Theory and Evaluative Reviews

On some two-way barriers between models and mechanisms

WILLIAM R. UTTAL

Arizona State University, Tempe, Arizona

A number of recent as well as classic ideas suggest that there are constraints and limits on the explanatory role that computational, mathematical, and neural net models of visual and other cognitive processes can play that have not been generally appreciated. These ideas come from mathematics, automata theory, chaos theory, thermodynamics, neurophysiology, and psychology. Collectively, these ideas suggest that the neural or cognitive mechanisms underlying many kinds of formal models are untestable and unverifiable. Models may be good descriptions of perceptual and other cognitive processes, but they cannot *in principle* be reductive explanations nor can we use them to predict behavior at the molar level from what we know of the neural primitives. This discussion is an effort to clarify the appropriate meanings of these models, not to dissuade workers from forging ahead in the modeling endeavor, which I acknowledge is progressing and is making possible our increasingly deep appreciation of plausible and interesting cognitive processes.

What do formal models of cognitive processes mean? What can models do, and what can they not do? My purpose in this article is to consider these questions for a wide variety of formal models—mathematical, computational, neural network, statistical, cognitive flowchart, and symbolic, as well as others that are just arriving on the scene.¹ The past decade, in particular, has seen an explosion of interest in the development of formal models of cognitive processes. But it must not be overlooked that this kind of modeling has been with us virtually since the first availability of digital computer technology. (See, e.g., Farley & Clark, 1954; Rosenblatt, 1962; Selfridge, 1959; and Widrow, 1962, among others; and for a comprehensive review of the history of this field and the mathematical relationships among the various theories, see also Grossberg, 1988. Anderson & Rosenfeld's 1988 compendium of significant papers in neurocomputing is also a good review of the history of one cluster of formal models.) The recent impetus for the renewed interest in computer simulations and mathematical models of cognitive pro-

cesses, however, has been generated by, among other instances, the work of Anderson (1968) and of Grossberg (e.g., 1968, 1969; summarized in his 1982 book), Marr's important book (1982), and one of the newest and most influential works: the twin volumes by Rumelhart, McClelland, and the PDP Research Group (1986) and McClelland, Rumelhart, and the PDP Research Group (1986).

Contemporary modeling efforts have already successfully produced formal analogues of many perceptual and other cognitive processes. Processes have been "simulated," "analogized," and "represented" by programs and mathematical formulae that seem to *perform* behaviorally in much the same way as do the organic systems under study. Nevertheless, these impressive simulations must be evaluated in terms of what they *mean*, not just in terms of how good a functional fit to the time course of some behavior has been achieved. This is an often overlooked epistemological aspect of many contemporary models.

In this article, I will use a vocabulary that may have some special meaning to me and therefore must be defined more precisely. In particular, my use of the terms *reductionism*, *explanation*, and *description* must be clarified before we can proceed with this discussion. Such definitions have a rich history in philosophy and psychology, and certainly they have been subject to many different interpretations. Cummins (1983), Dretske (1988), and Haugeland (1981) are among a growing cohort of philosophers who have dealt with the meanings of these words.

This work was supported by Contract N00014-88-K0603 from the Office of Naval Research. I would like to express my thanks to several anonymous reviewers of an earlier draft of this article. In addition, I wish specifically to thank Stevan Harnad, Joseph Lappin, and James T. Townsend for their detailed and insightful commentaries, which helped make this document a clearer and more worthwhile contribution than it otherwise would have been. Correspondence should be sent to the following address: William R. Uttal, Department of Psychology, Arizona State University, Tempe, AZ 85287.

Fully appreciating the frailty of any attempt to define terms like these to the satisfaction of all, I nevertheless propose that the following meanings be understood in this article:

By *reductive explanation*, I mean a precise, unique, and valid statement of the particular internal mechanisms (neural, logical, or cognitive) by means of which a system carries out its function. *Reductionism*, in this sense, implies an analysis of a system into its components.

By *description*, I mean a mathematical (or neural or computational) representation of the functions, processes, or transformations of a system that maps the course of a process (i.e., its behavior), but in a way that is separable from assumptions about the specific internal physical mechanisms that might implement that function. I will argue that most models are, at best, descriptions no greater in explanatory content than an equation fitting a curve.

I distinguish between a *formal model* and the more general notion of a *model*, in terms of the degree to which specific assumptions are made about the relationships among the components. Of course, there is a continuum of models from the formal to the informal, but the kinds of formal models with which I am concerned here all involve specific statements of how the parts interact. Specifically, they involve mathematics, computer programs, or neural nets.

In the present article, I present an alternative interpretation of the relationships between formal models and the underlying physical reality—the neural substrate—on the one hand, and the psychological reality—the cognitive processes—on the other. I examine some issues and ideas that speak to the validity, verification, and testability of formal models in a reductionistic sense. That is, I am concerned with the following question: What limits, if any, exist concerning what these models can say about the internal workings of the mind on the one hand and the brain on the other?

The epistemological inquiries exemplified by the present article are necessary, because formal models and their attendant neural and cognitive corollaries are often misinterpreted as valid reductionistic explanations that realistically describe the actual function of the brain or of cognition. I am not concerned here with the details of how models work or what algorithmic functions must be carried out to more or less accurately simulate cognitive processes. Rather, I am concerned with what formal models may or may not have to say in their role as putative explanations of the systems they analogize. The issue at hand is this: What, in fundamental principle, can we expect of a formal model in terms of its ability to *explain* as opposed to its ability to *describe*?

The constrained, limited, and conservative answers I offer to this question should be familiar to some of my readers, and should not come as tremendous shocks or surprises to anyone. The concept that there might be specific limits to scientific measurement or achievement has wide currency in other scientific fields. Physics lives comfortably with indeterminacy principles, the speed of

light constraint, the impossibility of perpetual motion, and the need for statistical analyses of large aggregations of particles. Unlike physicists, reductionist cognitive scientists seem sometimes to behave as if their models can be conclusively explanatory and that it is only a matter of time, effort, and new technology before a full analysis of the mechanisms underlying brain–mind relationships will be forthcoming.

It is important that my readers appreciate the essence of the epistemological argument that I make here. That essence is that *formal models cum theories of this genre can have nothing definitive to say about the physiological or logical mechanisms by means of which these processes are carried out; they are neutral, in fundamental principle, with regard to internal mechanisms*. I am also asserting that *it is not simply a matter of technical refractoriness or the logistics of complexity but one of fundamental principle* that impedes our attempts to draw verifiable conclusions concerning internal workings from external behavior—the scientific approach that has come to be called *reductionism*.

It should also be noted in this context that an accumulation of inadequate information does not necessarily lead to a correct conclusion. Though there have been many references in the psychological literature to the idea of “converging evidence,” nothing has been established that supports the contention that something that is in principle impossible would become possible if a large number of different experiments were carried out, each of which was also incapable of peering into internal mechanisms. This is the stuff of superstition, not science.

The question of what models and theories mean has been at the core of the philosophy of science at least since the days of John Stuart Mill. Mill influenced subsequent scholars such as the very modern Ernst Mach, who asserted that “scientific investigations do not penetrate nature; they organize in economical fashion that which experience presents to us” (as quoted on p. 291 of Vol. 3 of Edwards, 1967).²

Recently, contemporary philosophers of science have analyzed the problem of explanation and reductionism specifically in terms of the fundamental nature of the relationship between mind and brain and, in other contexts, they have considered the role that formal (computational, mathematical, and neural network) models may play in explanation. Fodor (1968) has made the same point about analogy that I will make here. Speaking of a theorem proving machine, he asserts that “the machine would be inadequate as an explanation of theorem proving insofar as it does not do proofs in the same way as a human mathematician does them” (p. 136).

Massaro (1986, 1988) has also raised questions about the role played by one kind of formal model—connectionism—in psychological theory. He notes that some connectionist models may be too powerful in the same sense that Fourier analysis is too powerful to indicate anything specific about internal structure. His point argues against the wide applicability of a type of contemporary psychological the-

ory that has gained a wide, but uncritical, popularity in recent years.

Similarly, Dennett's (1981) "type intentionalism" theory asserts that "every mental event is some functional, physical event or other, and the types are captured not by any reductionist language, but by a regimentation of the very terms we ordinarily use . . ." (p. xix). Dennett suggests this point of view as a further development of an earlier "Turing Machine Functionalism"—the idea that although all mental states are physical, they need not be embodied in the same physical mechanism even if modeled by the same Turing algorithm. That is, different mechanisms, having the same functions, could be explained in terms of some physically neutral language.

Cummins (1983) makes the same point in a different way. In distinguishing between instantiation and reduction, he notes that "uniform instantiation is not sufficient for reduction, since a variety of modal truths will hold for the target property that do not hold of the instantiation" (p. 24).

