Zygon | *Journal of* RELIGION AND SCIENCE VOL. 11, NO. 1

HUMAN VALUES, FREE WILL, AND THE CONSCIOUS MIND

by George Edgin Pugh

Experience over the last decade in the design of automated planning systems has done a great deal to clarify the principles that evolution seems to have used in the design of the human brain. It now seems probable that both the brain and the computerized planning systems are based on a basic design concept which I describe as a "value-driven decision system." Experience in the automation of complex planning problems has shown that, as the problems become more complex, the alternative system-design concepts become progressively less feasible, and the practical design alternatives begin to converge on the basic concept of a value-driven decision system. It seems probable that evolution encountered similar problems in the design of biological control systems, so that over the ages the control systems (the brain) for the more advanced species began to converge on this same basic design concept.

This point of view may have important implications not only for our understanding of the human brain but also for our understand-

George Edgin Pugh is principal scientist at General Research Corporation, McLean, Virginia. This paper is from the forthcoming book, *On the Origin of Human Values,* by George Edgin Pugh. This material is reproduced with the permission of Dr. Pugh and the publisher, Basic Books, Inc.

[*Zygon,* vol. 11, no. 1 (March 1976).]

ing of human motivations and origin of human values. I have recently completed a manuscript for a book which will develop a more comprehensive treatment of these relationships.[1] The purpose of this paper is to outline some of the most important implications of this new, decision-science perspective as it affects both the theory of human values and some old philosophical issues such as the paradox of free will and the mystery of consciousness.[2]

Because this point of view is based on an interpretation of the brain as a value-driven decision system, it is necessary to sketch briefly the main elements of the concept, but I will avoid duplicating the more nearly complete treatment in the forthcoming book.

The Value-driven Decision System

The concept of the value-driven decision system has provided the design framework for a number of automated and semiautomated planning systems. Such systems have been developed to deal with a variety of complex planning and scheduling problems—for example, investment planning, production scheduling, or the development of hypothetical, strategic war plans. The resulting systems can be viewed as artificial decision systems. The systems adapt in a purposeful way to changes in the problem environment. They generate courses of action and detailed plans that are creatively adaptive to the specified problem.

These systems make use of a "mental model" of the problem which they use both to identify feasible courses of action and to project (or simulate) the probable consequences of various decision alternatives. To choose between alternatives, they make use of a value assignment algorithm which provides a way of estimating the quantitative value (or "desirability") of the projected consequences for any course of action. Although the specific design of such systems varies widely, the systems always incorporate at least the following five functional components: (1) a data input system, (2) a data-processing system which uses the input data to update the mental model of the problem environment, (3) a procedure for searching through possible decision alternatives, (4) a system which uses the model to project or simulate the probable consequences of a decision, and (5) a value assignment algorithm which defines the system's scale of preferences for different alternative outcomes.

The primary values that provide the system with its ultimate criteria of decision must be built into the system by the designer; and the resulting behavior of the system can be understood only in terms of this built-in value structure. These primary values are an essential part of the system design. If the human brain is to be interpreted as a

value-driven decision system, then biological evolution must have played the role of system designer. It follows that evolution's behavioral plan for each species must be defined by an underlying system of values, which is an essential part of the design of the brain, and there must be an underlying system of "innate" human values motivating people to behave like people.

The Source of Human Values

Decision systems normally use two types of values: primary values and secondary values. The primary values for any decision system are those that are built into the system by the designer. They define the system's ultimate criteria for evaluating decisions. Secondary values may be developed by the decision system itself as a practical aid to problem solving. The secondary values normally reflect the primary values. Both types of values are used as criteria for decisions. The analysis of the human brain as a decision system indicates that our subjective, valuative sensations (e.g., the unpleasantness of pain and hunger, the pleasant taste of good food, or the pleasure of a sexual experience) are manifestations of a built-in, primary human value system. Like all innate characteristics, these primary values are products of biological evolution. They are probably built into the hardware of the modern human brain, almost exactly as they evolved during our primitive, prehuman past. To understand the human decision system, we must try to understand these innate or built-in values.

Human beings also make use of secondary values. Whereas pain and hunger are simple manifestations of our primary, built-in value system, our moral, ethical, and social principles are all examples of "secondary" values. In ordinary conversation, when we speak of human values, we are almost always concerned with the secondary rather than the primary values. The primary values are so much a part of our "reality" that we take them for granted and rarely perceive them as "values." The secondary values, on the other hand, are products of personal and cultural experience with the environment and with the primary human value system. Because the secondary values are fundamentally products of rational thought, they are appropriate subjects for social debate. According to the decision-theory perspective, however, the built-in primary human value system is the ultimate source of all human values.

The primary values are innate or instinctive. They are built into the human decision system. They cannot be deduced by rational thought, and the individual may not even be consciously aware of their existence. These innate primary values are reflected in both the emotions and the traditional biological drives. The primary values are not con-

stant over time but are carefully orchestrated in response to various stimuli in accordance with genetically inherited rules. For example, the hunger sensation is an innate response to a deficiency in nourishment, pain is an innate response to dangerous heat or physical pressure, fear can be an innate response to specific stimuli such as a sudden noise or a strange person or animal. Because these values respond to internal and external stimuli in accordance with genetically determined rules, it seems appropriate to refer to these values as "instinctive."[3] The resulting behavior, of course, is far from instinctive, but it is motivated by instinctive values.

