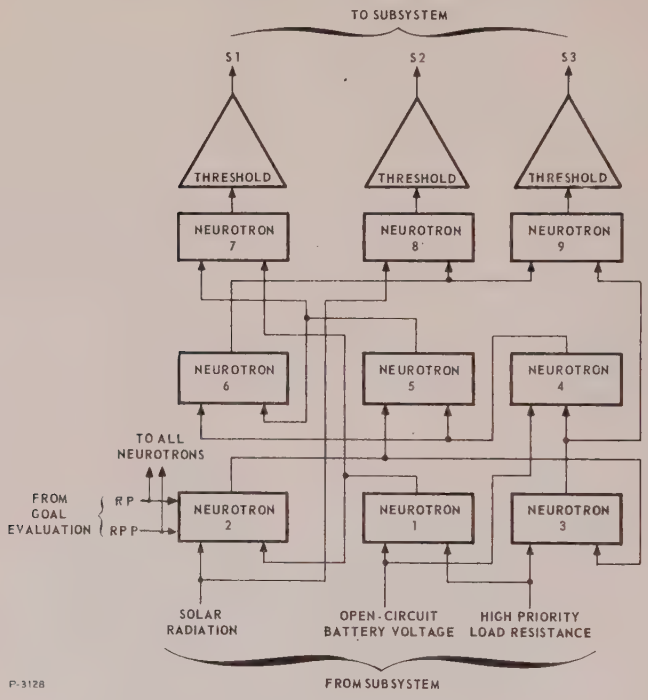


Fig. 10. Nine-Neurotron Genotype Network

METHOD OF CLOSED-LOOP SIMULATION

A general purpose digital computer was used to simulate the spacecraft energy allocation subsystem, the goal evaluation circuit, the programmed controller, and the self-adaptive controller. The subsystem configuration was that shown in Figure 4, with solar panels and the battery simulated in accordance with empirically derived expressions. A flow diagram of the overall closed-loop system is shown in Figure 11.

Referring to the figure, the initial computation determines the various environmental and load factors according to one of the schedules pre-

sented in Figure 3. Next, the battery condition is determined by assuming a fixed discharge rate over the previous incremental time unit. The interconnection between subsystem components is then established by interrogating the state of switches $S1$, $S2$, and $S3$ previously commanded by the controller, and solutions for the resultant currents and voltages are obtained. Finally, the cost and the goal indices are computed, completing the subsystem computations.

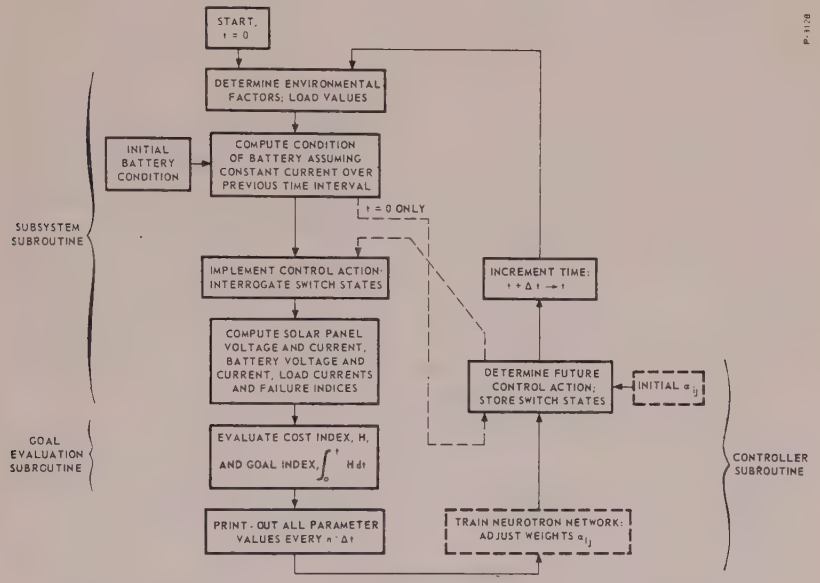


Fig. 11. Flow Diagram for Energy Allocation Closed-Loop Simulation

The controller subroutine is then entered, and a set of future switch settings is computed according to the rules governing the programmed controller or the self-adaptive controller. The entire procedure is then repeated for succeeding time increments, thereby approximating a continuous system.