Other equally distinguished scholars have distinctly different views about explanation. Churchland (1986), in strongly arguing for the possibility of reduction from mind to brain, misinterprets the kinds of reasons that will be most compelling in countering the arguments that there will ultimately be a verified neuroreductionistic theory of mind. The "common theme," according to her, of current objections to mind-brain reduction is the fallacious argument that "mental states are not physical states" (p. 346). I agree with her on the inadequateness of any form of this argument. However, there are some much more powerful physical arguments for asserting that neuroreductionism is impossible without denying that mind is a process of a physical substrate, as will shortly be presented.

Rumelhart, McClelland, and the PDP Research Group (1986) argue for the utility of their modeling approach because models of the type they offer "hold out the hope of offering computationally sufficient and psychologically accurate mechanistic accounts of the phenomena of human cognition" (p. 11). In opposition, it should be noted that although sufficient to *simulate* some processes, neither the PDP models nor any other formal model can be shown to also be necessary and, thus, unique. Thus, there is no way that they can be verified as being "psychologically accurate" in any sense beyond that of description.

By way of introducing the following discussion, it must also be understood that I appreciate that it is unlikely that many of the claims that I make can be rigorously proven in a strictly mathematical sense. I expect this article to be controversial because of the logical complexity of the problem as well as because of the prior commitment of individuals to some of the key premises of our current scientific Zeitgeist. I present these ideas as an ensemble of suggestions that raises issues not well enunciated in contemporary psychological science. It is the ensemble, rather than the individual argument, that has convinced me that

we must reconsider some of these logical and philosophical premises.

Some Possible Constraints on What Models Can Do

The issues of explanation and description considered here are closely related, but not identical, to the question asked by Teller (1984) in her important discussion of "linking propositions." The kind of neuroreductionistic linking propositions with which Teller was concerned dealt with the nature of the relationships invoked by sensory psychobiologists between neurophysiological and psychophysical data. Is it plausible, for example, to claim that a particular neural response "is not inconsistent with," "accounts for," "explains," or "is" the equivalent of a particular psychophysical phenomenon? Teller carefully defined the kinds of relationships that could occur between the neural and psychophysical domains and concluded by briefly raising the issue of the "empirical testability of linking propositions." She did not propose an answer (although she suggests that testing should be possible in some cases), but her metatheoretical study laid important groundwork for the analogous question I now consider. Formal models are, in a sense, more elaborate "linking hypotheses," in that they describe a precise relationship between the formal structure of a computer program or mathematical algorithm and a cognitive process, instead of the relationship between neuronal and psychophysical measures. By showing similarity in process, the conjecture is that they demonstrate similarity in structure. The following ideas suggest why this conjecture may not be valid.

The combinatorial explosion: Problems arising from numerosity. Let us first consider the complexity of the brain in simple numerical terms. At low magnification, the nervous system appears as a wonderfully complex aggregate of a number of interacting centers and nuclei. Yet even at this relatively gross level (which is probably not the appropriate one at which to consider the information-processing aspects of cognition), the known degree of interconnectedness is large and will surely be found to be larger as anatomists and electrophysiologists track the connections among the increasing number of centers now known to be involved in even as relatively straightforward cognitive functions as are exemplified by some visual processes. The works of Allman (1981), Kaas (1978), and Van Essen (1985) illustrate this complexity. Allman (1981) notes that more than 14 visual areas have already been discovered in the brain of the owl monkey; Van Essen (1985) provides a map (facing his p. 268) of the macaque cortex, suggesting that as many as 40 regions may have at least partial visual function in that animal. Van Essen goes on to note that although the human brain has not been mapped in as detailed a manner, if the same kind of structural organization should hold in humans as it does in simpler primates, there would be room for "literally dozens of visual areas in the human brain" (p. 283).

The situation is made even more complicated by the intricate series of interconnections among these many visual regions. Van Essen (1985) lists 92 distinct interconnecting pathways among these visual areas in the macaque that have been reported by neuroanatomists working on visual brain anatomy.

The question one must ask in the context of brain system analysis is the following: Is it plausible to carry out an analysis of a system of this complexity (40 areas interconnected by 90 or so tracts)? One way to approach this question is to assume that the interconnecting pathways are simple—that is, to consider them to be homogeneous communication channels conveying a collective positive or negative valence analogous to electrostatic charge or gravitational force. (Realistically, however, it must be remembered that the system of visually involved nuclei is not simply organized; the areas that are connected seem to communicate by means of reciprocal, multichannel tracts consisting of many complexly encoded, independent fibers.) This simplification will allow us to initially consider this problem in a manner analogous to how one would consider the multibody problems of mechanics and physics.

The results of a few comparisons are not encouraging for the prospects of any enterprise analyzing such a system. Astronomers have taught us that, in general, a deterministic, exact solution is not available for even as small a system as three mutually interacting bodies (unless one of the three has negligible mass) given current methods of analysis. This classic proposition was shown to be true, in general, a hundred years ago by Henri Poincaré. Therefore, it would not be possible, in principle, to deal with a problem of the complexity of the 40 or so nuclei of the visual brain in such a direct manner if the organizational analogy I draw between brain nuclei and planets is appropriate. The system of simultaneous equations describing this “multibody” interaction would be vastly too complicated to be solved directly, even by a very powerful computer.

Astronomers, on the other hand, have developed some methods for dealing with interacting objects when their number is very large (>200), by considering their discrete forces as the inseparable elements of a continuous field of force. Thus, large numbers of interacting forces can be thought of as collectively contributing to a field with a density that can be specified at all of the regions in a space. It is this potential density that then becomes the critical factor in the multidimensional Fourier calculations that are used for this kind of analysis. In this manner, quite satisfactory approximate solutions to the multiple interactions of the forces from the many point-like objects that collectively produced the hypothetical potential field can be determined and the overall behavior of the system estimated.

There is a difficulty with even this kind of approximation, however. The mathematical procedure for producing an approximate global solution to a multibody problem of this magnitude requires that the forces exerted by and

the behavior of the individual objects be submerged into those of the field. Furthermore, approximate field solutions to the multibody problem do not work when the number of objects is less than 200 and the density of the objects becomes too low for a field to be simulated. In the intermediate case ($3 < n < 200$), the only available solution to the multibody problem is that of actual physical or numerical simulation. In other words, the system must actually be constructed in the form of a computer program (or some other kind of mechanical analogue) and then allowed to run its course as certain perturbing inputs are applied to it. Depending on the nature of the system, this approach can sometimes provide a useful solution, but computability usually depends greatly upon the uniformity and simplicity of the network and upon interconnecting forces that can be represented by relatively simple functions. Otherwise, the situation becomes analogous to the general three- or n -body problem and can only be represented by an unmanageable and probably unsolvable system of equations.

Hence, although it may be possible statistically to describe what happens to the ensemble of objects (e.g., the stars in a galaxy or the neurons in a cortical nucleus), the behavior of the individual element (i.e., the star or the neuron) is not specifiable. Thus, just as is the case with the pressure of gas in a rigid container, the free-running electroencephalograph, the impulsive stimulus-evoked brain potential, or the behavior of a social system, the behavior of the components cannot be deduced from the ensemble activity of the entire system. In all such cases, the various involved sciences use molar attributes (e.g., pressure, EEG voltage, social movements, psychological phenomena) to describe the ensemble's behavior. All suffer from an inability to reduce those molar attributes to the behavior of the individual components, simply because there are so many of them and the number of possible interactions grows rapidly as additional components are added.

The problem of complexity has been discussed so far in the context of the 40 or so nuclei of the brain's visual system. Yet even at this level of complexity, it seems to be just as intractable as a rigorous model of, for example, the dynamics of the solar system. The problem of neural interactions becomes far more combinatorially explosive when we consider networks of the individual neurons that compose these nuclei. The unsolvable network of 40 nuclei connected by 90 or so tracts is replaced by 10^{13} neurons, each potentially connected to a thousand or more of its fellows in a disorderly and irregular way.³ There is no simple and direct way to calculate the number of plausible neural subnetworks of 10^{13} neurons and 10^{16} possible connections that may well be present, because we do not know very much about the size of a “psychologically interesting neural array.” For the sake of discussion, however, let us assume that 10^6 neurons make up a network that is “interesting” in the sense that it constitutes an autonomous module capable of some semblance of plausible psychological behavior. Combinatorial math-

ematics can then be used to show that the number of possible networks of $n = 10^{13}$ elements taken $m = 10^6$ at a time must be very large. Specifically, the approximation formula $(ne/m)^m$ estimates the number of combinations of n elements taken m at a time. Even ignoring the constant e , the value of this expression is approximately 10^7 raised to the millionth power—a formidable number, indeed!⁴ For all practical purposes, therefore, there are innumerable combinations of a million neurons possible even in small sections of the brain. The problem here is that with so many alternatives, it is unlikely that any neural investigation could unravel the network, and no model could ever bring order to the underlying mechanisms of the brain. Redundant circuits, as well as simple numerosity, would always obscure the brain's true organizational details.