Actual human behavior is very strongly affected not only by the primary value system but also by the mental model of the world environment and by a very complex network of secondary values that have evolved as a consequence of personal and social experience with the environment. This decision-science perspective, therefore, may open the door to a more scientific way of understanding and evaluating our social and ethical values.

In most historical societies the transmission of information about values has been one of the most important functions of education. In the United States, as a result of our principle of freedom of religion, the responsibility for such instruction has been left primarily to the church and the family. Now with the decline in the influence of established religion a large fraction of the population fails to receive any education concerning their own human values. This unnecessary ignorance of a subject that is of the highest personal importance is a real handicap both to the individual and to society at large. Perhaps the availability of a scientific approach to the topic will make possible a more objective and scientific education in the field of human values.

If there is any field of study that should pass the test of "relevance," it is the relationship of "human values" to human decisions. The need for a better understanding of human values has probably never been more acute than at present in our rapidly changing society. Traditional methods for dealing with valuative issues have amounted to little more than trial and error, and they are far too slow and unreliable for an environment of rapid change. As a result of scientific and technological progress, many traditional value commitments no longer seem adequate or relevant. While scientific progress has tended to undermine the traditional ethical and religious perspective, it has failed to produce any generally acceptable replacement for traditional ideas. With the erosion of old religious and ethical convictions there is concern that society will be left without effective guidance or control.

The lack of credible value criteria seems apparent in both our per-

sonal lives and our public policy. It is apparent in the increased crime, violence, and general amorality of the cities. It is evident in the sense of meaninglessness and despair that so many feel in their personal lives. It is reflected in the widespread concern that public policy has become disconnected from human objectives and that our social institutions are pursuing abstract and meaningless economic goals while ignoring fundamental issues of real human value.

R. W. Sperry, professor of psychobiology at California Institute of Technology and one of the foremost scientists in the field of brain research, has become a vigorous advocate of a scientific approach to the problem of human values. More than any other scientific writer, Sperry seems to have recognized that human values originate in the basic design of the human brain. When human values are recognized as an essential part of nature's system design for the brain, they become a natural subject for scientific study. Sperry states his case very clearly:

> I tend to rate the problem of human values Number One for science in the 1970's, above the more concrete crisis problems like poverty, population, energy, or pollution on the following grounds: First, all these crisis conditions are man-made and very largely products of human values. Further, they are not correctable on any long-term basis without first changing the underlying human value priorities involved. And finally, the more strategic way to remedy these conditions is to go after the social value priorities directly in advance, rather than waiting for the value changes to be forced by changing conditions. Otherwise we are doomed from here on to live always on the margins of intolerability, for it is not until things get rather intolerable that the voting majority gets around to changing its established values. It is apparent, further, that other approaches in our crisis problems already receive plenty of attention. It is the human value factor that has been selectively neglected and even considered, in principle, to be "off limits" to science.[4]

Of course, the "social value priorities" mentioned above are secondary human values. The secondary values are products of the human mind, interacting in the context and memories of sociocultural systems, and are therefore subject to change on the basis of changes in the sociocultural system and increasingly by factually informed rational thought. It is the "human value priorities" which need to be brought into harmony, both with the modern technological environment and with the innate (or primary) human values.

Because the innate human values are built into the human mind as part of our genetic inheritance, they are not subject to change either by rational persuasion or by social pressure. The primary values originally evolved in a primitive, prehuman society very different from our modern urban environment. There is reason to believe that much

of the discontent and alienation we find in modern society may be the result of our failure to ensure that the modern social environment remains compatible with our ancient and innate human values.

The Brain and Introspective Experience

In humans the cognitive decision process is intimately linked with our sense of awareness or consciousness. Indeed, it seems likely that consciousness itself has evolved as a natural (perhaps inevitable) by-product of the value-driven decision system. In our everyday speech we speak of consciousness as if it were a well-defined object. Yet there are reasons for believing that consciousness is a physically distributed and diffuse property of the brain. The content of consciousness at any single moment is quite limited. But as we shift our attention we can bring into consciousness a wide variety of alternative elements from specific parts of present sensual experience to a preoccupation with conceptual or theoretical problems.

These subjective experiences are quite consistent with what we should expect as a consequence of the design principles for a large cybernetic system. Compared to the solid-state junctions used in modern computers, the neurons of the brain have quite slow response characteristics, but they are much smaller and more compactly arranged than can now be achieved with the junctions even in modern integrated circuit computers. To achieve cybernetic efficiency in a system using such components it is essential to make extensive use of parallel processing methods.

It seems likely that the brain operates much like a large associative processing computer which includes not only a central processor but also numerous peripheral processors that can carry out specialized data-processing functions without interfering with the central processor. The operation of a large data-processing system using peripheral processors is organized and coordinated through a central control unit which sets priorities and allocates tasks to the peripheral units. It seems probable that the center of consciousness serves an analogous role in the human brain. Our ability to shift our attention and bring different things into consciousness may correspond to the selective interaction of this central control with various peripheral processors and with various parts of a large associative memory. An examination of the structure of the brain from this point of view suggests that the center of consciousness may be in the general vicinity of the thalamus.

Of course, many readers will claim that this functional discussion does not really address the riddle of consciousness and that it only moves the problem to another level in an infinite logical regression.

They will say that in postulating the thalamus as a central control system we have really postulated a little decision system inside our main decision system and have thereby only moved the basic dilemma of consciousness from the main decision system to the little decision system. And then, they will ask, what about a microdecision system inside the little decision system?