To obtain a comparative evaluation of the alternate controllers, a number of computer runs were made with each run subjecting both controllers to an identical test. Each test was characterized by a specific schedule of parameter variations, and performance data was obtained for the three Neurotron network configurations, in addition to the programmed controller.

SUMMARY OF RESULTS

The results obtained from closed-loop computer simulations of both programmed and self-adaptive (Neurotron) control of the spacecraft energy allocation subsystem are summarized in the following paragraphs. The performance of a given closed-loop simulation was rated according to the final value of the goal index. For operation of the subsystem over the interval $0 \leqq t \leqq T$, the final value of this index is given by

$$\int_0^T H dt$$

The smaller this integral, the better the actual performance of the subsystem in terms of originally specified goal requirements (as defined by choice of the cost constants in the expression for H). The values of this cost integral were determined for each test, and this data provided the basis for comparison of the two types of control. Only goal indices for simulations run under identical test conditions were directly compared, and in terms of the original cost constants, H_b, H_{hp}, and H_{lp}, the percentage improvement in one type of control over another was given by the corresponding reduction in this index.

Table I illustrates representative test conditions to which each controller was subjected during the evaluation program. The environmental schedules were those presented in Figure 3, and are of increasing complexity progressing from *A* to *D*. The subsystem configuration was as shown

Table I: Representative Test Conditions Used in Comparative Evaluation

Test Condition Number	Environmental Schedule	Battery Capacity (Ampere Hours)	Leakage Resistance (Ohms)	Charge Resistance, R_A(Ohms)	Comments
1	A	5	1,030	1	Nominal Conditions
2	B	23	1,030	1	Increased Battery Capacity
3	C	5	1,030	1	Nominal Conditions
4	C	23	1,030	1	Increased Battery Capacity
5	C	5	50	1	Degraded Battery
6	C	5	10	1	Severely Degraded Battery
7	C	5	1,030	20	Degraded Charge Path
8	D	5	1,030	1	Nominal Conditions
9	D	23	1,030	1	Increased Battery Capacity

in Figure 4, with nominal battery capacity of five ampere-hours, and nominal values of R_A and battery leakage resistance of one ohm and 1030 ohms, respectively. As shown in the table, departures from these nominal values were made in some of the tests. This was done for the purpose of investigating the effects of increased battery capacity (tests 2, 4, and 9), the effects of battery degradation due to a low-resistance leakage path to ground (tests 5 and 6), and the effects of battery charge current limited by a large value of R_A, as might occur due to degraded relay contacts or corroded battery connections (test 7).

Several closed-loop runs were made under each test condition for the purpose of evaluating the various Neurotron control network configurations. Initially, the use of nine Neurotron networks was investigated. It was subsequently found that nearly comparable performance was achieved using a three-Neurotron network, and further effort was then concentrated on evaluating this control configuration.

A summary of the results of a variety of representative closed-loop runs is contained in Table II. This table is a tabulation of the accumulated costs (goal indices) over 12 hours of simulated operation for each run;

Table II: Accumulated Costs After 12 Hours Operation Under Different Test Conditions

Test Condition Number	Type of Controller			
	Programmed	3-Neurotron Monotype	9-Neurotron Monotype	9-Neurotron Genotype
1	4057	2964 (—27%)		
2	2253	984 (—56%)		
3	1998	1122 (—44%)		984 (—51%)
4	1020	1050 (+3%)		
5	2417	1889 (—22%)		
6	4468	4367 (—2%)		
7	2953	2560 (—13%)		2446 (—17%)
8	3667	3372 (—8%)	3919 (+7%)	4977 (+36%)
9	3338	2937 (—12%)		2740 (—18%)

included are results obtained from programmed control and from various types of self-adaptive control under the test conditions tabulated in Table I. Also indicated in the table is the percentage change in the accumulated cost (relative to the value obtained for programmed control)

as a result of using each type of Neurotron controller. Negative percent values represent cost reductions, or improvements in performance, and positive values indicate performance degradation.