The enormity of the combinatorial problem faced in dealing with realistic neural networks is understood by mathematicians, but it is usually ignored by neuroreductionistically oriented cognitive scientists. Mathematicians like Stockmeyer and Chandra (1979), for example, in an illuminating consideration of the size of computational algorithms, note the existence of "simple" combinatorial problems that, although they involve only a few elements ($n \sim 64$) and are demonstrably solvable in principle, would necessitate a "computer as large as the universe running for at least as long as the age of the universe" for their solution. Thus, they are asserting that although this computer might have as many as 10^{126} logical elements—a number approximately equal to the number of protons in the universe—it still would not be able to solve problems encountered in even simple games like chess and checkers in anything corresponding to a reasonable amount of time in a complete and rigorous manner.⁵ The point is that analysis of such a system at the neural level is algorithmically unlikely, because of its enormous complexity and numerosity at this critical level of informational interaction—the level at which cognition is most likely encoded.

Stockmeyer and Chandra's (1979) point should not be misunderstood. They are not asserting that such combinatorial problems are infinite or unsolvable in principle, but rather that although some problems of this genre are finite and solvable in principle, they remain rigorously unsolvable in practice: The computational algorithms required to evaluate the many possible descendent possibilities of even the simple sequential process typified by a board game are beyond the power of any conceivable computer.

It seems clear that the nervous system is an example of a "game" much more complex than the ones Stockmeyer and Chandra (1979) mention, and that a computational analysis or model based on the simulated interaction of the discrete parts of even local regions is practically impossible to achieve unless very severe constraints, such as an almost crystalline regularity and totally regular and simple forces of interaction, are imposed.

Thus, the few really successful (and truly neural) network theories depend a great deal on the presence of reg-

ularity and repetition of the elements in the network for their successes.⁶ The algebraic model (e.g., the classic Hartline and Ratliff model—see Ratliff, 1965, p. 110—of the Limulus eye) and the differential equations of some of the more modern versions (e.g., those used by Cohen & Grossberg, 1983) are all analyses that assume uniform arrangement of the constituent model neurons. The behavior of one neuron is prototypical of all of those in the network, although interactions within the network produce the differential behavior of each neuron.

It was the presence of a homogeneous and regular network of interconnections among ommatidia that permitted Ratliff and Hartline (1959) to simulate the spatial Mach band by observing the response of a single neuron while moving a light source over a stationary Limulus retina. Unfortunately, this success was also the source of a great deal of unjustified optimism that these early accomplishments could be generalized to the irregular networks that are typical of most parts of the vertebrate nervous system.

Mathematics and nature are one-way systems: Entropy cannot be reversed to order in closed systems. Nature is governed by fundamental laws that may also be interpreted as contradicting the dreams of neuroreductionist and computational modelers. One of these fundamental laws is the second law of thermodynamics, which, in the domain of physics, can be stated as follows: No body can give up heat to a body of higher temperature. In the domain of information theory, a counterpart of this law is: Isolated systems tend towards a state of greater randomness or greater probability than the one in which they currently reside.

One major implication of this fundamental law of nature is that there is a directionality to time. That is, the flow of events cannot be reversed to its origins. In other words, while it is very easy to mix the batter for a cake, you cannot "unmix" it. This has led the second law of thermodynamics to be referred to as "Time's Arrow" by such scholars as Blum (1951).

Although physicists have not yet rigorously established this widely accepted physical principle (few reputable scientists would waste their time, however, developing a perpetual motion machine), it may also have a psychological logical equivalent. That is, the molar behavior of a highly evolved, self-organizing, complex system cannot be analyzed or reduced to its logical and elemental mechanical (i.e., neural) underpinnings. The degree or level of complexity at which this conjecture becomes operative for the nervous system is uncertain, of course, but what I say below about chaos may speak directly to this issue. A chaotic situation, like a system with a high degree of entropy, is characterized by randomness, and randomness suggests unanalyzability.

If correct, this corollary of psychoneural nonanalyzability would hold, it should be noted, even though there is no question that brain-mind systems can be created and can develop. The conjecture I make is that thermodynamics implies that although simple laws can lead to complex behavior, the behavior of the complex system is not reducible to the events that generated it. There may

be some debate about the degree of complexity at which this conjecture would hold (i.e., where does entropic complexity begin to exert its influence), but I maintain that an understanding of the exact internal mechanisms of a system with as high a degree of entropy or irregular complexity as the brain is prohibited by the second law's implication of irreversibility.

The behavior of familiar physical systems provides support for this conjecture. When observing the behavior of a gas, we cannot say anything about the current state of individual particles solely on the basis of the ensemble attributes such as pressure or volume. To observe the state of the individual particles we must turn to local measurements of position and velocity. Simple numerosity (of the particles) and entropic irregularity then play havoc with any attempt to detail the attributes of the individual particles as well as to go from those individual states to the molar measure. Furthermore, there would always be a vast number of possible alternative pathways from "then" to "now" even if we did not have to deal with the numerosity problem. The discussion of chaos theory to be presented later elaborates on this point.

Thermodynamics thus suggests that the longed for informational bridges may not exist between microscopic and macroscopic levels of analysis of complex ensemble systems because of the irreversibility axiom. I am further suggesting that the brain is analogous to a highly entropic system. Therefore, it is not a priori true that the attributes of one level of a complex system like the mind are analyzable or representable in the language of another level (for example, neural nets or computer programs), even if the two languages should denote exactly the same physical reality.

The black box: Problems arising if the brain is a closed automaton—Moore's theorem. Can we hope to reduce cognitive processes to their underlying mechanisms? Can we use mathematics to uniquely determine internal structure from external observation of function? These are two of the intimately related questions with which this article is concerned. In this section, I will deal with another aspect of these perplexities—the difficulty that engineers refer to as the "black box" problem.

One of the most compelling arguments against the idea that we can reduce the externally observed behavior of a system to its internal mechanisms is to be found in the work of E. F. Moore (1956). Moore is a mathematician who has worked in the field of *finite automata* theory. A finite automaton is a device that is deterministic (as opposed to probabilistic, random, or stochastic) and has a finite number of internal states, a finite number of possible input "symbols," and a finite number of output "symbols."

One of Moore's (1956) theorems (Theorem 2) is directly related to the argument made here that the reductive (as opposed to the descriptive) aspects of formal models cannot, in principle, be tested, verified, or validated.

THEOREM 2: *Given any machine S and any multiple experiment performed on S, there exist other machines experimen-*

tally distinguishable from S for which the original experiment would have had the same outcome. (p. 140)

Moore goes on to assert:

This result means that it will never be possible to perform experiments on a completely unknown machine which will suffice to identify it from among a class of all sequential machines. (p. 140)

It is certainly debatable whether or not Moore's theorem is applicable to the modeling of human cognitive processes. But, since the brain meets all of the definitional criteria mentioned previously, it seems that the mind-brain can be considered to be a finite automaton (even if very large numbers of states and neurons are involved, the number is finite). Hence, this theorem, because of its broad applicability to the class of systems to which we humans seem to belong, raises severe questions about the goals of some of the work currently carried out by theoretical and experimental psychologists and cognitive modelers. It is particularly limiting for cognitive scientists who attempt to understand, measure, and define the exact nature of internal, representational, and functional mechanisms by behavioral means alone. But the concept implicit in Moore's theorem was also reflected in some of the classic arguments in peripheral sensory coding theory—for example, in the long-lasting controversy between trichromatic and opponent mechanisms dating from the time of a simpler kind of psychophysics. There too, the internal mechanisms could not be determined from behavioral testing and awaited electrophysiological techniques for their resolution. Another example of a related controversy is that between the proponents of serial and parallel processing, respectively, in various cognitive tasks. Again, the implication of this theorem is that the issue may be unresolvable by means of any behavioral test.

Moore's theoretical answer to the question of the analyzability of automata may mean that we can never, in principle, decide between alternative representational models for memory and perception. If his theorem is right, and if it can be validly generalized to human mentation, then many of psychology's reductive goals become unobtainable. Indeed, psychology's situation may be even worse than that of Moore's generalized automaton. In our case, we do not have direct access to what are arguably the essential outputs of the system—the mental responses or percepts. Visual perception, for example, is a private experience that is not directly sharable with others. Ideally, science deals with the explicitly measurable, yet in our science it is only an indirect indicator—overt behavior—that is measured.⁷ The underlying cognitive processes, the responses of true interest, are not directly accessible to either the psychologist or the observer.

