Reprinted from

PROGRESS IN BRAIN RESEARCH

VOLUME 17

CYBERNETICS OF THE NERVOUS SYSTEM

Edited by

N. WIENER

and

J. P. SCHADÉ

ELSEVIER PUBLISHING COMPANY

AMSTERDAM–LONDON–NEW YORK

1965
On Cybernetics, Information Processing, and Thinking

M. E. MARON*

The RAND Corporation, Santa Monica, Calif. (U.S.A.)

INTRODUCTION

Norbert Wiener's great intellectual capacities erupted at an early age to make him a child prodigy, but unlike many child prodigies, especially those with a flair for mathematics, Wiener's intellectual powers neither faded nor froze and he emerged as one of our great applied mathematicians. We honor not only an outstanding mathematician, but also the herald of a revolution in technology and science, the influence of which has been felt in such areas as automation, education, and economics, with possible reverberations shaking the structure of society itself. However, the full impact and implications of the revolution of cybernetics, with which his name is associated, will not be recognized properly for some years to come, which is the way with most intellectual revolutions. Norbert Wiener has changed our way of looking at and grappling with problems. What is especially significant is that the scope of his revolution includes some of man's basic problems, perhaps the most important of which is the problem of the brain, how men think, and how to understand the activity of knowing.

1. ORIGINS AND SCOPE OF CYBERNETICS

The background

What is cybernetics? Perhaps the best answer has been provided by Norbert Wiener himself in the introduction to Cybernetics (Wiener, 1948), his classic work on the subject. Wiener relates how the ideas for cybernetics grew out of a seminar at Harvard on the methodology of science. The participants in the seminar were researchers from diverse fields, whose participation reflected their common belief that new and exciting areas for scientific exploration lay in the borderlands between established fields— in those areas where the problems could be attacked successfully only by an interdisciplinary group of scientists representing different fields of specialization.

* The views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors.
With the onset of World War II, Prof. Wiener found himself involved in just such a multidisciplinary problem, viz., the problem of designing an automatic system to feed aircraft-tracking radar data into a computing machine which, in turn, would predict the future course of the aircraft being tracked and then generate and transmit control signals to aim the guns at the moving target. Thus, Prof. Wiener, together with Julian Bigelow, became immersed in a set of new and complex problems relating to the theory of prediction and the design of machines that could embody the theory. In the course of this study they came to the very important realization of the critical role of feedback in voluntary human motor activity. Prof. Wiener and Julian Bigelow conjectured that the human mechanisms which control motor activity could suffer from improper feedback and show the same ‘symptoms’ as known mechanical systems, and that the mechanisms for manipulating feedback information in automatic machines are similar to those in that part of the brain where motor activity is controlled.* If this were so, both systems should malfunction in similar ways, and apparently they did.

With this insight came the recognition that the problems of communication and of control are intricately bound together. An analysis of control problems involves analysis of the underlying communication problems, and the latter, in turn, center around the fundamental concept of the message. With the problem of analyzing messages comes the problem of noise, and this led Wiener to questions of the relationship between amount of information and entropy, and their roles as measures of degree of organization and disorganization of a system.

For the historian of cybernetics, the following is a key passage:

"...as far back as four years ago, the group of scientists about Dr. Rosenblueth and myself had already become aware of the essential unity of the set of problems centering about communication, control, and statistical mechanics, whether in the machine or in living tissue. On the other hand, we were seriously hampered by the lack of unity of the literature concerning these problems, and by the absence of any common terminology, or even of a single name for the field. After much consideration, we have come to the conclusion that all of the existing terminology has too heavy a bias to one side or another to serve the future development of the field as well as it should; and as happens so often to scientists, we have been forced to coin at least one artificial neo-Greek expression to fill the gap. We have decided to call the entire field of control and communication theory, whether by machine or in the animal, by the name Cybernetics ...." (Wiener, 1948, p. 19).

In that splendid personal account describing the circumstances which gave rise to cybernetics, Prof. Wiener goes on to describe the relationships of other subjects to the broad field of cybernetics, e.g. neurophysiology and information processing in the cerebral cortex, mathematical logic and its application to the analysis of neural switching circuits, computers and their use in constructing models of brain mech-

---

* A joint paper discussing feedback and purposeful behavior was published in 1943 (Rosenblueth et al., 1943).

References p. 136
Thus, we find in his book the groundwork and beginnings of a broad interdisciplinary study of all phases of information processing and control systems and of the implications and ramifications of this for such subjects as the psychology of perception, psychiatry and memory disorders, linguistics, and sociology. Since cybernetics erupted on the scientific scene in 1948, much work has been carried out in the field. Let us look briefly at some of those more recent developments which are relevant to an information-flow study of brain mechanisms and of artificial mechanisms designed to exhibit intelligent behavior.

**Subsequent developments**

**Theory of information**

If cybernetics is concerned with those sciences, studies, techniques, and mechanisms which relate to communication and control, the concept of information must play a central role. What sort of a commodity is information? What are its properties and what would we expect to find in a *theory* of information? Messages, of course, have many properties — their information may be precise, complex, redundant, timely, improbable, effective, etc. It is reasonable to assume that we need a different measure for determining the 'amount' of each different property in a message. In 1948, Claude Shannon, a mathematician and former student of Prof. Wiener, published a theory of communication (Shannon, 1948; Shannon and Weaver, 1949), which for the first time quantitatively explicated one of these measures. Shannon's theory allows one to define precisely the notion of the capacity of a communication channel, as well as the capacity of an information store, so that it is possible to answer the question: How much more information can be conveyed over one communication system than over another? Shannon's theory provides a means of determining the *amount* of information generated by an information source, real or artificial. The important notion of a code for conveying messages is treated in a formal way so that one can talk of the efficiency of a given coding system. And one can decide whether or not a given coding system (relative to a given information source) is the best possible one. Shannon's theory clarified the key concept of *noise* and showed how *redundancy* can be introduced in coding schemes so as to combat the destructive effects of noise on messages. And finally, he showed some very interesting and surprising relationships between noise, redundancy, channel capacity, and error-free transmission of messages.