With the exception of certain isolated cases, the Neurotron controller provided a general improvement in subsystem performance over the programmed controller. As mentioned earlier, the degree of improvement achieved in any particular case is directly proportional to the percent reduction in the accumulated cost. The variation in the comparative results can be attributed to two factors. First, the Neurotron controller is probabilistic in operation and its performance in any given test is therefore subject to statistical variation. This is particularly apparent in the operation of the nine-Neurotron controller. Second, the programmed controller design is closer to optimum for certain sets of test conditions than for others. Therefore, its performance relative to the Neurotron controller is dependent upon the degree of departure of the test conditions from the a priori optimum set.

Figures 12 and 13 illustrate the performance of various controllers as a function of time, for representative test runs. Figure 12 is a plot of the cost function, H, over 12 hours of simulated operation of both the programmed and three-Neurotron controllers under test condition 1 (see Tables I and II). The uppermost plot shows the variation of the solar radiation with time according to environmental schedule A (with time scale applicable to both H plots). It may be seen that the cost function resulting from both types of control tends to follow the cyclical variation of the solar radiation, with higher costs occurring during periods of eclipse, as expected. Of significance is the fact that the cost under Neurotron control is consistently less than the cost under programmed control. This accounts for the substantial difference between the respective accumulated costs (the areas under each curve). It should be noted that the cost differential between the two curves grows as time passes. This is due to a faster battery discharge in the case of the programmed controller, and indicates that the Neurotron controller is more efficient at replacing the battery energy which is consumed during eclipse periods.

A representative plot of a run made under more complex environmental conditions is shown in Figure 13. Included are cost function plots for the programmed controller, the three-Neurotron monotype controller, and the nine-Neurotron genotype controller, all subjected to test con-