I believe this concern is unfounded. Electronic computer systems also have control systems. Moreover, they operate autonomously, without an infinite sequence of control systems. The central control is usually quite simple, and it does not require a control for the control, etc. On this basis it seems reasonable to assume that the sensation of consciousness may indeed arise from the interaction of the control center with the peripheral processing units and that it does not require an infinite regression of control systems to provide an explanation.

On the other hand, the true physical basis of this sensation of consciousness remains a mystery.[5] It is even unclear whether the mystery is philosophical, metaphysical, or scientific. This basic theoretical dilemma about the origin of consciousness will be addressed in more detail in a later section.

In our subjective experience things are always happening which we did not plan. Ideas come to us which we cannot rationally explain. These experiences convince us that the subconscious plays an important role in our intellectual lives. How can we reconcile this subjective experience with our understanding of the brain as a system?

Much of our subjective experience with subconscious mental processes can be attributed to the presence of many, almost autonomous, information-processing centers. These processing centers communicate with the conscious mind (and probably with one another) only when they have something relevant to say. The unexpected messages from these processing centers are experienced subjectively as a profound mystery.

In a large cybernetic system where the basic components (or neurons) are quite slow (relative to electronic components), efficient operation demands extensive use of peripheral processing subsystems. This is necessary to allow simultaneous processing of different types and classes of information. Such parallel operation increases the total processing capacity by bringing a large number of neurons simultaneously into the activity.

The peripheral processors are not, however, the only source of subconsciously generated information. A number of qualitatively different system functions can be identified which contribute in different ways to our awareness of a subconscious mind.

1. *The Operations of Instinctive Drives and Values.* The selection of primary goals and objectives is not under conscious control. The innate values (or drives) can produce almost uncontrollable urges that may be inconsistent with our rational objectives. They can also provide an intuitive sense of "right" or "wrong." This intuitive conscience may be perceived to be of either subconscious or supernatural origin, depending on the religious persuasion of the individual.

2. *Automatic Analysis of Sensory Data.* The sensory data received by the conscious mind have already been subjected to much logical analysis by other system components. This is particularly true of visual and auditory information. Basic two-dimensional visual images are interpreted as three-dimensional structures outside the realm of the conscious mind. The perceived colors of objects are automatically compensated for the redness or blueness of illumination before the information reaches the conscious mind. These automatic interpretations of the input data can sometimes be in error. Nevertheless, the information is presented to the conscious mind as if the interpretation were an essential part of the original sensory data. The procedures used in this prior processing of sensory data are not accessible to conscious thought (except, of course, after scientific discovery), and they provide another subjective manifestation of a subconscious mind.

3. *Operation of the Associative Memory.* The recall of past experiences cannot be voluntarily commanded. An associative memory requires a certain correspondence between the interrogation and the stored retrieval keys in order to recover the stored information. Since the precise retrieval keys are not usually known by the conscious mind, the success of retrieval (or recall) is somewhat unpredictable. Either success of failure of recall can come as a surprise. In the case of frequently used information, the retrieval keys may become reliably available to the conscious mind. However, the recall keys for little-used information may be difficult to locate. They may be encountered almost accidentally, as we succeed in retrieving related information, including the desired retrieval keys. Thus, by a chain of association, we may ultimately recall the desired information.

4. *Operation of Peripheral Processors.* In all probability the brain includes a complex hierarchy of peripheral processors that carry out analysis and comparisons at different levels of abstraction. The processors that operate at low levels in the hierarchy may never be directly accessible to the conscious mind. Others that operate at a higher level may report to the conscious mind only when they have relevant results to report. The results generated by these processors can be delivered to the conscious mind at unexpected times long after the

original stimulus for the activity. Scientists frequently report that critical ideas have "come to them" suddenly when they were relaxed or engaged in quite unrelated activities. The sense of suddenly "knowing the answer" can be very strong even before there is conscious awareness of the answer itself. In such cases, the individual may feel extremely confident that the answer is valid, despite the fact that much work may remain for the conscious mind, in order formally to prove the result.

All of the above phenomena, which once seemed to be impenetrable mysteries, now appear to be a natural consequence of the semiautonomous peripheral processors that are needed to provide parallel processing in a large cybernetic system. Although the relationships just described remain speculative (because we do not have a detailed physiological understanding of the neural structures involved), the circumstantial evidence for this type of design concept seems very strong.

The Paradox of Free Will

The interpretation of the human brain as an automatic decision system provides a rather simple resolution (at least in scientific terms) of the long-standing controversy concerning free will versus determinism. The traditional discussions of free will have been concerned with reconciling two extreme points of view. On one side, there are those who wish to demonstrate the existence of an almost metaphysical entity known as the will. The will is presumed to exercise ultimate and unpredictable control over human behavior without regard for (and perhaps even in violation of) physical laws. On the other side, there are those who view the mind as an essentially predictable physical system which fatalistically follows deterministic laws. The decision-system concept offers a well-defined scientific model which occupies a position intermediate between these two traditional extremes. Whether this model confirms or denies free will depends, of course, on how we define free will. Rather than become involved in an unproductive discussion of the definition of free will, I will try to focus attention on those traditional attributes of free will which seem to be scientifically observable.

From this perspective there seem to be three key issues. First, does the system make decisions or choices? Second, are the resulting decisions "free" choices? Third, are the choices predictable? The decision-system model provides rather direct answers to each of the above questions.