**Computer technology and applications**

Since the digital computer is a mechanism for the physical (and logical) processing of information, computer theory and technology occupy a sizable part of the cybernetic stage. The computer field includes not only digital technology and theory of computer design and organization, but also includes the facets of computer applications. Following World War II, and beginning just before the publication of *Cybernetics* and Shannon's *Mathematical Theory of Communication*, there started an almost explosive development of digital computer technology. Some early digital machines were developed secretly during the war, but once the general knowledge of
computers and their components and their potentials became known, there was a race to develop new switching components and circuitry and new types of information storage media. More recently, we have seen the development and refinement of very-high-speed circuits and switching components and comparable gains have been made in the perfection of high-density, rapid-access memory media. Finally, associated with these developments, there has appeared a host of sophisticated input–output (and analog–digital conversion) equipment to aid in the task of communicating to and from a computer.

While the component and systems engineers were pushing rapidly ahead, their more theoretically oriented colleagues were working at the corresponding theory of information machines — including automata theory and the theory of Turing machines. The theory of machines includes switching theory and what might be called the art of computer design and organization. These theoretical probings have led to the beginnings of a theory of reliability dealing with the question of how best to organize a large number of unreliable switching components to do a particular job so that the system as a whole is extremely reliable. Other theoretical aspects of computer organization have to do with programming theory, development of new computer languages, and the theory of algorithms.

Whereas initially the computer was looked on primarily as a machine designed to execute complex mathematical computations at high speed, we have since come to recognize the universality and flexibility of computers as information machines. It is this great universality that makes the modern computer a general-purpose research vehicle — a tool giving scientists the ability to perform arbitrary experiments at high speed. One of the more significant uses of computers now, and increasingly so in the future, lies in the fact that experiments can be executed by machines through simulation. Subsequently, we shall describe how a computer can be used to aid both psychologists and neurophysiologists by simulating aspects of the structure of the brain and the behavior of human beings. A clarification of the use of computing machines in these types of applications reveals the impact of cybernetics not only on psychology and neurophysiology, but also on the theory of knowledge itself.

2. INFORMATION PROCESSING IN ARTIFICIAL AND NATURAL SYSTEMS

The brain as a computing machine

Since the publication of Wiener's Cybernetics, comparisons have been made between the human brain and electronic computing machines on at least three different levels: components, coding and logical organization, and information processing. As the concepts of information theory, the computer, automata, and control theory have become developed and refined, attempts have been made to apply these notions to describe and interpret information processing in biological systems (McCulloch, 1949; Ashby, 1952; Sluckin, 1954; Von Neumann, 1958; Wooldridge, 1963).

When making such comparisons, the brain is viewed as an information transforming device which accepts data from the external world, interprets and operates on these

References p. 136
data, and produces control information. In the jargon of the computer engineer, the brain is a black box which receives input signals via the receptors from the various sensory organs, and operates on these signals to generate other output signals which finally go to effectors causing appropriate muscles and glands to function. The information loop is closed via the external world and thus the brain is interpreted as a complex automatic regulating device which allows its owner to react so as to maintain a certain stability of success in his interaction with the local environment. Given this representation of the brain as a complex computer, attempts have been made to describe the information-flow paths and the logical organization of its major subsystems.

In the brain the neuron is usually considered as the basic information processing unit. One of the first formal models of a neuron was described by McCulloch and Pitts in 1943 (McCulloch and Pitts, 1943), when they argued that the all-or-none principle of neuronal activity allowed one to interpret the behavior of a nerve cell in terms of truth-functional logic. Thus, just as Shannon in 1938 had applied truth-functional logic to the analysis and synthesis of switching circuits (Shannon, 1938), McCulloch and Pitts interpreted the behavior of neurons as two-state switching devices. We now know that the behavior of nerve cells is more complex than that implied by the early McCulloch–Pitts model, and even the most elaborate of recent models suggested by computer technology for neurons account for only a fraction of the richness and complexity of actual neuronal behavior.

One of the major simplifying assumptions of the McCulloch–Pitts model was that the pulses in a neural net are synchronized — i.e. that the action spikes generated and propagated by individual nerve cells correspond to those in a binary computer where each gate is pulsed by a signal from the central timing clock. This means that the neuron can fire (if at all) only at regular discrete intervals, and this interpretation leads to a so-called 'binary quantized' coding system. However, other coding schemes have been suggested, e.g. pulse interval modulation. Using different coding schemes (MacKay and McCulloch, 1952) one can compare the theoretical channel capacity of biological 'cables'. The problem of discovering the coding schemes used by the various sensory channels has attracted attention, and with information analyses of sensory and other channels there has been work on interpreting the various biological input–output and conversion mechanisms in terms of computer counterparts.