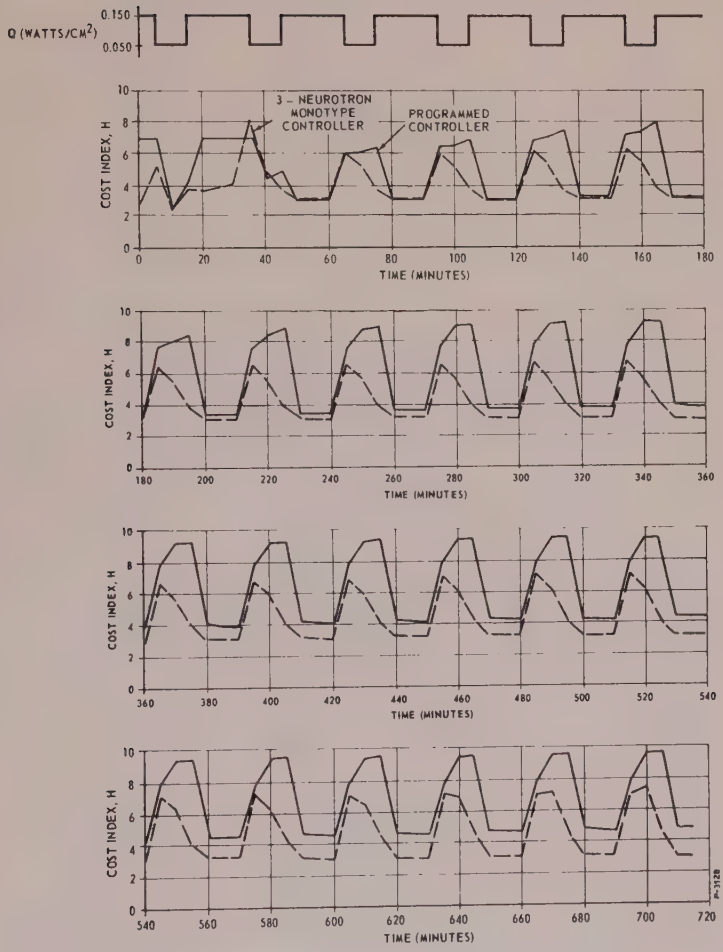


Fig. 12. Cost Index Versus Time, Test Condition I

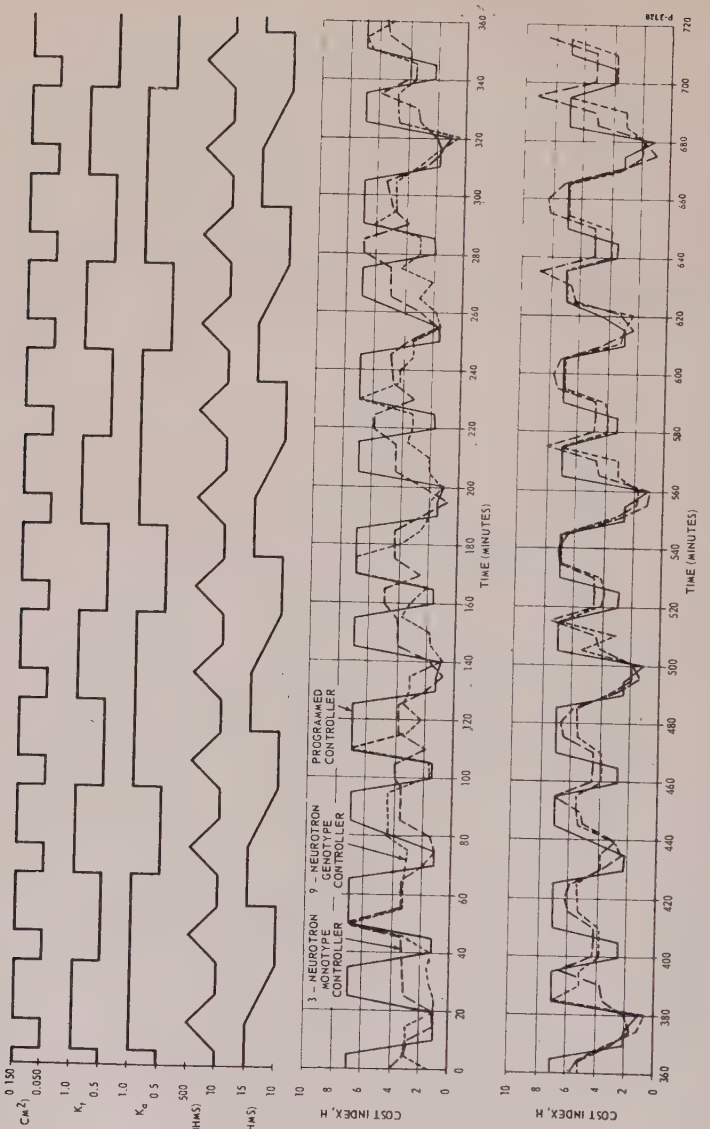


Fig. 13. Cost Index Versus Time, Test Condition 9

dition 9. Although a cyclical pattern may be discerned in both of these figures, the cost variations are more complex than those of Figure 12. It may be seen that on the average, the cost curves for both Neurotron controllers are below those for the programmed controller. This may be verified by reference to the corresponding accumulated cost values in Table 2.

CONCLUSIONS

This paper has presented the results of a variety of computer simulations comparing the effectiveness of self-adaptive (Neurotron) control of the energy allocation subsystem with that of programmed control. Using the value of the goal index as a criterion, the Neurotron controller, compared with the programmed controller, provided a general improvement in subsystem performance for the test conditions considered. Use of the nine-Neurotron genotype controller resulted in the greatest improvement in certain cases, but operation was subject to considerably more variance than for the three-Neurotron controller. It was found that, in general, the three-Neurotron controller provided nearly comparable performance improvement, and was more consistent in operation.

The results presented herein form part of a series of practical endorsements of the Neurotron concept, which offers the advantage of faster and more efficient control of complex systems. Similar conclusions were reached in a program investigating the application of this concept for flight control. The accumulation of further data from practical supporting investigations is encouraged by these results.

REFERENCES

1. "Specification Brochure No. 60A," Solar Systems, Incorporated, Skokie, Illinois.
2. "Hermetically Sealed Nickel-Cadmium Batteries for the Orbiting Astronomical Observatory Satellite," R. Shair and W. Gray, presented at the ARS Space Power Systems Conference, Santa Monica, California, September 1962.
3. "Eveready Battery Applications and Engineering Data," Union Carbide Corporation, New York, New York.
4–9. "Theory of Probability State Variable Systems," ASD-TDR-63-664, Adaptronics, Incorporated, Volumes 1 through 6, respectively, December 1963.

DONALD R. TAYLOR, JR.*

Philco Corporation

Communications and Electronics Division

Willow Grove, Pa.