Surprisingly, Moore's theorem seems to recommend a kind of dualistic, behavioristic epistemology (as opposed to the monistic metaphysics—i.e., the idea that the mind is nothing but a process of the brain—to which all psychologists implicitly subscribe) without denying the existence of mind or cognition or whatever other term one

would prefer to use to describe those hidden brain processes. That is, his theorem suggests that we will always have to be satisfied with some kind of a neobehaviorism in which it is accepted that psychological processes cannot be reduced to neural nets or logic-like functions.

Perhaps the most controversial point that I shall make in this article is that, in my opinion, psychological studies are neutral with regard to the nature of internal mechanisms in the same way as mathematics. None of these approaches, therefore, can deal directly with the essential level of analysis of high-level cognitive processes—the very complex, intricate, and tangled interconnections of the neural network that are the specific responsible mechanisms. While they may be powerful heuristics and contain implications of a general kind, they can never, according to this point of view, be definitive.⁸

The intrapersonal privacy of cognitive processing and its attendant implied positivistic behaviorism is, therefore, just another way of expressing the engineer's "black box" proscription that puts into more familiar psychological terms the formalisms of Moore's very important theorem. The essence of any of these expressions of this controversial, central theme of our science is that it is not possible to uniquely determine the internal mechanisms of a closed system (the brain) by input-output (behavioral) analyses alone. Moore says that the specific internal mechanism can only be defined by looking inside the black box. However, even that strategy is not always sufficient, as will be indicated in the next section.

Chaos: Problems arising when deterministic systems give rise to apparent randomness. If it follows from Moore's theorem that it is necessary to look inside the black box to identify specific mechanisms, does this mean that we will always thereby be able to understand any given mechanism should we actually be able to "take the top off"? The answer to this question is assuredly *no*. We have already considered several reasons why this negative conclusion is inevitable—complexity, numerosity, and the combinatorial explosion. However, complexity and numerosity are just practical problems. It is not inconceivable, in spite of Stockmeyer and Chandra's view, that someday we might invent an ultrafast transform or a super computer that would allow us to overcome even that handicap. I shall now argue, however, that the achievable goals of modeling and reductionism are far more constrained than is implied by the practical barrier of relatively uncomplicated, though numerically overwhelming, combinatorics. It is currently asserted that the necessary original information (i.e., the initial state of the components of the system and the history of their interactions) that led to a particular outcome is not just too ponderous to grasp, but actually *no longer exists* in the present condition of any system of the complexity we are considering. Similarly, the current state of such a system is not just difficult to understand, but is impossible to understand.

These are the messages of modern chaos theory. They state that, for fundamental reasons, the behavior of complex systems may be both unpredictable with respect to

the future and irreducible to their underlying mechanisms.⁹ If these implications of chaos theory can be verified, then there exists a two-way barrier between explanatory theory and mechanism that is insurmountable, not just challenging. The implication is that we can neither reduce the psychologically molar to the neurally microscopic nor predict the molar from the microscopic.

In brief, chaos theory may be interpreted as arguing that our world is governed by the following postulates:

1. Perfectly deterministic molecular laws may govern the evolution of very complex behavior based on the myriad properties and interactions of very simple components.

2. This complex set of deterministic interactions, however, may result in a system that is chaotic and apparently random because of the pyramiding of small uncertainties. Thus, it may not be possible, in principle, to predict the future state of a system even if one knows the full details of its initial state (Grebogi, Ott, & Yorke, 1987).

3. The apparently chaotic system does not exhibit a pure form of randomness, however. It may still produce organized behavior, however unpredictable, at a macroscopic level. The point is that any examination of the microscopic components of the system may display apparent randomness in spite of that molar order.

4. Neither the apparently random behavior of the ensemble of microscopic elements nor the ordered global or molar behavior of the overall system may be reversible to their initial-state origins, because the information describing the interactions that led to the final state from the initial state is lost. It is lost not only because there is no possibility of maintaining the history of the many small interactions and uncertainties, but also because small uncertainties have pyramided into an information state in which the information literally no longer exists.

5. Chaos theory may be very useful in helping to locate and identify global forces operating on the system that are otherwise hidden. "Regular," "strange," or "chaotic" *attractors* (i.e., global forces emerging from the organization of the system, which act to direct its molar behavior) may be discerned in very complex, nonlinear systems. It may also be possible to predict global outcomes (such as the shapes of snowflakes or the dripping of water from a faucet). However, it must be appreciated that, in this case, chaos theory deals with the molar outcome and not with the microscopic details of the underlying structure. In this sense, it is very much like psychophysics or any other cognitive approach. Chaos theory tells us about the overall behavior of the system, but eschews any attempts at reductive analysis. Neither the fact that an attractor may be a point nor that other pertinent field-like forces may be well described detracts from the generalization that a microscopic analysis may be impossible.

Embedded in these postulates of chaos theory, there is a very important concept—apparent randomness—that ought to be made explicit. We must distinguish a dichotomy, in this case, between a system that is truly ran-

dom and one that is apparently random. The brain is obviously not truly random; our mental and physical realities constitute an existence proof that the apparent chaotic confusion of neural activity often leads to organized, coherent behavior. On the other hand, there is an apparent randomness about the intricate pattern of microscopic cellular activity that underlies that organization. When we examine the microstructure of a complex system like the interacting neurons of the brain, we typically find disorderly response patterns that cannot be distinguished from true randomness in spite of the system's molar functionality.¹⁰

Chaos theory, therefore, is essentially a theory of macroscopic function that is antireductionistic in essence. It deals with macroscopic "attractors" and forces, not with the microdetails of the organization of a system. It deals with nonlinear systems, and does so perhaps better than any other approach to that complex problem; but it does so in a holistic manner. Farmer, one of the leaders in the development of chaos theory, is quoted by Gleick (1987) as saying:

The trend in science, and physics in particular, has been towards reductionism, a constant breaking things down into bitty pieces. What people are finally realizing is that that process has a dead end to it. Scientists are much more interested in the idea that the whole can be greater than the sum of the parts. (p. 134)

Chaos theory thus suggests another source of difficulty in the understanding of complex systems, even when an enclosing "black box" can be "opened" to internal examination. These ideas have been elaborated by Crutchfield, Farmer, Packard, and Shaw (1986), in an especially clear discussion in which they explicitly state the implications of chaos theory for prediction. According to them, chaos "implies new fundamental limits on the ability to make predictions." They agree that microscopic determination is not always extrapolatable to detailed macroscopic outcomes. Although these authors go on to assert that some classes of random processes may actually be opened to a deeper understanding if one uses the mathematics of chaos theory, their thesis can also be read as implying that fundamental limits will emerge from this analysis that will constrain our ability to understand the breakup of smoke, the weather, the erratics of fluid motion, or the operation of a very large number of neurons in a cognitive process. In each of these cases, the output is so complicated that it must basically be considered to be random, thus precluding top-down reductionistic analysis as well as bottom-up predictions. Although chaos theory suggests that it is possible to designate some global attributes that integrate these apparently random microscopic states into some kind of a useful statistical measure (e.g., pressure, the ratio of turbulent to laminar flow, the electroencephalograph, or perceptual phenomena), reduction to the behavior of individual elements is also impossible according to their interpretation of chaos theory!

Most of the constraints on reductionism that I have already discussed emphasize complexity and suggest limits on analyzability *in practice*. Chaos theory suggests that

there may be even more serious limits *in principle*. These systems are not just complex and multivariate; they produce apparently random—that is, intrinsically unanalyzable—behavior.

Crutchfield, Farmer, Packard, and Shaw's (1986) discussion of chaos is not just mathematical esoterica; it is directly relevant to and specifically damning for some of the implicit programmatic goals of both neuroreductionistic and mathematical or computational theorists. This can be made most clear by the following excerpt from their work:

Chaos brings a new challenge to the reductionist view that a system can be understood by breaking it down and studying each piece. This view has been prevalent in science in part because there are so many systems for which the behavior of the whole is the sum of its parts. Chaos demonstrates, however, that a system can have complicated behavior that emerges as a consequence of simple, nonlinear interaction of only a few components.