There is also the question of memory and how past events are coded and recorded in the brain. Estimates have been made by Von Neumann (Von Neumann, 1958) and others as to the total information capacity of the brain. Various types of experiments (for example, Penfield, 1958) suggest that a very large proportion of our experiences are coded, recorded, and under certain conditions can be subject to recall. It has been proposed that information may be stored by means of many different kinds of mechanisms ranging from circulating loops, analogous to the delay-line memories of computer technology, to variable thresholds, modifiable synaptic weights (Maron, 1961), variable time delay mechanisms (MacKay, 1962a), DNA memory mechanisms, and neural glia (Schmitt, 1962). Each of these has its own properties of permanence and access speed. We have no lack of proposals for biological memory mechanisms,
but we have yet to see an adequate theory of recall—on the mechanism level. Other organizational comparisons have been made (Von Neumann, 1956, 1958), showing how the brain might embody redundancy in the form of circuits and codes, how computing in the nervous system is organized so as to balance precision, logical depth, reliability, etc. Finally, numerous feedback and control mechanisms have been suggested as models to account for the automatic internal control of breathing, heart action, kidney function, blood sugar concentration, body temperature control, appetite, etc.

Thus, on the level of components, codes, and logical organization of underlying structure, the concepts and language of information theory and cybernetics have been used by some in an attempt to analyze brain mechanisms in terms of artificial information processing mechanisms. But, at this level of comparison between the brain and the computer, the notions have yet to suggest any really deep neurophysiological hypotheses. It is one thing to understand the logic of neural firing, the coding of basic cortical messages, or the mechanisms of storage and recall in the brain, and quite another problem to relate these mechanisms to complex intelligent behavior.

The other level of comparison between the brain and a computer relates to their functional similarities. Both are information transforming mechanisms which accept input data and, as a function of previously stored data, derive output conclusions. They both process information in order to derive outputs of a certain kind. There is a rationale to the input–output relation in both the brain of a normal person and a computer with a meaningful program. This functional similarity suggests that the logic of information processing and automata theory applied at first solely to computers might be applied also to the analysis of biological information machines and that it might help to bridge the gap between the psychology of behavior and the physiology of those mechanisms which produce it. MacKay has described this conceptual bridge as follows:

"Now the language of information and control, in which the theory of automata is framed, is conceptually intermediate between those of psychology and physiology. It belongs in a sense to both fields, inasmuch as questions of information transfer arise in both. It seems possible therefore that the theory of an automaton, specifically designed to be a research model of the human organism considered as an information-flow system might provide a common meeting-ground and a common language in which hypotheses might be framed and tested and progressively refined with the help of clues from both sides." (MacKay, 1956, p. 31).

What is the logic of the relationship between behavior and the information processing mechanism which produces it?

The logic of information processing

There are three separate but related problems that must be clarified in order to understand the sense in which brain–computer comparisons are not merely interesting, but potentially valuable. On the one hand, one may look inside the brain, so to speak,

References p. 136
and compare its structure and the organization of its structure with that of a computing machine. On the other hand, it is possible to treat each as a black box which cannot be opened, and compare the kinds of input–output relationships that are possible in the brain and a computer. Finally, there is the question of how to relate input and output behavior with the internal mechanisms which produce them.

In the case of a human we ordinarily talk of observing his behavior — *i.e.* the relationship between stimuli and responses — but we might equally well move upstream and consider the input and output signal patterns which correspond to stimuli and responses. We can consider signal patterns in an abstract way as expressions of a logical language. And we will find it most convenient to clarify the relationships between input and output expressions if we assume that they belong to the language of elementary truth–functional logic.

In what ways might input and output expressions be related? Consider the case where we know the truth relationships between input and output for all possible values. That is, for every allowable truth value of the input expression, we know the corresponding truth value of the outputs. It is now possible to formulate a description of this relationship in terms of a complex sentence of truth–functional logic. We emphasize that there is a mechanical procedure for describing the truth relationship in terms of the grammar represented by the logical connectives, such as negation, conjunction, disjunction, etc. In fact, it is possible to translate any expression into an equivalent one which contains only occurrences of a single type of connective, *viz.*, the Sheffer stroke function or the Peirce function, both of which are *universal*.

Turing first pointed out (Turing, 1938) that it is convenient to clarify problems in logic by talking in terms of the mechanisms for implementing logical operations. We know that if we can construct the physical counterpart of the elementary operations of truth–functional logic in the form of, say, electrical switching circuits, then it is possible to build the physical counterpart of the logical expression which describes the truth relationship between input and output expressions. (Note that one may wish to base the truth value of a conclusion not only on current premises, but on past history as well. There are concepts in automata theory (see, for example, McNaughton, 1961), such as the *state* of an automaton, which permit past input history to be dealt with in a precise way so that the output is a function not only of current input values but also of the state.)

We have traced the transition from the statement of an input–output relationship to its description as an expression in truth–functional logic, and to its physical implementation as an automaton. This automaton may, in turn, be thought of as a special-purpose computer tailored to solve automatically the particular problem in question. But, all problems that are solved by the use of any special-purpose computer can also be solved on a universal general-purpose, stored-program, digital computer, whose essential characteristic is that it is able to interpret and execute a set of stored instructions (which are themselves reducible to elementary logical operations) and modify its course of computations during the execution of a problem. These remarks are intended to explain and justify the following statement of Ashby: “Name the behavior, and the logic of mechanisms leads inexorably to a program that with the
machine forms a whole mechanism that will show the behavior" (Ashby, 1962, p. 306). If one can give a complete and unambiguous description of behavior, this description is, in a sense, equivalent to a blueprint of an information processing mechanism (a computer program) for producing it. This is one aspect of the relationship between behavior and the underlying mechanisms for generating it. But it leads us to ask the following questions:

(a) To what extent can complex behavior be specified in complete and unambiguous detail, eventually in the language of logic?
(b) If there exist the equivalent of stored computer programs in the brain (Wool-dridge, 1963, p. 62) which control complex purposeful behavior, how do they originate? How does the logical chain (i.e. the program) get modified and corrected with experience? How do we account for learning via feedback from the environment?
(c) What is the logic, if any, of this process of discovery and could it be embodied in an artifact which could exhibit learning and intelligent behavior?