First, a decision system does in fact make "choices." It considers alternatives and decides which alternative to select. Apparently, dur-

ing the early stages of the free will controversy, it was assumed that the process of making choices could not be "explained" in terms of physical laws. Thus the mere existence of such a process was accepted as evidence for a metaphysical entity called the will. Our experience with artificial decision systems shows decisively that choice is not inconsistent with physical laws. It is a form of natural behavior which man shares both with other animals and with artificial decision systems.

The argument that freedom of choice does not exist was reiterated recently by B. F. Skinner.[6] The underlying reason for Skinner's belief that freedom of choice is an illusion seems rather obvious. His "operant conditioning" model of human behavior does not include the concept of choice or decision, so obviously it cannot include freedom of choice.

But Skinner makes his formal argument on an entirely different basis. He claims that free choice cannot exist because everything that man does is an automatic consequence of the interaction between his genetic inheritance and his experiences with the environment. The same argument could be made with equal force for a choice-making model of human behavior. If we were to apply this logic to the decision-system model, we would have to conclude that free choice does not exist because the individual will always select the decision alternatives that seem best in terms of his world model (which is a consequence of his experience with the environment) and his innate value system (which is a genetic inheritance). Certainly, the resulting choices in the decision-theory model must be accepted as "natural" consequences of experience and inheritance. But does this mean that the choices are not "free"?

This, of course, is our second key question, and the answer to this question depends on what we mean when we describe a choice as free. According to our decision-theory model, the human brain will always try to make choices that seem the most "desirable" in terms of a built-in, genetically inherited value system. Because the brain is not free to select its own primary value system, one might claim that the brain does not really have freedom of choice. From the decision-theory perspective, however, such an argument does not make much sense. Deliberate free choices can be made relative only to a built-in scale of values or some other primary decision criteria. In the absence of a criterion of decision there can be no real "choices" but only "chance" or random decisions. When we speak of freedom of choice, we surely are not referring to random, mindless decisions. If decisions are to be based on choice rather than chance, then there must exist some fundamental criterion of decision. It appears that if free choice

is possible at all, it is possible only in the context of an a priori value system.

This difficulty cannot be avoided by allowing a decision system to select its own value system, for such a selection itself can be only a random chance decision unless it is based on an a priori set of values. Thus we are led inevitably to the conclusion that choice can be meaningful only in the context of a preexisting set of values and that the fundamental values can never be supplied by the system itself. From a decision-science perspective, real choice is not possible except in the context of preexisting values. If our definition of free choice requires that the system must rationally select its own primary values, then free choice is a logical impossibility, and we must concede that we do not exercise free choice.

But in everyday conversation what we usually mean by free choice is that we are free to consider alternatives and to select what seems best. Obviously, by this definition, we do exercise freedom of choice. To claim that we do not have "freedom of choice" because we will always choose what seems best in terms of our own innate values is a commonsense absurdity. It is equivalent to saying that we do not have freedom of choice because we will always choose to do what we like! Although we did not choose our innate values, we certainly do make choices in terms of those values. Moreover, the choices are as free as it seems theoretically possible for a choice to be. Apparently, in terms of any reasonable commonsense definition, we must admit that freedom of choice exists.

But if we focus our attention on the deterministic extreme of the traditional free-will debate, the issue appears to center neither on the existence of choice nor on the freedom of the choices but rather on whether the choices are inevitable, deterministic consequences of natural laws. Obviously, in any scientific theory of behavior, choices would have to be explained as a consequence of natural laws. But it is an open question whether the choices are inevitable, predictable, or deterministic consequences of the natural laws.

We must therefore address the question of inevitability or predictability. In this respect biological systems differ fundamentally from present artificial decision systems. Of course, adherence to physical laws no longer requires a strict determinism. According to the laws of quantum mechanics, systems at the atomic and molecular level are predictable only in a statistical sense. Thus microscopic phenomena are subject to an irreducible uncertainty.

Modern electronic computers are engineered so that the predictability of their results will not be impaired by random thermal noise or by quantum uncertainties. However, biological cybernetic systems

are designed very differently. The all-or-nothing response of a single neuron in the brain depends on input from hundreds of other neurons, some of which will tend to facilitate a response while others will tend to inhibit a response. The response of the neuron thus depends on input from many, many sources. When the input signal is very close to the level required for a response, minor fluctuations in temperature, chemical composition, and possibly even quantum mechanical effects can determine the response of the neuron. The unavoidable uncertainties are thus amplified by the all-or-nothing response of neurons so that they can have important effects in the subsequent system behavior. Thus the brain is a mechanism capable of amplifying microscopic thermal and quantum fluctuations to the macroscopic level of human behavior. Even if two physically identical brains could be placed in exactly the same environment, the results produced by the brains would be different. Evidently, the operation of the brain is far from predictable, even in principle.

However, the practical difficulties in predicting the behavior of a complex decision system are acute even without the fundamental uncertainties. This can be seen by considering the behavior of an artificial system which is completely deterministic and therefore predictable, at least in principle. But how predictable is such a system in practice? The answer is that its behavior is really not very predictable, even to the designer of the system. The purpose of the system is to make or recommend decisions. If the resulting decisions could be easily predicted, there would be no need for the system. Of course, if two identical systems were prepared and provided with identical input data, the subsequent behavior would be identical. In a sense, each system could predict the behavior of the other. However, this is almost an academic observation, for the slightest difference in the two systems would completely destroy the predictability of the results.