The problem is becoming acute in a wide range of scientific disciplines, from describing microscopic physics to modeling macroscopic behavior of biological organisms. The ability to obtain detailed knowledge of a system's structure has undergone a tremendous advance in recent years, but the ability to integrate this knowledge has been stymied by the lack of a proper conceptual framework within which to describe qualitative behavior. *For example, even with a complete map of the nervous system of a simple organism, such as the nematode studied by Sidney Brenner of the University of Cambridge, the organism's behavior cannot be deduced.* Similarly, the hope that physics could be complete with an increasingly detailed understanding of fundamental physical forces and constituents is unfounded. *The interaction of components on one scale can lead to complex global behavior on a larger scale that in general cannot be deduced from knowledge of the individual components.* (p. 56; italics added)

It is, of course, important to discriminate between systems that are inherently indeterminate or unanalyzable and those that are not yet well understood. The point I am intentionally making throughout this article is that there are strong arguments that systems like the brain may never be fully understood, no matter how much effort or time is put into their study. The barrier is one of fundamental principle, not just computational load.

Models are only abstractions of full reality. In the context of establishing what a model means, it must be established just what a model is. Trivially, a model, whether made of neuromimes, mathematical terms, or plaster, is a reduced portrait or partial representation of a real something that is more complex, richer in detail, different in size, or existed at a different time. Like any other partial representation of an object, a model is less than that which it represents. A map does not include all of the information inherent in the real world that it models. Similarly, models, no matter how precisely they may predict function, do not include all of the details of the mechanical underpinnings of the system that they portray.

There is, however, a reverse side to this generality. Sometimes models may superimpose their own proper-

ties on our concept of the object being modeled. Hence, the modeler's conception of the object may reflect a property that the object does not actually possess. In this sense, therefore, the model may be more (rather than less) than the object it represents.

This line of thought has practical implications for psychophysical research. In some cases, the model of a phenomenon is so incomplete that it does not operate under the same constraints as the human observer does. The model may predict an unconstrained performance level that is never achieved (or is not achievable) by the observer. A clear example of this kind of discrepancy can be found in the work of Barlow (1978), in which it was shown that observers performed only half as well as an ideal model predicted they should in a texture discrimination task. Obviously, inefficiencies of one kind or another may be operative in the real situation that are not taken into account by the model.

The other kind of discrepancy between a model and reality becomes possible because the model is but an incomplete manifestation of the reality. This occurs when observers are able to take advantage of some attribute of the stimulus that has not been programmed into the model but that allows the observer to perform far better than predicted by the model. We have encountered such a situation in work in my laboratory, which was reported in Uttal, Davis, Welke, and Kakarala (1988). We used a simple three-dimensional surface-fitting process to model the ability of an observer to reconstruct the shape of a surface from a very sparse sample of dots randomly positioned (on that surface). The mathematics of the situation is straightforward. It is mathematically impossible to fit or recognize the exact stereoscopic surface with fewer than five dots if the surface is a quadric and fewer than seven dots if the surface is a cubic. Nevertheless, our observers performed significantly better than chance (12.5%), correctly naming one of the eight forms with as few as two dots (24%), and much better than that with three and four dots (40% and 50%, respectively).

A similar result was obtained in a slightly different field of perceptual research, in which motion rather than disparity was used as a cue for three-dimensional form. Braunstein, Hoffman, Shapiro, Andersen, and Bennett (1987) also found that their observers did better than had been predicted by their mathematical model. Under a rigidity constraint programmed into the model, the observers should have required three sequential views of four points to perceive the shape of a rotating dotted form. The observers, however, were able to do so with either smaller numbers of points or smaller numbers of views. Without a doubt, the observers in these experiments made use of other information that was not incorporated into the models. It is clear that the models that expected too little in both of these cases were as inadequate as models (e.g., that invoked by Barlow, 1978) that expect too much of the human visual system.

In retrospect, both of these two research groups (Braunstein et al., 1987; Uttal et al., 1988) were able to par-

tially explain their respective discrepancies and thus to improve the models. However, the salient point in the present context is that a model is not the same as the reality. No matter how good our models, there will be some residual difference between models and reality. To the extent that a discrepancy exists, this can also be interpreted as representing a barrier to explanation and limit the degree to which a model can be tested and verified. Of course, the goal of theory is to reduce these discrepancies, but the residual difficulty concerns whether or not any model can actually close the gap between itself and realities of the complexity with which we are concerned here. Some of the arguments presented earlier suggest that, in general and in principle, they may not be able to do so.

Perception is inferential: A deterministic, algorithmic model may simply not be relevant or realistic. Another possible difficulty with computational theories of any kind of cognitive process is that the tool used to implement the program and/or the mathematics in these simulations is often either a highly deterministic entity—the digital computer—or a system of equally deterministic analytic equations. These methods and tools of analysis are characterized by a highly *deductive* mode of operation.

Whether they are serial or parallel, electronic or optical, computer systems are designed to operate in a precisely causal and rigidly deterministic manner. Given a particular set of initial conditions, the final outcome will always be the same. That is, each subsequent state is the outcome of a precisely determined transformation of a preceding state. Acknowledgedly, this mode of operation may be quite suitable to describe what happens in the peripheral parts of a sensory system that are more concerned with the communication of sensory information than with the transformation of that information into perceptual experience.

However, if anything is well established in psychology, it is that the higher levels of human cognition do not operate in this strictly deterministic fashion, but by a much softer, *inductive*, kind of logic that is only loosely determined by the stimulus conditions. The perceiving brain seems to use processes (of which we have but the barest understanding) of inference, interpretation, and a kind of high-level adaptive logic that are probably not the same rules as those governing the algorithmic, deterministic computer or the analytical forms of mathematics usually applied in this domain. Such a set of inferential rules may be simulated to some degree on a computer, but there is little evidence that the generalization from one type of problem to another that is prototypical of human cognition occurs in contemporary computer simulations. (See Dreyfus's 1979 discussion of the key issue of generalization and pyramiding of results in artificial intelligence studies. His work also has implications for the formal modeling of cognitive processes.) Such generalization should certainly be a hallmark of any model that presents itself as a valid explanation of organic mentation.

Whether thought is inferential or mediated (i.e., rationalist) as opposed to determinist (i.e., empiricist) has concerned psychologists and philosophers for centuries. The kind of indirect, mediated cognition that is now generally accepted to characterize our mentation was, in previous incarnations, known as "unconscious inference," as proposed by Helmholtz (1925/1968). It is also of the same genre as the formal and rigorous statement of inference made by Bennett, Hoffman, and Prakesh (1989), in what is certain to become one of the most controversial and influential studies in contemporary psychology. One thesis of this important work is not only that perception is inferential, but also that it is not, in general, either deductively valid or unbiased. Absence of deductive validity would mean that the rules of standard logic (and presumably the rules of computational or mathematical logic) do not hold for the mechanisms underlying our thoughts and perceptions. To truly mirror the real world of cognition, our models should be inferential and rationalist; unfortunately, they have tended to be determinist and empiricist.¹¹

My point here is that no current computational or connectionist model comes close to operating as the mind does. All of the algorithmic and deterministic strategies that characterize this current theoretical effort are, in terms of their basic premises, different in kind from the adaptive and inferential processes exhibited by the brain. Computational models are, according to this point of view, not even good first approximations (at their most fundamental axiomatic roots) of the underlying mechanisms of cognition. Thus, *a priori*, they must be invalid as reductive explanations of mental processes, however well they describe the behavior.

None of this should be interpreted to mean that complex, adaptive, inferential behavior cannot emerge from the concatenation of elements that are individually deterministic. We know that it can. Indeed, knowing (i.e., our own self-awareness) is itself an existence proof of that hypothesis.¹² But no one has yet shown any satisfactory bridges between the deterministic neural elements and inferential thought of the kind that could bridge that gap, and many of the issues raised so far suggest that such a quest would be unsuccessful.

The cumulative microcosm does not equal the macrocosm. There is no denying the enormous accomplishments of chemists and biologists in understanding the components or, as computer engineers call it, the "technology" of the cognitive engine between our ears. We have enjoyed a revolution in the last half century in both neurochemistry and neurophysiology. However, all of this elegant knowledge may be irrelevant to the understanding of cognitive processes, since it tells us little about the essential level of psychoneural equivalence—the details of the interneuronal interactions. Although physiologists and biochemists working in the neurosciences at this microcosmic level often claim that their studies have molar cognitive or behavioral implications, and psychologists likewise claim that their findings have neural implications,

a more critical look at both assertions suggests that they are not justified. Neurochemical studies of how a synapse works or of the macromolecules involved in the propagation of the action potential, although stupendous achievements in their own right, are comparable to studies of the materials of which a transistor is made. It is the organization of neurons that counts, however. Neurophysiological studies, by practical necessity, deal only with the interconnections of a very small number of neurons and thus may be irrelevant.

The essential point is that cognition arises out of the organization of these components, not out of their metabolic characteristics or their local means of communication. The biochemical and physiological properties of a neural network may be totally irrelevant to its informational attributes.