This final question takes us away from neurophysiology to the problem of artificial intelligence.

3. THE PROBLEM OF ARTIFICIAL INTELLIGENCE

Initial characterization

The goal of artificial intelligence is to find design principles for an intelligent, thinking artifact — i.e. an artificially constructed entity which can solve new and different problems, learn from experience, respond appropriately under a variety of different circumstances, and, thus, behave in ways that justify our calling it intelligent. This initial, rough statement of the problem of artificial intelligence appears to encompass all that is relevant to the problem of knowledge itself: how we think and plan, value and behave. This being the case, how does one distinguish between psychology, neurophysiology, artificial intelligence, and related disciplines?

Psychology is alleged to be the study of how we think and learn and behave, and its subject matter is different from that of artificial intelligence in that psychology deals with the behavior of the actual biological artifact and not with all possible intelligent entities. Also, its subject is more limited in that most psychologists would argue that it deals with behavior and not with the underlying mechanisms which mediate between perception and behavior.

Neurophysiology, of course, is the study of those underlying signal processing mechanisms (viz., the brain, the nervous system, etc.) which enable intelligent behavior. And since its subject matter concerns the actual biological system, neurophysiology is indeed relevant to artificial intelligence but, of course, not identical with it. We might put the relationship between these two pursuits in the following way: If one had a complete neurophysiological theory that spelled out in detail how the human brain is organized to process information so as to allow its owner to think and act intelligently, then, in principle, one could use that theory to design an intelligent artifact composed of artificial (non-protein) components, assuming, of course, that

References p. 136
whatever, if anything, is informationally relevant about protein could be embodied in the artificial counterpart. Notice, however, the converse relationship: If we did have design principles for an intelligent artifact, these principles might be radically different from those incorporated in the human mechanism. Therefore, neurophysiology and artificial intelligence are distinct studies, but are related in the sense that an understanding of the functioning of the former contributes to that of the latter, although the converse does not necessarily hold. It is because the study of artificial intelligence is not unrelated to either psychology or neurophysiology that workers have looked to these and other traditional disciplines for insights into the problem of designing an intelligent artifact.

Methods of attack

How might one proceed to construct a device whose behavior would resemble that of a thinking human? What clues might guide efforts to simulate human problem solving behavior, learning, etc.? At least two different kinds of approach are possible; the difference between them becomes more clear in terms of a distinction, made by Donald MacKay (MacKay, 1963b), between the forging of a product and the forging of its generator. The product corresponds to human behavior and its generator corresponds to the information-flow mechanisms which causally determine behavior. If one wanted to imitate the sound of a motor engine, for example, he could observe in minute detail the pressure waveform of the sound, and seek to reproduce it by means of some form of curve-following device driving a hi-fi loudspeaker. He might continue to observe and correct his imitation by matching it against the genuine output and altering details of the curve being followed so as to approximate more and more closely to the original. However, he might proceed in a quite different way, by trying to build a model of the engine which would become a generator of the desired output product. This difference in approach is exemplified in the research conducted so far in artificial intelligence.

Although the work of Newell, Shaw, and Simon was motivated primarily by an interest in the development of a psychological theory, their early work on the Logic Theorist (Newell et al., 1957) represented a step in the direction of artificial intelligence. They attempted to simulate one type of human problem solving behavior by programming a digital computer to derive theorems of elementary logic. Although there is a decision procedure for truth-functional logic, they ignored that procedure and programmed the computer to use methods allegedly similar to those used by humans. They gave the computer the Whitehead–Russell axioms for truth-functional logic, a set of transformation rules (such as modus ponens, substitution, etc.), and a set of techniques, called ‘heuristics’, on how to sequence transformation rules to obtain a proof of a given expression. The heuristics or advice was to correspond with the set of tricks and techniques, not always conscious or explicit, such as working backwards, using a reductio-type proof, knowing when to give up on one line of approach and seek out another alternative, etc., which a human apparently uses. Newell, Shaw, and Simon attempted to clarify and elaborate these techniques and advice
so as to make them more precise in the form of a routine which the machine could interpret and execute. Such a computer routine, called a 'heuristic program', would differ from conventional programs in that it might never solve a given problem. Instead it would (hopefully) simulate the problem solving behavior of a human in the sense that it would employ similar sorts of problem solving techniques and methods. It would grope, use trial and error, set up goals and sub-goals, etc.

The adequacy of a problem solving program is tested by matching it against a human subject in the following way: As the human solves a problem he is asked to describe his 'thoughts' as he gropes toward the solution. This report is tape-recorded and typed out for visual inspection. Likewise, the machine is programmed to print out periodically intermediate steps as it proceeds to solve the same problem. Then, these two reports are matched against one another to determine how closely they agree. Where they do not agree, an attempt is made to modify the computer program so that on subsequent problems the outputs will more closely approximate the method employed by the human. Here then is one example of the several attempts to get a machine to simulate human problem solving behavior by forging the product.