Almost any decision system will encounter numerous points where there is approximate indifference between alternatives. The final selection of one of the alternatives then depends on trivial details, such as the way numbers are rounded in the machine, the sequence in which the alternatives are encountered, or the number of significant figures used to represent values. Once a different decision has been made at such an indifference point, the subsequent behavior of the two systems will diverge, for they are not longer facing the same problems.

Most computer scientists are very familiar with this type of unpredictability. A very typical example occurred when a production scheduler was transferred to a different computer with a slightly different level of accuracy in the representation of numbers. Despite the

fact that no change was made in the basic system design, it was impossible to duplicate the schedules produced on the previous computer. Although the resulting schedules might start out being identical and might remain so for a few weeks of the schedule, sooner or later a decision of near indifference would be taken differently. Thereafter there would be little resemblance between the two schedules.

Fundamentally, the difficulty in predicting the behavior of a decision system arises because there are thousands of alternative courses of behavior that are almost equally optimum. The choice between such alternatives is largely a matter of chance. When the first "adequate" alternative is found depends on many details both of the structure of the world model and of the order in which decision alternatives are examined. As a practical matter, the accuracy of knowledge required to predict how an organism will resolve such choices of near indifference can probably never be attained.

Thus a science of human behavior can never do more than to identify a plausible range of action alternatives. It is unlikely that it will ever be possible to predict behavior very fully. Human beings operating in an environment of other human beings are continuously making unpredictable personal decisions in an environment that is intrinsically unpredictable. The state of any individual's mind, as well as the state of his values, is the result of his experience during a long, complex chain of past decisions. The level of predictability that can be expected without an unattainable precision of knowledge is necessarily very low.

From the preceding discussion we can conclude, first, that human beings do make choices; second, that the choices are as free as it is theoretically possible for choices to be; and, third, that the choices are intrinsically unpredictable. At best, human decisions can be only imperfectly predicted. Obviously, one can continue to argue about the definition of free will, and, in the context of some of the definitions, one can deny its existence. However, if we accept a definition based on the concepts of choice and unpredictability, then it appears that the affirmation of free will corresponds better with reality than does its denial.

The Mystery of Consciousness

Most of the objectively analyzed functions of the human decision system can just as well be described as functions of the conscious mind. These objective functions of man's decision system, which I have analyzed in detail elsewhere,[7] include (1) the reception of input sensory data, (2) the assignment of "values" to experience, (3) data storage and retrieval (or memory), (4) objectification and symbol use,

(5) building and refinement of the world model, (6) the allocation of intellectual effort, (7) simulation—using the model, (8) decision —based on the outcome of simulation, and (9) control of action. From a purely functional point of view, these seem to define the main activities of the conscious mind. Yet, paradoxically, they do not even include the concept of consciousness! There is at present no consensus with regard to this paradox. There are at least three conflicting points of view. Probably the best way to clarify the nature of the problem is to outline briefly these three alternatives.

1. *The Functional View.* According to this view, a theory of consciousness needs only to encompass the functional activities of consciousness. The above basic functions of the human mind do, in fact, encompass both the objective and subjective functions, and thus they provide an adequate explanation of consciousness. After all, if a system can receive information about the environment, recall and compare past experiences, sense the quality of experience, make conceptual models of the environment, project the consequences of alternative courses of action, and then select and implement preferred courses of action, what else does consciousness entail?

However, this purely functional view does not seem entirely satisfactory to most people. According to the functional view, any system which performed these basic functions would be conscious. A computer programming expert could easily design computer programs which, at least in a rudimentary way, would have all of these characteristics. However, even if a program were designed to have all of these characteristics, even at a very sophisticated level, most experts would see no reason to expect that the program would acquire consciousness. Indeed, it is difficult to see how a computer program could acquire any subjective sensations at all. It is even more difficult to see how it could acquire the sensation of wholeness and continuity of experience that we call consciousness.

Supporters of the functional view, however, would reject these objections on the good scientific grounds that we have no way of knowing whether the computer system is conscious. Indeed, there is no reason to believe that the system would behave differently if it were conscious.

Obviously, a large part of the difficulty lies in the fact that there is no known way for an external observer to determine whether a system is conscious or not. This functional perspective, however, is probably the most simple and internally consistent theoretical view, so I will return to it after considering the other points of view.

2. *The Metaphysical View.* According to this view, consciousness does not occur and cannot occur in any machine. It is a unique

characteristic of complex biological systems, and it may be unique to the human brain. It is a metaphysical concept, perhaps related to the "soul," and it cannot be explained by any physical laws. The obvious objection to this concept is that it is a scientific cop-out. It provides no explanation, and, if scientific investigation were founded on this premise, it would almost guarantee that no theory would be provided in the future.

3. *The Biological View.* According to this view, consciousness does not occur, and cannot occur, in computers or computer programs as we know them today. The programs might seem to behave as if conscious and might even assert that they are conscious when asked. But they would not "really" experience the subjective sensations of consciousness that are common to the higher animals.

The reason these computer systems cannot experience consciousness is that they are fundamentally different from biological systems in their cybernetic structure. Present digital computer systems are essentially sequential processing systems. While they have large high-speed memories (where results can be filed and later retrieved), and they may have several peripheral processing units, their actual processing in any unit is done sequentially one step at a time. The results of each processing step are dependent only on the status of a very narrowly defined set of system elements. This narrow, disciplined approach to data processing makes it much easier for human programmers to control the computer and to obtain predictable results.