A Bioelectric Pattern Recognition Control for Prosthesis

ABSTRACT

A powered myoelectrically controlled prosthetic arm has been demonstrated which utilizes pattern recognition techniques to control the motions of the artificial limb on the basis of the myoelectric signals detected at multiple sites on the surface of the skin. Its principal significance is the demonstration of the feasibility of using integrated circuit and pattern recognition techniques for processing the myoelectric signals to achieve simultaneous multi-axis control of an externally powered device.

INTRODUCTION

Studies in pattern recognition have been extensively applied in many areas but most notably in the visual recognition field. That these techniques might be employed with great utility to analyze, process, and harness surface myoelectric signals for function control was pointed out some two years ago by W. L. Wasserman in papers presented here and abroad.[1,2,3] This has been supported and confirmed in subsequent basic investigations of electromyographic [EMG) phenomena conducted under the auspices of the Office of Naval Research. Work reported by Finley, Wirta, and others[4,5,6,7] has described patterns of myoelectric activity associated with specific body motions and their correlations with torque and displacement.[8] Development of specific instumentation to obtain numerical values to describe the signals has led quite directly to computer analysis and complete confirmation of the original thesis.[9] Establishing the feasibility of external power control has promoted the support of engineering application studies directed toward prosthetics and has been carried out by Philco for Temple University on Vocational Rehabilitation

* Now at Temple University — Moss Hospital, Biomedical Engineering Center, Philadelphia, Pennsylvania

Administration funding. The result was the demonstration of feasibility with an experimental model in which the functions of elbow flexion and terminal device rotation was implemented.

TECHNICAL BACKGROUND

The Myocoder, which was developed and built by Philco for the specific purpose of obtaining quantified myoelectric values, has been described in several of the cited references. (See Fig. 1.) The equipment consists

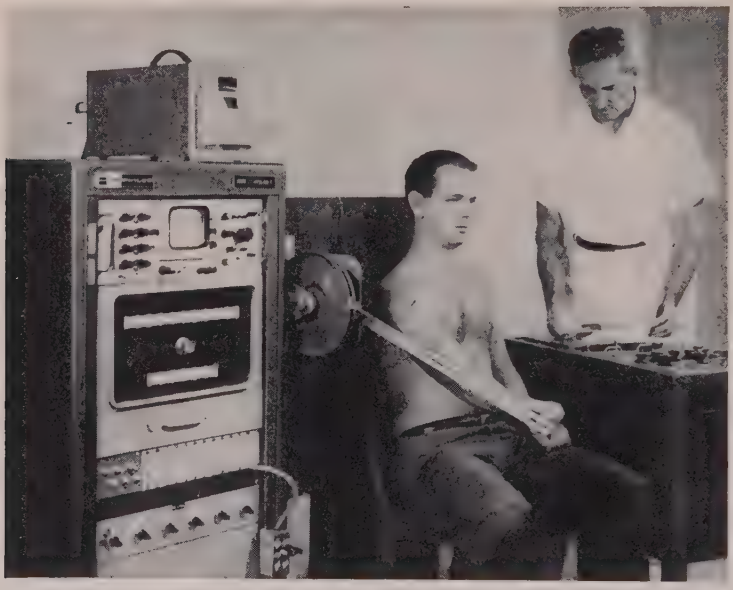


Fig. 1. The Myocoder renders myoelectric signals into numerical values for analysis.

of a bank of six high-gain amplifiers, rectifying and integrating circuits, analog to digital conversion circuits for the resultant signal parameters for all six channels, and a monitoring oscilloscope of the storage type for ease in superficial observation. A paper printout of two decimal digits for each of the six channels, and control timing of the repetitive sampling and integration period, completed the equipment complement. Although the relatively simple signal processing and conversion function tends to underemphasize the importance of the equipment, the Myocoder has

made it possible to render large quantities of difficult data into a form tractable for subsequent data reduction and manipulation.

A somewhat simplified version of the "Multinorm"[10,5] computer program developed for use in visual recognition work has been used to perform linear and quadratic discriminant analysis on myoelectric patterns observed during execution of physical body motions[8] and has shown an ability to provide excellent separation with only the linear portion operating when five to ten time frames of myoelectric data were supplied during an entire task execution. For the purposes of the experimental model design, which must operate in real time and not after the fact, a single time frame of six signals had to be used. Results of computer testing of the six variable linear discriminant functions operating on design and new data samples yielded imperfect but impressive enough results to justify their incorporation into a demonstration device.