If one were to take the other approach, he would attempt to forge the generator by modeling the informationally relevant aspects of the brain and nervous system. And, in fact, during the past decade, several groups have been studying so-called 'brain models' by simulating on digital computers various aspects of biological information-flow mechanisms. The strategy has been to ignore some aspects of neuronal behavior*, simplify other aspects, and then introduce assumptions or hypotheses about how cells are modified with 'experience'. The goal is to examine the consequences of these assumptions to see whether or not they can account for pattern recognition, learning, etc. It turns out that if one tries to analyze assumptions about large numbers of interconnected neuron-like elements, the complexity of the situation makes any kind of an analytic understanding practically impossible. Complexity rules out the usual use of standard logical and mathematical tools of analysis. One needs to see the model in action, so to speak, and the large-scale computer is just the research instrument which allows it to be so examined. The description of a neural network can be translated into a computer routine which allows the machine to simulate the behavior of the network in question. Computer simulation allows the researcher to modify his assumptions by merely changing the values of certain parameters in the routine, which in turn cause the simulated network to exhibit different behavior. Thus, the researcher can formulate and examine a variety of hypotheses about neurons and try to find rules which will enable networks to generate interesting behavioral properties. Here then are two different ways of studying the problem of artificial intelligence, both of which depend heavily on the digital computer in its role as a general-purpose simulator.

* It should be pointed out, of course, that the so-called neuron doctrine which asserts that information processing is executed exclusively by nerve cells, action spikes, etc., is not free from controversy. See, for example, Bullock, 1959; Galambos, 1961; and Hyden, 1962.

References p. 136
Thinking vs. brain activity

In what sense does the simulation of complex neuron-like nets constitute a reasonable attempt to forge the generator of human behavior? The actual brain mechanism of the human is, of course, exceedingly more complex than any net so far simulated, not only in terms of the numbers of elements but also in terms of the behavioral richness of the individual cells. But even more crucial than these differences is the fact that network studies start with the behavior of cell-like elements and proceed to synthesize nets of such elements which display interesting properties, whereas a more fundamental approach would be to look first at the necessary information-flow requirements for any mechanism which is to underlie the intelligent behavior. It is to this question of the logical organization of a generator that we shall return subsequently. For now we ask the following question about those other attempts — such as automatic theorem proving, game playing (Newell and Simon, 1961), etc. — which allege that the computer simulates human thinking: In what sense is it legitimate to say that automatic problem solving on a computer corresponds to human thinking? Is it even sensible (as opposed to true or false) to assert that the machine is thinking when it executes its stored program and subsequently prints out the correct results (along with the methods used in obtaining those results)? Does it not seem, on the surface at least, that what goes on in the ‘mind’ of a thinking human is not even remotely similar to what goes on in a digital computer when it sequentially interprets and executes those routines which lead from problem to solution?

These questions, of course, bring up the age-old problem of the relationship between the brain and the mind: the problem of relating mechanical, physical processes, such as nerve signals on axons, cell firing, etc., to such mental activities as having ideas, thoughts, sensations, feelings, etc. The mystery of the relationship between the mind and the brain has been expressed by Sherrington as follows:

“When I turn my gaze skyward I see the flattened dome of the sky and the sun’s brilliant disc and a hundred other visible things underneath it. What are the steps which bring this about? A pencil of light from the sun enters the eye and is focused there on the retina. It gives rise to a change, which in turn travels to the nerve-layer at the top of the brain. The whole chain of these events, from the sun to the top of my brain, is physical. Each step is an electrical reaction. But now there succeeds a change wholly unlike any which led up to it, and wholly inexplicable by us. A visual scene presents itself to the mind; I see the dome of sky and the sun in it, and a hundred other visual things beside. In fact, I perceive a picture of the world around me. When this visual scene appears I ought, I suppose, to feel startled; but I am too accustomed to feel even surprised.” (Sherrington, 1950, p. 3).

Implicit in this description is what Ryle (Ryle, 1949, Chap. 1) calls the official doctrine about the nature and place of minds. Briefly, this doctrine asserts that every normal person has a body (including his brain) and also a mind. Bodies and brains have mass, occupy space, obey the physical (natural) laws of mechanics, chemistry, etc., and their behavior is, in principle, public and can be studied by qualified observers. Minds, on
the other hand, are not physical entities, do not occupy space, are not subject to mechanical laws, and their 'workings' can be observed only by the person whose mind it is. Given this characterization of the two entities, 'minds' and 'brain-bodies', the question is how can these influence one another. Ryle states:

"What the mind wills, the legs, arms and tongue execute; what affects the ear and the eye has something to do with what the mind perceives; grimaces and smiles betray the mind's moods and bodily castigations lead, it is hoped, to moral improvement. But the actual transactions between the episodes of the private history and those of the public history remain mysterious, since by definition they can belong to neither series. They could not be reported among the happenings described in some one else's biography of that person's overt career. They can be inspected neither by introspection nor by laboratory experiment. They are theoretical shuttlecocks which are forever being bandied from the physiologists back to the psychologist and from the psychologist back to the physiologist." (Ryle, 1949, p. 12).

We cannot consider here the classic paradoxes and difficulties surrounding the mind–brain problem. The literature on this subject is enormous* and its roots can be traced back to the writings of Plato, although Ryle credits Descartes with the 'official' view described above. Attempts to come to grips with the problem of the relationship between the body and the mind have been made by physicists (see, for example, Schrodinger, 1958) and physiologists** as well as philosophers. And the concepts of cybernetics, which have renewed and stimulated analysis of the prospect of designing an intelligent artifact, have also aroused considerable interest in the mind–brain problem.