The essential logical operations in the brain are probably much more complex. Information, images, and concepts can flow through the neural network in a completely parallel, wavelike form which is unlike anything that can occur in present digital computers. The phenomena of consciousness, therefore, could be a special, as yet not understood, consequence of this informal, parallel-processing approach to data analysis. According to the biological concept, therefore, the subjective sensation of consciousness is a natural, but as yet not understood, consequence of the foregoing functions of consciousness in a cybernetic system like the brain. But such consciousness would not be expected to occur in the narrowly disciplined operation of present-day computers.

The obvious objection to the biological concept is that it is not really an explanation. It offers the hope of an explanation, but, unlike the functional concept, it does not purport to provide one.

The Missing Test. It was mentioned earlier that one of the most critical problems facing a theory of consciousness lies in the lack of an objective test for the existence of consciousness. A useful scientific theory must make predictions that can be tested. The functional view

does in fact make predictions. It predicts that any cybernetic system with certain specified properties will be conscious. Unfortunately, however, the prediction is not subject to experimental verification because there is no known test for the existence of consciousness.

The seriousness of this difficulty can best be understood by imagining circumstances where such a test would be needed. Suppose one were to construct a complex robot which encompassed all the previously listed functions of consciousness. With appropriate choice of design parameters we would expect that the robot might behave much like an intelligent animal. With appropriate input/output routines, the system could probably be designed to communicate in English. Terry Winograd, in fact, developed a computer program capable of limited communication in natural English language.[8] Thus one would be able to carry on a conversation with such a robot. One might ask the robot if it is conscious. In response, the robot might say, "Yes."

Supporters of the functional view might believe the robot, but most other scientists would not. They would reject consciousness in the robot on the grounds that they know the system design and know of no mechanism by which any component or group of components in the system could acquire the sense of whole experience that we call consciousness. But supporters of the functional view would claim that consciousness is a distributed sensation of the whole system, which cannot be localized to any single component or any single point of the system logic.

While most people will reject the robot's claim of consciousness, they will readily accept the same claim by another person or even by another intelligent animal. Basically, we are willing to attribute consciousness to people (and perhaps to animals) because they seem similar to ourselves, and we "know" we are conscious.

The logical strength of the functional view can best be perceived by reversing the encounter with the robot and allowing the robot to ask if the man is conscious. When the man answers "yes," the robot can proceed to ask what he means by consciousness and what the attributes are of his experience that he identifies as consciousness. As the man responds, the robot murmurs, "Just like me!"

The Missing Definition. The issue of consciousness raises profound questions about how much one can or should ask of a theory. For example, it is apparent that the paradox of free will is largely a result of problems in the definition of free will. Is it possible that similar problems are involved in the definition of consciousness? Perhaps if we could obtain an adequate definition, we could also provide an adequate theory.

To address this question we can imagine a scenario involving a brilliant computer designer and his skeptical supervisor. The computer expert claims he can make a computer system that will be conscious. The supervisor does not believe him, so they agree to a test. The supervisor then writes down a complete list of all the characteristics the system must have if it is to meet his definition of consciousness. Some months later the designer returns and reports success in the project. To prove his point, he puts on a demonstration which shows that all the specifications have been met. He then challenges the supervisor.

"Now," he asks, "do you believe it is conscious?"

"No!"

"Why not? It meets all of your requirements!"

"Well, I'm not sure, but I think I must have omitted some important characteristics in my definition of consciousness. Let me think for a while."

A few days later the supervisor has collected some additional critical specifications and he gives them to the designer. The designer goes back to work and returns a few months later to report success. Again, he puts on a demonstration, and again the supervisor is unsatisfied. The process is repeated several times until the supervisor can no longer think of any additional requirements for his definition. The designer claims success, but the supervisor still is not satisfied. The designer wants to know why.

"It met all your criteria; you can't even think of any more! Why do you stubbornly refuse to believe it is conscious!"

"Well, I just can't believe that that pile of wires and circuit elements is conscious."

"In what way would it behave differently if it were conscious?"

"I don't know. I admit it acts as if it were conscious! It shows emotion; it reasons, it remembers; it seems to have a highly developed ego! It even asks the right questions when we don't give it complete information. But I still don't think it is conscious."

"Then, for heaven's sake, why don't you believe it is conscious?"

"I don't really know. But I do know how you designed it. We talked about all the components. You never showed me any part of the design that would give it consciousness. So I don't believe it is conscious. I believe it is just a complex computer program and an assembly of electronic components. I don't know what I omitted in my definition of consciousness, but I am sure I forgot something."

The dialogue is imaginary, but it could easily be real. Why is the supervisor still skeptical? Basically, he wants to know whether the system is "really" conscious. He wants to know how it "really" feels,

not how it says it feels. There is no way he can find out. All his complicated specifications are really efforts to specify how the system must feel. In the end he can only observe how it acts; he can never know how it "really" feels—if indeed it feels at all.

There is one critical experiment that he would have to do, to be sure. He would have to get inside the system and be the computer. That he can never do. The ultimate test of consciousness can be applied only to ourselves. We can never be sure that any one else is "really" conscious. We are willing to accept that they probably feel as we do on the circumstantial evidence that they look and act much as we do.