Understanding neurons or their underlying biochemistry, therefore, is not tantamount to understanding cognition; that understanding would come from the discovery of the laws governing "reasonably" large interacting ensembles of neurons. As we have seen, combinatorics, thermodynamics, and chaos theory suggest that such an analysis of psychologically significant networks may not be a realizable goal.

There is a substantial controversy concerning how much we really know about the salient cellular interactions, mechanisms, and forces that underlie complex neural interactions. Does such knowledge presently exist? Some steps have purportedly been taken in this direction (e.g., Hawkins & Kandel, 1984). Hawkins, Kandel, and their colleagues have made many other important contributions to our understanding of some simple kinds of learning in invertebrates. They have elucidated possible neural circuits responsible for conditioning and habituation, and they have even made contributions to our understanding of the kinds of chemical processes that explain changes in synaptic conductivity and that might therefore mediate changes in these networks associated with experience. But, contrary to some others in the field, Hawkins and Kandel are extremely constrained in their interpretation of the results. They say:

We do not provide any data suggesting that higher orders of conditioning must necessarily emerge from the basic cellular mechanisms of more elementary forms of learning. Nor would we argue that participation of the cellular mechanisms that we have outlined here in higher orders of conditioning would provide evidence for their role in yet more sophisticated types of learning. (p. 389)

Thus, even beyond the problems associated with the extremely large numbers of small interactions, it can be argued that we do not yet know the nature of the microscopic rules in vertebrate neuronal systems. This may be a surmountable challenge in the future; in principle, electrophysiologists like Kandel and his colleagues can be expected to make progress in this direction. But it is yet to be established whether they can do so in practice for any

vertebrate central nervous system, given the variability of the neural structure from specimen to specimen and the possibility of many redundant encodings of the same process.

The germane question here is not what we know, but rather what we can know. It is not yet clear that we can know what is needed or that past progress is really what it seems to have been at first glance when extrapolated from the neurophysiological to the cognitive domains.

If one looks at the past history of neural net modeling, it is clear that only in the simplest cases, such as the relatively regular neural nets of the *Limulus* eye or the cerebellum, has any simulation been verified. Even then, the theoretical accomplishments have been rare. Many so-called "neural net" theories, although perfectly good descriptions of some cognitive process, are based on neural (as opposed to generalized parallel) networks in only the loosest sense. Often they only represent discrete embodiments of mathematical (i.e., descriptive) formulae and are, in fact, neutral with regard to internal mechanisms. In other cases, very simple networks consisting of only a few neurons have been used to simulate the function of the much more complex networks that almost certainly must be instantiated in real organisms. (See, e.g., Anderson, 1972; Grossberg, 1982.) These simple systems often do a creditable job of simulating or imitating some of the molar process aspects of a cognitive or perceptual transformation, but it is likely that they do so by means of vastly different logics than those used by the enormous ramifications of neurons in the brain. Theories or models using such regular or minimal systems need not operate by means of the same rules that a complex, irregular system consisting of many idiosyncratically interconnected components does, even if their behaviors may be superficially the same.

A further issue germane to this problem is that the appropriate level of analysis critical for the information-processing aspects of the brain are still uncertain: Networks of neurons seem to me to be the most primitive salient level. But what are the important higher echelons of interactions? Even the simplest perceptual process may involve so many neurons that it will represent a fully chaotic system in the sense described earlier. Are the rules the same at all levels of anatomic organization and magnification? It has been suggested that the two general domains at which science operates—the macroscopic, human scale and the microscopic, invisible scale—operate under the influences of laws that may be quite different. Prigogine (1980) reminds us of Eddington's (1928) suggestion that there are primary laws that control the behavior of individual particles and secondary laws that regulate the behavior of ensembles. Of course, this may only be a reflection of the practical combinatorial difficulty of computing the results of a large number of interactions or of the limits of contemporary theory. It may also reflect a misinterpretation of a quantitative change as being a qualitative one (i.e., concatenated multiple interactions may only appear to be different in kind than the individual

interactions). On the other hand, perhaps Eddington also had an insight into something more fundamental that is also reflected in thermodynamics, chaos theory, and psychology, among other sciences that deal with complex systems arising from the interaction of myriad components and operating in the macroscopic world. Perhaps, even though there may be no difference in quality in some absolute sense, there are differences in the ways in which we must approach problems at the two different levels. These differences may "only" be the practical outcome of complexity and nonlinearity; but this would make them no less real in terms of the epistemology governing our studies.

I will not pursue this speculation further, but whatever the future may hold, the bridges that we wish to build between the microcosm and the macrocosm certainly do not yet exist: We have not yet been able to build satisfactory explanations of Eddington's "secondary" laws in terms of the "primary" ones. Neither brain actions nor statistical mechanics have been reduced to simple deterministic laws. Chaos theory and combinatorics suggest that they may not be reducible.

Mathematics is descriptive, not reductive. I will now reiterate a point made briefly in the introduction to this article and alluded to earlier. Mathematics is an enormously powerful tool, but it is not omnipotent. The insights given to us by classical analytical methods as well as by contemporary computational techniques have been profound. Nevertheless, it is important to be clear about just what can be done with mathematics and what cannot.

It is a well-appreciated fact among students of quantum physics that some solutions to differential equations are physically meaningless. That is, some of the solutions of equations purporting to model physical systems have no significance but are merely superfluous and extraneous baggage that emerges from the mathematics rather than from the physics of the real world being described by that particular notation. It is not necessary to look to abuse mathematics to find other examples. We have rules to cover some of these cases, such as the prohibition against dividing by zero. Without that rule, Newton's inverse square law would provide energy forever for the world from the power exerted at the point of contact of two mutually attracted bodies.

The point is that all parts of a mathematical model may not be relevant to the system being modeled. In the same way, mathematical models (or neural models couched in mathematical terms) may also fall victim to this difficulty and denote mechanisms that are not likely to be found within the brain.

A closely related, if not identical, attribute of mathematics is that this powerful technique is intrinsically non-reductive. This powerful system of thought can work in situations in which the premises that govern the physical reality are totally distinct and separable from those governing the mathematical manipulations. One obvious example is the all-too-powerful Fourier analytical method. It is sometimes overlooked by enthusiasts that Fourier-type

analysis allows the reduction of any one-, two-, or three-dimensional pattern that has a finite number of discontinuities and maxima and minima to sets of superpositionable orthogonal (usually sinusoidal) functions. *This form of analysis holds, even though there is nothing that corresponds to sinusoidal generating mechanisms in the physical system being analyzed.* Indeed, the great power of the Fourier method is that any function, regardless of its actual underlying mechanism, can be analyzed as if it was the actual physical result of the superposition of the outputs of a set of real orthogonal function generators. This strength, however, is also the source of the great weakness of Fourier analysis when it comes to its role in explanation and reduction and the source of its seductive power to mislead. All too often, contemporary theoreticians impute actual mechanical reality to what is just a mathematical abstraction.

Fourier analysis, like any other mathematical tool, is a means of describing physical systems by providing a set of numerical values to represent some unquantified feature, such as shape. The actual physical reality may be totally different from the terms of this powerful descriptive device and totally opaque to reductive analysis.

Other instances of the premises of mathematics being separable from those of the system being described are trivially evident in any kind of curve-fitting algorithm. All that is required for a perfect fit of a model to a mechanism is that there be sufficient degrees of freedom in the mathematics to accommodate the behavior of the system.¹³ But again, there is no reason to assume that a polynomial equation (to pick a trivial example) that fits some process is actually physically instantiated as a sum of a series in the physical system. One has only to work briefly with analog computers to appreciate the fact that the assumptions about process and mechanism are separable and distinct, and that the relation between homologs is not the same as that between analogs.

It is important to emphasize that what I say in the present context is also true of all other forms of theory that aspire to play an explanatory, as opposed to a descriptive role, in science. All models may be accused of fitting the data spuriously if they are being used as explanations, but a fit is not spurious if description is all that is being attempted. Then, the fit is either satisfactory or not, and a precise criterion for choosing among the various alternatives exists. Implications about internal mechanisms must be handled separately. It is in this context that this article is intended to serve as a gentle reminder, since I doubt that there is general disagreement with the principle enunciated here, only a casual forgetfulness.

What, then, does mathematics do for us? My answer is that mathematics is a superb way of *describing* a system. It allows us to transform the uninterpretable into a quantitative, interpretable form and often to describe and predict the present and future course of a system. In some cases, it even allows us to determine the laws governing interactions among objects or transformations in their form. Mathematics can simplify, beautify, quantify, and,

perhaps most important of all, force a precision of language and thinking upon the theoretician that raises scientific inquiry from speculation to analysis. The latter attribute would be enough, even if mathematics had no other virtues, to guarantee its continued value for all kinds of intellectual pursuits, however compelling the arguments for its limits might seem to be.