The notion of complementary descriptions

Perhaps, the difficulty in explaining the mind–brain interaction problem arises from an inadequacy in our theory of explanation. Perhaps, our experiences concerning thinking and brain activity cannot be integrated into a single, complete, unified description formulated in a single language. This is not to say that there are two kinds of reality, but rather that there are (at least) two kinds of descriptions of the one reality. Niels Bohr (1958), and others, have suggested that a solution to the mind–brain problem has its analogue in quantum mechanics where the impossibility of deciding between a particle or wave theory of matter has led to the notion of complementary descriptions. Arguing from what might be considered a more general notion of complementarity, Donald MacKay (1958), has suggested that one can, in principle, describe the activities of knowing in two quite different, but complementary, languages: one language expresses the process from the standpoint of the 'actor' who is doing the thinking — having the mental experiences; the other language expresses the

---

* For an indication of some fairly recent discussions see, for example, Hook, 1960, and Scher, 1962.
** See Eccles, 1953; also see the discussions by Adrian, Le Gros Clark, Zuckerman, Slater, Brain, and Penfield in Laslett, 1950.

References p. 136
situation from the standpoint of the 'spectator' — i.e. the external observer who can make behavioral observations and who, in principle, has evidence concerning the underlying neurophysiological events, for example, via implanted micro-electrodes, etc. The important point is that descriptions formulated in these two different languages are not in conflict. Neither is necessarily incomplete in its own terms but at the same time neither is exclusive of the other. MacKay has written:

“The false dualism which used to be expressed in the question 'how can matter produce mind' would now seem to have its origin in the genuine dualism of conceptual frames of reference, defined respectively for the viewpoint of actor and spectator.” (MacKay, 1951, p. 118).

Let us clarify briefly the implications of this way of looking at brain activity and thought processes. People, real or artificial, who think, feel, have sensations, etc., can describe these activities in actor language. Correlated with the actor's description of knowing, there are the physical and behavioral indications which can be studied by an external observer and formulated in spectator language. This implies that the designer of an intelligent artifact need not worry about how to bring in some 'extra mind-stuff' in order to carry out his job. Rather, his task is first to characterize the behavioral indications of thinking, knowing, perceiving, understanding, etc. then to determine the logical organization of the information-flow mechanisms in an artifact which could generate the behavior in question, not ad hoc, but as a natural consequence of the identity of the logical organization.

The generator as a prediction machine

How must an artifact be internally organized, from an information-flow point of view, in order to generate the type of behavior which would be compatible with an 'outsider's' description of an intelligent, thinking artifact? What are the essential, logical (information processing) ingredients required in the design of a generator of intelligent behavior? We would argue that a key characteristic of intelligent, understanding-like behavior is the ability to make reasonable predictions. The intelligent person and the intelligent artifact must be able to predict in order to behave appropriately. The predictions in question cover the range of the activities of the artifact from games and mathematics to conversation, and they relate not only to expectations but also to preparations for action. The intelligent artifact must predict what to expect and how to respond, and the predictions must be not only for the present and immediate future, but for more distant futures as well. It is because the ability to predict (linguistically, to make inferences) is presupposed by intelligent behavior that we may characterize the generator of intelligent behavior as a prediction (or inference) machine (Maron, 1963).

What is prediction and what are the logical grounds for making predictions? A prediction, which has grounds for its assertion, is the logical consequence of an empirical hypothesis. It is the hypothesis which gives the prediction its logical grounds. The linguistic form of an hypothesis is a conditional statement of the type: 'If A, then B, with probability p'. In order to obtain hypotheses from which useful predictions
can be deduced, it is desirable to have the value of the corresponding probabilities be moved as close to 1 or to 0 as possible. To do this one must find relevant items of information and include reference to these in the antecedent of the hypothesis. As a result, the corresponding conditional statement often has a complex antecedent describing a complex state of affairs which must be satisfied in order for the consequent to hold with the probability $p$. Given the initial conditions (i.e. the antecedent is satisfied), a prediction is the assertion of the consequent of an hypothesis of the type described above. Again, predictions of the type we are considering are logical consequences of hypotheses, which themselves may describe complex relationships between events.

This brings us to the question, what is the logic of hypothesis formation? How are empirical hypotheses formed, modified, corrected, etc.? Are there rules of logic which prescribe how to synthesize hypotheses — how to make the logical leap from evidence to hypothesis? We know, of course, that there is no strict logic of discovery in this sense which gives rules that guarantee how to discover hypotheses. Although there are no such rules of logic now — nor can there ever be — we do have a methodology of empirical science. The instrument for forming and improving hypotheses is provided by the scientific method itself.

Roughly speaking, this methodology advises that in order to improve hypotheses one must test their logical consequences against the relevant facts and, depending on the outcome (i.e. whether and how the prediction in question was disconfirmed), modify the hypothesis. Then the cycle of deduction, prediction, testing, and modification continues. Hopefully, the process converges in the sense that the probability of the hypothesis in question can be pushed ever closer to 1 or 0 unless, of course, we are talking about hypotheses describing statistical processes.

The methodology of hypothesis forming, testing, and modifying can be described in the language of feedback and control. Consider the schematization in Fig. 1, which suggests that a switching circuit with feedback from its environment can be considered as the implementation of a self-correcting technique of the type described above. Needless to say, of course, the schematic representation is grossly oversimplified and we have offered no details of the ‘box’ which modifies the hypothesis in

References p. 136
the most probable direction. However, to the extent that there are heuristics concerning how to filter and use relevant evidence, how to make cross inductions, to appraise and improve inductions, etc., it can be done (Reichenbach, 1949).

Two questions may be raised: Is this information processing logic which we have outlined actually implemented in the brain? Can one construct an artifact along these rough guidelines which will be able to learn to interact successfully with its environment and exhibit intelligent behavior?