EXPERIMENTAL CONTROL CIRCUIT

By utilizing integrated circuit amplifier assemblies and cordwood component placement, the complete electronics package, with the exception of the electrode assemblies, was fitted into a space measuring one by two by three inches—about the size of a king-sized package of cigarettes. (See Fig. 2.) Functionally the unit was divided into three compartments: EMG amplifier, recognition circuit, and threshold decision circuit. This unit, the electrode assemblies, and the motor control relays (mounted at the motors) comprised the complete system, as diagrammed in Figure 3. Power is supplied by a nickel-cadmium rechargeable battery.

1. Electrodes

Muscle sites for electrode monitoring were chosen to include musculature normally associated either directly or indirectly with the arm motions which were to be later identified. The anterior deltoid, posterior deltoid, long head of the biceps, short head of the biceps, lateral aspect of the triceps, and the pronator teres were used as signal pickup locations. In order that these sites could be easily reidentified, the motor point of each muscle was located by using a conventional muscle stimulator. The point on the skin immediately over the point of maximum response to electrical stimulation was marked and the electrode pair was fastened,

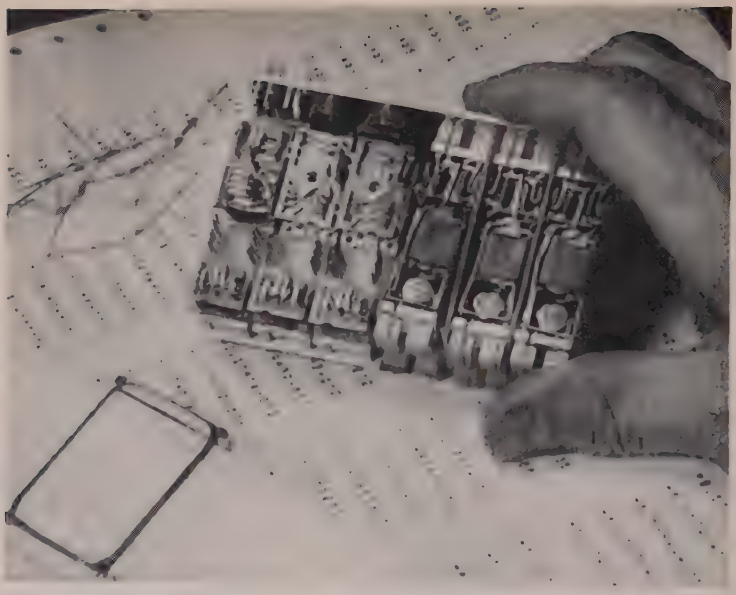


Fig. 2. The experimental control circuit package contains EMG amplifiers, signal processors and recognition circuits.

Fig. 3. The circuitry controlled a powered prosthetic arm from surface signals on the upper arm.

using double sided adhesive. Electrode pairs were located longitudinally with respect to the muscle and straddling the motor point with a separation of two centimeters.

The double sided adhesive material used for fastening the shielded silver disc electrodes was about 1/16 inch thick. Contact with the skin was made through a confined volume of electrode paste in a 1/4 inch hole punched in the tape under each electrode. Reliable contact was achieved with this arrangement.

Figure 4 is a photograph of the shielded electrode assemblies. These appear center and right. On the left is an integrated circuit amplifier addition used with the electrodes in another study.