But mathematics is not a panacea, or even a help, for those seeking to explain internal mechanisms. In the words of Cutting (1986), "mathematics is too powerful to provide constraints on information; it models truth and drives with equal felicity" (p. xi). No mathematical procedure can peer deep into a "black box" that has been produced either because the system is really closed or because of the complexity of the interaction or numerosity of its components, and create information about internal structure where there is none. In short, mathematics is so powerful that it is incapacitated in a fundamental way: We tend to infer more than it implies.

The conclusion to which we are directed by these considerations is that there may very well be a vast difference between the real psychobiological mechanisms and the inferences about mechanisms drawn from formal models. In addition, there may be no way to measure or evaluate those differences. My assertion is that computational, mathematical, or neural net models can, in principle, say nothing rigorous about internal mechanical reality in a reductionistic sense—they are neutral in practice and principle—regardless of how perfectly they describe function or process. They may be elegant and fully descriptive, but they can never be explanatory in the reductive manner suggested by some contemporary theoreticians.

Gödel's incompleteness theorem. Though the following interpretation of this important mathematical theorem and its relevance to neuroreductionism is in some disrepute, one can argue that self-analysis (i.e., by a brain or any other kind of self-organizing system as well as mathematics) can never be effective. Gödel (1962) established that there are always going to be theorems within any mathematical system that cannot be proven by the axioms of that system. This proof, according to some, can be extended to mechanical and biological systems by way of analogy. Thus, it is often asserted that the brain cannot analyze itself and indeed, that self-organizing automata cannot simulate themselves (see, e.g., the discussion on this topic by Hofstadter, 1985). Even if one is not willing to make the leap from mathematics to neurophysiology, some would argue that Gödel's theorem leads directly to the conclusion that mathematics must be limited in terms of its applicability to cognition simply because of the nature of mathematics.¹⁴

Conclusion

In the preceding section, I have discussed some of the reasons that computational, mathematical, and neural models cannot be tested and verified as valid reductive explanations of the internal workings of the nervous system or of the mind. Though they may serve as powerful de-

scriptions and ingenious heuristics, they are fundamentally neutral about the internal mechanisms of all kinds of cognition.

I do not think that these limitations are as ominous for my fellow psychobiologists as they may at first seem. They do not portend the end of psychology, the neurosciences, or artificial intelligence research; they only clarify and redefine what it is that we are about. Our colleagues in the physical sciences are increasingly coming to appreciate the limits of their methodology and theory, and they have flourished in spite of what are virtually the same kinds of limits as those that should discipline theories of the mind. Indeed, some of the most significant contributions of the physical sciences have consisted of the identification of those limits. They have also modified their views about nature to incorporate indeterminacy and probability rather than strict causation. The distinction between model and reality is becoming understood. They also appreciate the message of chaos theory that, as Prigogine (1980) has said, "The future is not included in the past. Even in physics, as in sociology, only various possible 'scenarios' can be predicted" (p. xvii).

Physics has thus come to appreciate its own epistemological limits. Mathematicians, who have also long known that there are some theorems that cannot be proven or are indeterminate, live comfortably with those limits in their activities. Why, then, should we psychologists not also accept the existence of similar constraints, limits, and indeterminacies in our part of the physical world? The systems we deal with are so much more complex than the ones dealt with by physicists that progress is necessarily slower than it is with the simple systems they study. To suggest that we can achieve with the brain what they could not achieve with as simple a mechanism as the solar system is an inexcusable arrogance. To assume that we can validate the reductionistic implications of a theory that is at best a process analog seems to me to reflect a lack of understanding of how science works. Raising false hopes or setting theoretical goals that have already been proven to be unobtainable (e.g., by automata, chaos, or computational theorems) reflects an ignorance of very relevant contemporary developments in related fields of science.

In conclusion, I want to counter a major potential misinterpretation of my argument. By no means should it be interpreted as a pessimistic view of our science. The strategies now being pursued for studying the brain and mind are timely and interesting, and, if properly interpreted and understood, they are among the jewels in contemporary science's crown. Admittedly, however, I have spoken mainly about what these models do *not* mean and what they *cannot* do. There is a further set of important epistemological issues, however, that must also be confronted. These have to do with what these models *do* mean and what they *can* do. I have argued that the main thing they do is to describe function and molar behavior. This idea can be particularized further. Formal models are

capable of precisely specifying the transformations that occur between the inputs and outputs of a system. They can quantify otherwise vague notions of the amount of information that can be processed. They can also allow us to test various alternative input conditions and determine whether or not the system is likely to respond in the same way to them. Indeed, they may allow us to predict the stability of a system that carries out the transformations that must occur.

There are other important reasons why we should proceed with the exercise of formal modeling:

1. First, mathematical, computational, and neural net modeling, however limited, is vastly superior to the prequantitative, verbal, and imprecise speculations that characterized most earlier forms of psychological theorizing about cognitive processes. Formal models, constructed in the form of computer programs, mathematical equations, or even physical networks demand a precise definition of the rules that might govern interactions of system components as well as the exact nature of the components themselves. They can, at least, distinguish between the plausible and the implausible, even if they do not allow us to distinguish between the real and the plausible. This kind of reasoning and reflection is, therefore, a huge leap forward toward the understanding that we seek, in terms of the precision of the language that we use as well as the exact nature of the processes of interaction.

2. A second reason for proceeding with formal modeling is that this is all that is necessary for some purposes. In spite of the ideas presented here that reductive analysis of the brain is practically and theoretically impossible, that we cannot validate or verify any model of cognition in a truly reductive manner (except in the most trivial cases of peripheral sensory communication), and that there thus are good reasons for changing our ideas about what models mean, there are other goals of formal modeling. We may wish only to imitate or describe some human information process. If so, all we really have to know is what kind of transformation occurs during the process and not necessarily how the nervous system carries it out. Thus we can manufacture a perfectly good artificial intelligence system capable of performing useful visual functions, for example, by observing and modeling human vision, making (however unverifiable they may be) informed assumptions about what are, at best, analogous internal mechanisms, and writing a program using any one of a large number of functionally equivalent algorithms. It is not likely that the system so developed will be able, in the foreseeable future, to perform with the same flexibility and competence as natural vision (although there is no obvious *a priori* reason why it should not), but it may function at a perfectly satisfactory level for some practical task or give a perfectly good mathematical description of the relevant phenomenology to a perceptual psychologist.

3. A third argument for formal modeling is that it is intellectually elegant and satisfying. It would not be

enough to throw up our hands in despair that we can never achieve complete analytic explanation. Psychologists should note that the implications of the second law of thermodynamics regarding the impossibility of a perceptual motion machine did not dissuade mechanical scientists from proceeding with their studies, nor did molecular numerosity deter them from studying gas dynamics. Interaction with nature in this manner has often produced heuristics that stimulate beautifully inventive and aesthetically pleasing solutions to problems in unexpected ways, even if they are not exactly the same as the ways in which a real system may accomplish the same task.

4. A fourth reason to pursue this work is that even though contemporary models may consist of description more than reductive explanation, description itself is of extreme value. Description may tell us about the historical course of a process, and it may allow us to extrapolate the course of events into the future. To be able to predict events, even though the underlying mechanism remains impenetrable, is itself a perfectly remarkable achievement.

5. Finally, however sharp the discontinuity between the knowable and the unknowable may be, it is clear that we have not come close to the limits of understanding to which I have alluded. We are progressing and improving our understanding of perception and other cognitive processes. There has been a constant closure on what I believe is the "truth" of the neural and logical basis of perception (if not of cognition in general). We have progressed from electrotonic field theories (Koffka, 1935) to single-cell models (Barlow, 1972), to analytic function models (Marr, 1982), to parallel neural net theories (McClelland et al., 1986; Rumelhart et al., 1986; Uttal, 1975, 1981) of how the brain works—and this is a trend in the right direction. The fact that a barrier may exist somewhere ahead should not deter us from striving to approach that barrier as closely as possible.

On the other hand, the enumeration of these reasons for going ahead with this type of modeling should not obscure the central thesis of this article—that no formal model is verifiable, validatable, or even testable with regard to internal mechanisms. As elegant as it may be, and as well as it may describe a process, no model can be excluded or included as a truly reductive, explanatory theory of any cognitive process by means of the kinds of tests of necessity and sufficiency that are made in other kinds of scientific enterprises. Though this conclusion may be disappointing to some psychological theoreticians, I believe it to be generally accepted throughout other areas of science. In the long run, it will help us on our way to the next phase of understanding what these wonderful models really mean.