Models and the notion of an internal representation

One sense of 'model' is a logical representation or set of hypotheses which, when suitably interpreted, correctly describe some empirical events and their relationships. In order to correctly and consistently predict, one must have a model of those features of the environment in question. The notion that our brain (the generator of intelligent behavior) physically embodies a model of the environment was first suggested by Craik as early as 1943. In his book The Nature of Explanation he wrote:

"My hypothesis then is that thought models, or parallels, reality — that its essential feature is not 'the mind', 'the self', 'sense data', nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation." (Craik, 1952, p. 57).

And again:

"If the organism carries a 'small scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, and more competent manner to the emergencies which face it." (p. 61).

The question of how an artifact might most economically embody a model of its world has been discussed with clarity and depth by MacKay (1951, 1952, 1954, 1956, and 1963a). He has argued that since awareness of external events or features implies a 'conditional readiness' to take them into account in the organization of any relevant action, the organizing structure which maintains or controls this conditional readiness can itself serve as an internal representation of the world. For this to be possible, signals received via sense organs must be able to 'irritate' the organizing structure, which in turn must be equipped to 'match' or back them off by an adaptive response which can ipso facto represent the events or features of the world giving rise to them*. That is, the artifact knows about things and events to the extent that it is prepared to react adaptively to them; and to be prepared to react correctly (i.e. to be in a correct state of expectation) means, in a sense, that incoming signals will not put the artifact

* MacKay has pointed out that the more stable the environment, the more redundant will be the signal-pattern to which adaptive reaction is required, and the more economical it becomes to develop an internal generator system to match it and organize an appropriate conditional readiness.
'off balance'. Stated differently, this means that its current predictions will not be falsified. The key point here is that the predictions can be thought of as internally generated signal patterns which will match those signals generated by the external events being predicted. Thus, external events (objects, etc.) are represented internally by those signal patterns which are internally generated to match those generated by the corresponding external events.

“The point of special relevance here is that insofar as an adaptive response is successful, the automatic movements of the goal-directed control centre will in quite a precise sense mirror the current state of affairs, or rather those features of the state of affairs to which it is responding adaptively. The state of the control-centre is in effect a representation of those features of the state of affairs, just as much as a code-pattern of binary symbols could be.” (MacKay, 1954, p. 191).

In a changing environment, what determines how the internal signal patterns should be generated? The answer, of course, is that there is an hierarchical structure to the logical organization of the model which an intelligent artifact embodies, just as there is an hierarchical structure in science itself. This means that whereas there are signal sequences generated on a first level corresponding to predictions about events, other signal sequences, which in turn control the lower-level sequence generators, correspond to predictions about hypotheses. Thus, the generator of intelligent behavior in an artifact is constantly producing those signal sequences which correspond to its expectations. The internally generated signal patterns are continually being matched against the received externally generated signals. The mismatch information is fed back to modify the way in which the patterns are internally produced (i.e. hypotheses are modified). The feedback data modify the ‘organizers’ or internally generated signals, and so on up the hierarchy. It is along lines such as these that one may begin to outline the logical organization of the ‘brain mechanism’ of an intelligent entity.

4. CYBERNETICS, NEUROPHYSIOLOGY, AND EPistemology

Cybernetics and neurophysiology

The publication of Prof. Wiener’s book on cybernetics (1948) was followed by the growing awareness on the part of engineers and physiologists that both the computing machine and the nervous system can be analyzed and described in the language of communication and control, because both are, in fact, information processing systems. There was a rush on the part of some computer engineers to find out about biological switching components, and the circuit organization of the brain. Were there clues from neurophysiology which could be taken wholesale and used by designers of intelligent artifacts? It turned out that at least two different kinds of problems prevented the engineers from picking up gold nuggets gratis. In the first place, there are linguistic and philosophical obstacles blocking the road to a clear discussion of thinking and brain activity. In the second place, the problem of the brain is enormously difficult, involving as it does, one of the most complex systems that we may ever
encounter. We have only the first glimmerings of how the brain is organized so as to allow its owner to know the world.

An interesting facet of this relationship however, is that circumstances have caused the coin to be reversed, in the following sense: The cyberneticist who is concerned with the problem of artificial intelligence and the logic of information processing may be able to suggest general hypotheses for the neurophysiologist. And the cyberneticist's way of looking at things may provide a fruitful framework within which the neurophysiologist may gain a better understanding of natural brain mechanisms and hence guide the selection and interpretation of his data. (It is one thing to study the transport of a chemical substance across a cell membrane strictly in terms of the relevant biochemistry, and it is quite different to interpret the same process as the transmission of a message whose function it is to modify information storage. Compare, by analogy, the way in which a computer might be described on the one hand by a physicist, and on the other hand by a computer engineer or programmer.)

In order to describe and analyze an intelligent entity, one is forced to use such concepts as perception, recognition, knowing, thinking, learning, understanding, etc. These concepts, in a sense, fall completely outside the province of neurophysiology, which is concerned with events 'in the small', i.e. the signal processing mechanisms. However, as we have said, an analysis of the information-flow organization of the generator of intelligent behavior provides the logical timbers needed to bridge the gap between signal processing and behavioral indications of knowing, etc.

The brain mechanism of an intelligent artifact embodies a model of its world. Things and events are represented internally as signal patterns which generate adaptive responses to the things in question. The artifact gains information about its world when its model is updated and modified, i.e. when its states of readiness for adaptive behavior are modified. This is a form of learning. Thinking has its correlate on the information-flow level, in that activity whereby the internal model is driven faster than real time so that the probable future may be examined and alternative behavioral responses evaluated. This activity is part of intelligent planning. And the activity of self-awareness has its correlate in the internal activity of the hierarchical information system whereby one part of the system embodies a model of another part (MacKay, 1952). If the brain mechanism of an artificial agent were organized along the lines outlined above, then it would behave in ways which an external observer would describe as indicative of thinking, perceiving, understanding, etc.*

* For a penetrating analysis of how personal language might be applied in talking about an artifact which did, in fact, exhibit all of the appropriate behavior expected of an intelligent person (see MacKay, 1962b).