If a theory is supposed to make predictions that can be verified, then perhaps the most we can ask of a theory of consciousness is that it be able to explain why a system acts as if it is conscious. Perhaps it is asking too much to require a theory to tell us whether a system "really" is conscious. The existence or nonexistence of an internal consciousness is not experimentally observable outside the system. It is observable only in a subjective sense. This does not mean that it is unreal; it is probably the most "real" of all our personal experience.

The preceding discussion makes no effort to select a preferred theory of consciousness. Although the theory of values is concerned with the cybernetic functions of consciousness, both the essence and the mechanisms of consciousness remain a profound mystery. Any or none of the foregoing concepts (the functional, the metaphysical, or the biological) could be correct. It is possible that in the future a theoretical perspective will evolve that will provide a more satisfying explanation of our personal sensation of consciousness. But such an explanation does not seem to be needed for the development of a theory of values.

The Possibility of a Dual Consciousness

I have suggested the thalamus as a probable center of consciousness of the human mind. However, the left half of the thalamus, which appears to be structurally almost completely separate from the right half, services only the left hemisphere of the forebrain, and this raises the strange possibility of a dual consciousness within the human brain. Is consciousness correspondingly divided? If so, does one side of the brain serve as a sort of peripheral processor for the other? In terms of the discussion of values, the issue is not critical. It does not really matter whether we are dealing with (1) a unified system (in which the central control is really unified but has two halves), or (2) two cooperative, independent systems in which each serves to assist the other, or (3) two cooperative, independent systems in which one

side leads and the other serves as a peripheral processor. Having already recognized the need for peripheral processors, we can treat the problem of the two halves of the brain simply as a special case of that very general concept.

Nevertheless, the question is very significant both in terms of our understanding of the human consciousness and in terms of our perception of the subconscious. By now a wide variety of experiments have confirmed that the two halves of the brain are in fact capable of operating independently. When the information transfer between the two cerebral hemispheres is stopped by surgically cutting the connecting fibers of the corpus callosum, the two sides of the brain appear to operate like two completely independent control centers.[9] This is consistent with what we would expect based on the physical structure of the thalamus, and it adds support to our structural interpretation of the system.

It is still not conclusive whether the two halves of the brain normally operate independently or as a single unit. The weight of evidence, however, seems to favor the hypothesis that the two sides operate rather independently. If there really is a hidden, independent intelligence in the nonverbal side of our brains, we know that we will never be sure whether it is "really" conscious. The most we will ever learn is whether it acts as if it is conscious. From the point of view of our conscious, verbal mind we would have to think of the nonverbal side as an imperfectly controlled peripheral processor. Intellectual activities and action decisions of the nonverbal half would appear to the verbal side to be subconscious or intuitive.

If this point of view is correct, then the scope of the conscious mind may actually be limited to only one hemisphere of the brain. The nonverbal hemisphere should then be considered to be a major subsystem of the brain that lies outside the mind.

The possibility of such a division within our own brains seems very difficult to believe. The only thing that makes it credible is the actual results from the split-brain experiments.

Relation to Idealized Model

One purpose of this paper has been to provide a better understanding of the relationship between the idealized model of a value-driven decision system and the imperfect realization of that model in the actual, human decision system. The human system suffers from many limitations because of design compromises necessary to realize a versatile decision system despite the limitations of the size of the skull and the size and processing speed of the neuron.

The human brain suffers from imperfect memory storage but,

more importantly, from imperfect ability to recall stored information. It suffers from imperfect control over peripheral processors and almost total lack of ability to monitor the activities of peripheral processors. Apparently, it suffers from difficulties of internal communication between the left half, which specializes in symbolic and verbal logic, and the right half, which specializes in analogue or representational logic. If suffers from limitations in the scope of activities that can be encompassed within conscious awareness at one time. This limitation is only partially overcome through the use of subconscious peripheral processors that can proceed automatically with well-established, habitual behavior.

Like any finite decision system, it suffers from an inability to think usefully very far ahead in an environment of uncertainty. For this reason, it has been supplied by evolution with a time-dependent, innate value system. These values are designed to motivate satisfactory behavior without excessive dependence on uncertain projections of the more distant future. Like any finite decision system, it must operate with imperfect and incomplete world models. It must make decisions on the basis of an exploration of a small number of action alternatives.

Within these practical limitations, however, it remains true to its basic concept as an optimizing, value-driven decision system. Many readers, of course, will object to the use of the word "optimizing" to characterize a system that so readily accepts action alternatives that are so far from optimum.

Herbert A. Simon in his challenging book, *Models of Man,* was very careful to point out that man does not operate on principles of optimization.[10] Human behavior, he points out, is better described as a policy of "sufficing," that is, of finding solutions that are adequate, but almost never optimum. If an adequate solution cannot be found, the individual will continue to "worry." He will explore and reexplore alternatives until a satisfactory solution is found or until the urgency of time forces acceptance of an unsatisfactory alternative.

The present perspective makes it clear that just such a policy of "sufficing" is an inevitable consequence of a broader policy of "optimization," in which finite, cybernetic resources must be conserved and allocated.

One of the most important characteristics of the human decision system is the dependence of behavior on the quality of available world models. This dependence was vividly illustrated in some experiments designed to measure actual human behavior in game situations. Considerable uniformity of behavior was observed among most of the subjects. However, the possibility of any universal conclusions was

destroyed by a few subjects who had studied game theory. These subjects played quite consistently in accordance with that theory. Obviously, successful prediction of behavior requires an understanding of both the goals or values and the underlying world model. A better world model will usually facilitate better or more effective behavior.