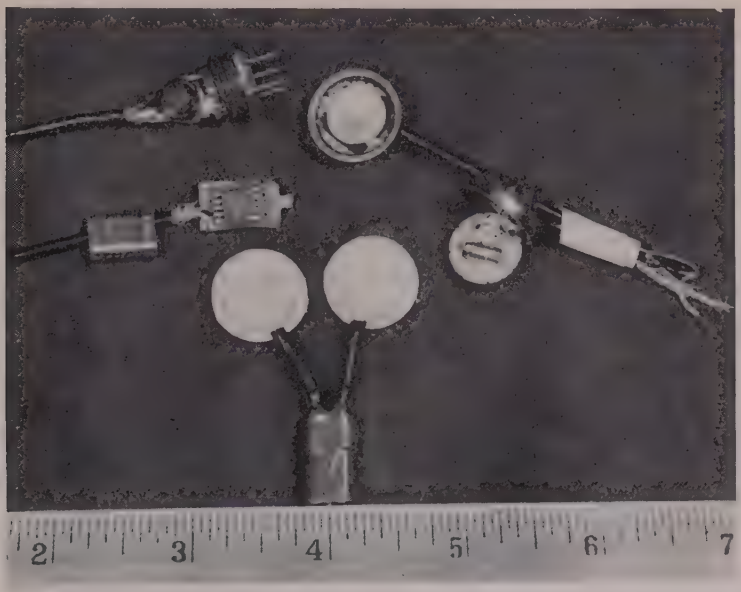


Fig. 4. Electrode assemblies were specially constructed (center and right) for low artifact.

2. Myoelectric Amplifiers

Average signal levels ranged from about ten microvolts to over a millivolt, depending on the particular muscle being monitored and the subject's physiological characteristics. The necessary gain to raise these signals to a respectable level for further processing was obtained in an

amplifier manifold consisting of six identical miniature assemblies measuring $\frac{1}{2}$ by $\frac{1}{2}$ by 2 inches. (Ref. Fig. 2.) The electrodes previously described, were used in pairs (i.e., a pair for each muscle site), to minimize power line interference. Each amplifier assembly, then consisted of a differential input stage followed by a differential to single ended transition and a final single ended output stage. These operations were performed by three integrated circuit amplifiers connected in cascade to give a gain which can be set by a trimming potentiometer to a maximum of about 100,000. The frequency limitation was set at 5,000 cycles on the high end and about 100 cycles on the low end. This adequately matched the conditions used in the data taking where the response was 30 cycles to 5,000. Compromise on the low end seemed not to have caused difficulty and tended to suppress residual power line pickup and to reduce the size of coupling capacitors used in the amplifier interconnections. Dual rectifiers and integrators were contained in each channel to provide both positive and negative smoothed signals as outputs to the recognition circuits. A 50 millisecond time constant was used in the filter to approximately match the integration time used in the data taking process with the Myocoder.

3. Recognition and Decision Circuits

The linear weighting values from the computer printout were realized physically for each class of motion by an array of six resistors which provided the proper conductance from the outputs of each of the six myoelectric channels to a common summing point. Positive or negative weighting was obtained by connecting the particular resistors to the appropriate polarity from the dual outputs of the myoelectric amplifier integrators. The summed weighted outputs from each channel were then compared with a pre-set level to provide a dichotomous output indication when the threshold level was exceeded. Two additional integrated circuit units and an additional output transistor were used for these functions. The weighting network, threshold setting potentiometer and comparator circuits were grouped together on a replaceable plug-in subassembly.

4. Power Circuits

While the complete electronics as described was sufficient to demonstrate operation of the control simply by using the decision outputs to turn on indicator lights, it was felt that a much more dramatic demon-

stration would be to actually modify a standard prosthesis by adding two motors as actuators. The modifications were made by replacing the lock joint and pull wires at the elbow with a DC motor, associated gear train, and ball screw. The manually rotated terminal device was similarly equipped with a smaller permanent magnet motor.

The resultant electromechanical arm weighed about three pounds and developed 50 inch pounds of torque for elbow flexion and 5 inch pounds for rotation of the terminal device. Relays located at the motors were controlled by the electronic circuits to effect actuation.

RESULTS

Operation of the control unit and artificial arm satisfied quite adequately the physical demonstration of the control technique. Pronation and supination functions were very reliably controlled by the person from whom the design data was obtained. Elbow flexion was easily obtainable but often actuated in addition to supination when supination only was intended. This correlated well with the performance predicted