**Cybernetics and epistemology**

We have said that intelligent behavior can be characterized in terms of knowing how to carry out a wide variety of tasks ranging from solving mathematical problems and games, to conducting sensible conversation. An intelligent artifact must be designed
to execute the activity of *knowing*. If we put the matter this way, what is the impact of artificial intelligence on epistemology? And, are there concepts which have emerged from cybernetics which are relevant for epistemology? The answer to the question is ‘yes’, but before we go into it a bit more, let us clarify the nature of epistemology.

First of all we must make the important distinction, which is often confounded, between psychology and epistemology. Whereas both have as their subjects the problems of knowledge and how thinking and knowing take place, the psychologist is concerned only with the *actual* processes. The epistemologist, on the other hand, is concerned with how knowing *should* take place so that its outcome will be logically justified. The best place to study the thinking process so as to provide its rational reconstruction as sought by epistemology is in scientific knowledge — the processes of going from observed data to theories which, in turn, are subjected to tests, are falsified, modified, etc., eventually giving rise to knowledge which has good grounds for belief. Reichenbach has said: “Analysis of science comprehends all the basic problems of traditional epistemology; it is therefore in the foreground of consideration when we speak of epistemology” (Reichenbach, 1938, p. 8). So we see that epistemology has as its task an inquiry into the foundations of empirical knowledge and a study and justification of the methods of empirical science, the instrument for predictive knowledge.

To return to artificial intelligence we remember our characterization of the generator of intelligent behavior as a prediction machine; a mechanism which embodies the methodology of science in order to form, test, appraise, modify hypotheses so as to improve its role as a predictor. This suggests immediately that there are at least two points of contact between artificial intelligence and epistemology. Firstly, of course, any methods for discovery and pursuit of knowledge taken from the epistemologist may be valuable to the designer of intelligent artifacts if he can implement them in ‘hardware’. Secondly, there are certain concepts from the theory of information which are useful for the development and refinement of not only epistemology but also the theory of intelligent machines. We refer, of course, to the key concepts of uncertainty and amount of information. We can stop only to sketch the outlines.

We can analyze the prediction machine as an information system which at any given time is in a state of uncertainty since, of course, it does not have secure knowledge of expected events, *i.e.* the probabilities associated with its expectations and subsequent responses are not unity. In principle, the state of uncertainty (of at least part) of the system could be measured. The prediction machine via its feedback connections with its environment, seeks to reduce the amount of its uncertainty. In fact, certain data if obtained by the artifact would provide more information at least in the sense that they would probably reduce a greater amount of uncertainty than other data. The information system could thus be organized to determine which kinds of data contain more information (in this sense), etc. Thus, we are talking about an epistemological problem having to do with the weight of evidence relative to various theories and the development of measures to determine the kinds and amounts of evidence which confirm or shake theories to a greater or lesser extent. What are efficient ways to remove uncertainty, to test and confirm theories, for gaining information

References p. 136
etc.? We are not suggesting that the present theory of information has such measures, nor are we stating that they exist in the literature of contemporary philosophy of science. But it seems clear that both the epistemologist and the cyberneticist are concerned with some of the same problems.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the help I received, on two counts, from Prof. Donald MacKay. First, his ideas have influenced my thinking and provided the impetus to cast this paper in its present form. Second, I thank him for his kindness in reading an earlier draft and suggesting several changes which served to clarify the discussion. However, I alone stand responsible for the paper as it now appears.

I am pleased to acknowledge the help of my wife Dorothy for her skill in suggesting more graceful and precise ways of expressing myself. Also, I am happy to thank Wade Holland of RAND for many editorial suggestions which have improved the form of the paper.

Finally, I am grateful to the management of The RAND Corporation for providing the environment and encouragement to write this paper.

SUMMARY

It is the purpose of this paper to examine the origins, development, and present status of those key cybernetic notions that provide an information-flow framework within which to attack one aspect of the question of how a person thinks; i.e. the question of the information mechanisms and processes which underlie and are correlated with thinking.

After an introductory survey of the scope and ramifications of the information sciences, the cybernetic way of looking at information processing in the nervous system is examined so as to see in what sense it provides new and sharp tools of analysis for the neurophysiologist. With this as background, the problem of artificial intelligence is considered and with that the logical and linguistic difficulties in talking about the relationship between thinking and brain activity. An information-flow model of an artificial brain mechanism is described whose activity, it is argued, is the correlate to activity such as perceiving, learning, thinking, knowing, etc.

This leads finally to a consideration of the impact of these notions on theoretical neurophysiology and its attempt to frame suitable hypotheses, and on epistemology which is concerned with the logical analysis of measures, methods, and techniques which can justify the activity of knowing.

REFERENCES

MacKay, D. M., (1963b); Discussion during a seminar on computers and comprehension. The RAND Corporation, Summer 1963.
McCulloch, W. S., and Pitts, W., (1943); A logical calculus of the ideas immanent in nervous activity. *Bull. Mathemat. Biophys.*, 5, 115–133.
Maron, M. E., (1963); *Artificial Intelligence and Brain Mechanisms*. The RAND Corporation, Santa Monica, Calif. RM-3522-PR.