It is worth emphasizing that a better model is not necessarily a more accurate or a more nearly complete representation of reality. It only needs to be more useful. It may be more useful if it is simpler, easier to understand, or easier to work with. The simpler the model, the easier it will be to examine alternatives. Obviously, the quality of decisions involves a trade-off between accuracy, which permits a better evaluation of individual alternatives, and simplicity, which permits more alternatives to be examined. The dependence of system behavior on the available world model explains, of course, the importance of education and the importance of simplifying theories.

One of the most important aspects of a world model is the self-model. A better model of self should allow better use of our own physical and mental resources. It should permit a better focusing of our efforts through a more accurate understanding of personal goals and objectives. It is hoped that, for some readers, the model described here will serve that purpose. One of the wisest pieces of advice ever given was contained in two words, "Know thyself."

The model of the human brain as an automatic decision system provides a suitable model only at one level in our intellectual hierarchy. Obviously, there are other models that are more suitable at other levels. We can hope and expect that the decision-system model will someday be "explained" in terms of more fundamental or reductionist physiological knowledge. Such a theory might "explain" both the decisions and the orchestration of values in terms of changes in synapse sensitivities and the detailed structure of the neural network. Such a model could be more nearly complete and more accurate, but it would not necessarily be more useful. The decision-system model of human behavior is offered here in the hope that it will provide a model that has practical value in human decision making.

The finite cybernetic limits of the human decision system explain a number of phenomena that seem superficially inconsistent with the behavior of an optimizing system. Human beings are suggestible, and they can often be misled by authority. In a mob or group environment, they will collectively commit atrocities that they would not consider as individuals. All of these weakness are consequences of limitations either in the value structure or the analysis capacity of a finite decision system.

A human being can consider no more than a small number of

action alternatives at any time, so he is inherently susceptible to suggestions. If a suggestion is offered which provides an adequate or "sufficing" alternative, the suggestion is likely to be accepted. The suggestion changes the state of the decision system. It makes available an alternative that otherwise might not have been consciously perceived. Thus the "suggestion" inevitably increases the probability that the suggested course of action will be chosen. By accepting a satisfactory suggestion, the individual avoids wasting intellectual effort to devise his own alternative.

Individuals are sometimes misled by authority because reliance on authority is often a good way to conserve cybernetic resources. If prior experience has shown that an authority is usually right, it will be more efficient to accept ideas from an authority than to develop independent ideas. After all, even independently generated ideas can be in error. It is simply a question of the individual's estimate about the odds favoring ideas from the authority, compared to ideas he might independently generate. In considering the odds, he must also consider the saving in cybernetic resources that is possible by accepting ideas from an authority. Of course, the issue is usually not consciously analyzed as above, but the intuitive decisions about the use of authority nevertheless reflect such considerations of cybernetic efficiency.

The foregoing considerations about cybernetic efficiency are often interwoven or confounded with value considerations. If may seem desirable to accept a suggestion in order to please or to accept the ideas of an authority to avoid adverse consequences. The time dependence of values and the instinctive social response of the value system to a group environment account for much of the apparent irrationality of mob action.

The emphasis in this paper on the limitations of the human mind should not obscure the achievement of evolution in the design of the human brain. Any real cybernetic system must be finite, and any finite cybernetic system would suffer in some degree from the human limitations. In such a design, compromises are inevitable, and the quality of the system will reflect the quality of the compromises.

NOTES

1. George Edgin Pugh, *On the Origin of Human Values* (New York: Basic Books, in press).

2. The material on free will and consciousness is based on ideas developed in Pugh (n. 1 above), chap. 7.

3. Psychologists have properly become very wary of using the word "instinctive" to describe any aspect of human behavior, so the use of the word in this context is likely to be controversial. As will be shown later, however, the innate values exhibit characteristics such that the word "instinctive" seems to be more applicable to these values than to any other aspect of human behavior.

4. R. W. Sperry, "Messages from the Laboratory," *Engineering and Science* (January 1974), p. 32.
5. For a summary of some recent speculations see John C. Eccles, *Brain and Conscious Experience* (New York: Springer-Verlag, 1966).
6. B. F. Skinner, *Beyond Freedom and Dignity* (New York: Alfred A. Knopf, 1971).
7. Pugh (n. 1 above), chap. 5.
8. Terry Winograd, *Understanding Natural Language* (New York: Academic Press, 1972).
9. R. W. Sperry, "The Great Cerebral Commissure," *Scientific American* (January 1964), pp. 42–52; Michael S. Gazzaniga, "The Split Brain in Man," *Scientific American* (August 1967), pp. 24–29; R. W. Sperry, Michael S. Gazzaniga, and J. E. Bogen, "Interhemispheric Relationships: The Neocortical Commissures; Syndromes of Hemispheric Disconnection," *Handbook of Clinical Neurology,* ed. P. J. Vinken and G. W. Bruyn (Amsterdam: North-Holland Publishing Co., 1969), 4:273–90; and Stuart J. Dimond and J. Graham Beaumont, *Hemisphere Function in the Human Brain* (New York: John Wiley & Sons, 1974).
10. Herbert A. Simon, *Models of Man* (New York: John Wiley & Sons, 1957).