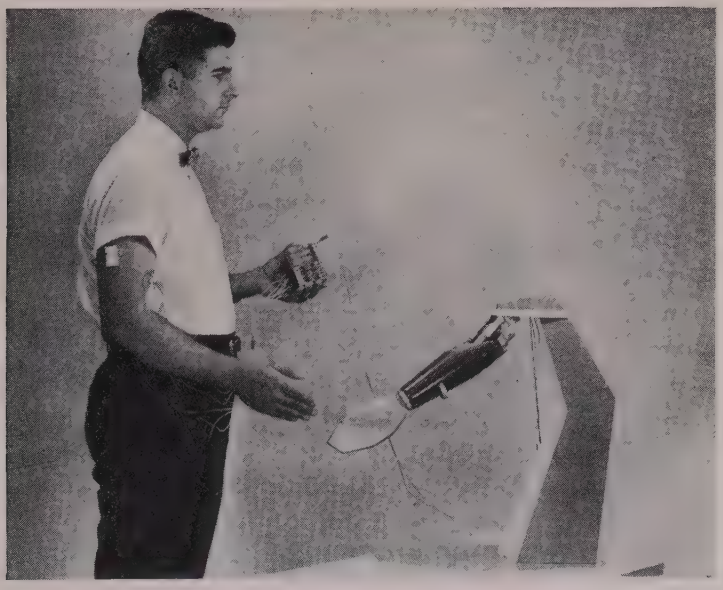


Fig. 5. The demonstration arm shown in elbow flexion.

by computer design runs which gave error rates of 3 per cent on both the pronation and supination functions but 10 per cent for flexion. A photograph of the arm in action is shown in Figure 5.

No particular difficulty in electrode adhesion or mechanical artifact was experienced during many removals and replacements. Although the electrode assemblies were worn for periods of eight hours a day for several days, no skin reactions were encountered.

Need for improvement in error rate, torque output, and motor noise level are most obvious. It is clear also that more practical electrode fastening arrangements, more efficient actuators, greater sophistication in signal processing, the addition of at least two more powered axes, reduction in size and weight, improved ruggedness and reliability, eventual sensory feedback, and proportional torque control are among the many factors which must be dealt with before a practical prosthesis application can be made.

Summary

An experimental myoelectric control using single layer discriminant networks was designed and fabricated to show the practical feasibility of the control technique. Using the device, a normal subject was able to control simultaneously two axes of motion in a powered prosthetic. Microelectronic components were utilized to attain a control unit size which was comfortably carried on the person. With greater sophistication in signal processing and further engineering effort, the technique promises to be extremely useful in many equipment control and human function amplification applications.

REFERENCES

1. Wassermann, Walter L., and Harrison, Lee III: The Application of Adaptive Pattern Recognition Techniques to Surface Electromyographic Phenomena. A paper delivered on September 26, 1964 to the International Ergonomics Association, Dortmund, Germany.
2. Wasserman, Walter L.: Electromyographic Function Control by Pattern Recognition. A paper delivered on February 13, 1965 to the American Institute of Aeronautics and Astronautics, Tampa, Florida.

3. Wasserman, Walter L. Human Amplifiers. A paper published in *International Science and Technology*, October 1964.
4. Finley, F. Ray. Electromyographic Patterns of Multiple Muscle Sources, presented to the Conference on Control of External Power in Upper Extremity Rehabilitation, Committee on Prosthetics Research and Development, National Research Council, April 8, 1965.
5. Finley, F. Ray, and Wirta, Roy W. Myocoder-Computer Study of Electromyographic Patterns, presented at the Congress of Physical Medicine and Rehabilitation held in Philadelphia, Penna. August 26, 1965.
6. Finley, F. Ray, and Wirta, Roy W. Myopotential Response and Force of Muscle Contraction as Analyzed by Myocoder-Computer Techniques, presented at the Congress of Physical Medicine and Rehabilitation, August 26, 1965.
7. Harrison, Lee III. Paralytic's Brain + Myocoder = Hope. *Electronics*, November 1964.
8. Myopotential Response and the Force of Muscle Contraction. Final Report. Contract Nonr 4292(00), Office of Naval Research, Psychological Sciences Division, Engineering Psychology Branch, Department of the Navy, Washington, D. C., November 1, 1965.
9. Harrison, Lee III. A Study to Investigate the Feasibility of Utilizing Electrical Potentials on the Surface of the Skin for Control Functions. Proceedings of the 1964 Seminars on Remotely Operated Special Equipment (Project ROSE), May 26–27, 1964, pp. 100–157. CONF 64508, Vol. I. USAEC Division of Technical Information.
10. Harley, T., *et al.* Semi-Automatic Imagery Screening Research Study and Experimental Investigation. ASTIA 410261 March 29, 1963.

Index of Names

Page numbers in bold type indicate contributions to this volume

Printed in Germany