In the meantime, we must not forget that some of the more mature fields of psychology that are less currently in vogue still have critically important roles to play. We must search out the facts, describe the phenomena, and measure the processes and transformations of mind and behavior to provide the fundamental material required for

the continued and fruitful pursuit of the formal modeling enterprise.

REFERENCES

- ALLMAN, J. (1981). Reconstructing of the evolution of the brain in primates through the use of comparative neurophysiological and neuroanatomical data. In E. Armstrong & D. Falk (Eds.), *Primate brain evolution: Methods and concepts* (pp. 13-28). New York: Plenum.
- ANDERSON, J. A. (1968). A memory model utilizing spatial correlation functions. *Kybernetik*, **5**, 113-119.
- ANDERSON, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences*, **14**, 197-220.
- ANDERSON, J. A., & ROSENFELD, E. (Eds.) (1988). *Neurocomputing: Foundations of research*. Cambridge, MA: MIT Press.
- BARLOW, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology. *Perception*, **1**, 371-394.
- BARLOW, H. B. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research*, **18**, 637-650.
- BENNETT, B. M., HOFFMAN, D. D., & PRAKESH, C. (1989). *Observer mechanics*. San Diego: Academic Press.
- BLUM, H. F. (1951). *Time's arrow and evolution*. Princeton, NJ: Princeton University Press.
- BRAUNSTEIN, M. L., HOFFMAN, D. D., SHAPIRO, L. R., ANDERSEN, G. J., & BENNETT, B. M. (1987). Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception & Performance*, **13**, 335-343.
- CHURCHLAND, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, MA: MIT Press.
- COHEN, M. A., & GROSSBERG, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *Transactions of the IEEE, SMC-13*, 815-826.
- CRUTCHFIELD, J. P., FARMER, J. D., PACKARD, N. H., & SHAW, R. S. (1986, December). Chaos. *Scientific American*, pp. 46-57.
- CUMMINS, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.
- CUTTING, J. E. (1986). *Perception with an eye for motion*. Cambridge, MA: MIT Press.
- DENNETT, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.
- DRETSKE, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
- DREYFUS, H. L. (1979). *What computers can't do: The limits of artificial intelligence* (rev. ed.). New York: Harper & Row.
- EDDINGTON, A. (1928). *The nature of the physical world*. New York: AMS Press.
- EDWARDS, P. (Ed.) (1967). *The encyclopedia of philosophy* (Vol. 5). New York: Collier Macmillan.
- FARLEY, B. G., & CLARK, W. A. (1954). Simulation of a self-organizing system by a digital computer. *Institute of Radio Engineers Transactions of Information Theory*, **4**, 76-84.
- FODOR, J. A. (1968). *Psychological explanation*. New York: Random House.
- GLEICK, S. (1987). New images of chaos that are stirring a science revolution. *Smithsonian*, **18**(9), 122-137.
- GÖDEL, K. (1962). On formally undecidable propositions. New York: Basic Books.
- GREBOGI, C., OTT, E., & YORKE, J. A. (1987). Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. *Science*, **238**, 632-638.
- GROSSBERG, S. (1968). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proceedings of the National Academy of Sciences*, **59**, 368-372.
- GROSSBERG, S. (1969). On learning of spatiotemporal patterns by networks with ordered sensory and motor components. *Studies in Applied Mathematics*, **49**, 135-166.
- GROSSBERG, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Boston: Reidel.

- GROSSBERG, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1, 17-61.
- HAUGELAND, J. (Ed.) (1981). *Mind design: Philosophy, psychology, artificial intelligence*. Cambridge, MA: MIT Press.
- HAWKINS, R. D., & KANDEL, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, 91, 375-391.
- HELMHOLTZ, H. VON. (1968). Excerpts from *Treatise on Physiological Optics* (3rd ed.). In R. M. Warren & R. P. Warren (Eds.), *Helmholtz on perception: Its physiology and development*. New York: Wiley. (Original work published 1925)
- HOFSTADTER, D. R. (1985). *Metamagical themes: Questing for the essence of mind and pattern*. New York: Basic Books.
- KAAS, J. H. (1978). The organization of visual cortex in primates. In C. Noback (Ed.), *Sensory systems of primates*. New York: Plenum.
- KOFFKA, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace.
- MARR, D. (1982). *Vision*. San Francisco: W. H. Freeman.
- MASSARO, D. W. (1986). The computer as a metaphor for psychological inquiry: Considerations and recommendations. *Behavior Research Methods, Instruments, & Computers*, 18, 73-92.
- MASSARO, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory & Language*, 27, 213-234.
- MCCLELLAND, J. L., RUMELHART, D. E., & THE PDP RESEARCH GROUP. (1986). *Parallel distributed processing: Vol. 2. Psychological and biological models*. Cambridge, MA: MIT Press.
- MOORE, E. F. (1956). Gedanken-experiments on sequential machines. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 129-153). Princeton, NJ: Princeton University Press.
- PRIGOGINE, I. (1980). *From being to becoming*. San Francisco: W. H. Freeman.
- RATLIFF, F. (1965). *Mach bands: Quantitative studies on neural networks in the retina*. San Francisco: Holden-Day.
- RATLIFF, F., & HARTLINE, H. K. (1959). The response of Limulus optic nerve fibers to patterns of illumination on the receptor mosaic. *Journal of General Physiology*, 42, 1241-1255.
- ROSENBLATT, F. (1962). *The principles of neurodynamics*. Washington, DC: Spartan Books.
- RUMELHART, D. E., MCCLELLAND, J. L., & THE PDP RESEARCH GROUP. (1986). *Parallel distributed processing: Vol. 1. Foundations*. Cambridge, MA: MIT Press.
- SEJNOWSKI, T. J., KOCH, K., & CHURCHLAND, P. S. (1988). Computational neuroscience. *Science*, 241, 1299-1306.
- SELFRIDGE, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanization of thought processes* (Proceedings of a symposium held at the National Physics Laboratory, November 1958; pp. 513-526). London: Her Majesty's Stationery Office.
- STOCKMEYER, L. J., & CHANDRA, A. K. (1979, May). Intrinsically difficult problems. *Scientific American*, pp. 140-159.
- TELLER, D. Y. (1984). Linking propositions. *Vision Research*, 24, 1233-1246.
- UTTAL, W. R. (1975). *An autocorrelation theory of form detection*. Hillsdale, NJ: Erlbaum.
- UTTAL, W. R. (1981). *A taxonomy of visual processes*. Hillsdale, NJ: Erlbaum.
- UTTAL, W. R. (1988). *On seeing forms*. Hillsdale, NJ: Erlbaum.
- UTTAL, W. R., DAVIS, N. S., WELKE, C., & KAKARALA, R. (1988). The reconstruction of static visual forms from sparse dotted samples. *Perception & Psychophysics*, 43, 223-240.
- VAN ESSEN, D. C. (1985). Functional organization of primate visual cortex. In A. Peters & E. G. Jones (Eds.), *Cerebral cortex* (Vol. 3, pp. 259-329). New York: Plenum.
- WIDROW, B. (1962). Generalization and information storage in networks of Adaline neurons. In M. C. Yovits, G. T. Jacobi, & H. Goldstein (Eds.), *Self-organizing systems*. Washington, DC: Spartan Books.
- of that system. The alternatives just listed are the ones most popular in contemporary science.
2. These words are virtually the same as those used by a group of eminent reductionists—Sejnowski, Koch, and Churchland (1988)—when they assert that models “should be considered a provisional framework for organizing possible ways of thinking about the nervous system” (p. 1305).
3. The number of neurons in the brain is a matter of debate. The traditional estimate is 10^{10} , but studies of the cerebellar cortex suggest that there are perhaps a thousand times as many fine-grained and regularly organized cells in that single region as in the rest of the brain in its entirety.
4. My thanks to James R. Johnston of Orion Research in San Francisco for suggesting this approximate solution to the combinatorial problem posed here. Both Johnston and I appreciate that this way of estimating the complexity of the nervous system is not completely satisfactory. However, it does give an idea of the enormous number of possible “mechanisms” that may exist in the brain.
5. Of course, computers do an increasingly good job at playing games like chess, but by methods that seem to be quite different from those used by humans. The exhaustive strategies used by the most powerful computer programs do not compare in basic principle to the holistic approach used by our nervous systems to play the same game. These two, strikingly different entities illustrate the great difference that can exist between a descriptive (analogous) model and a real system. Although the former may behave like the latter, it does so by entirely different internal mechanisms.
6. There is probably no better discussion of the family of relationships among various neural network models of many cognitive processes than Grossberg's (1988) elegant exposition of the principles and history of this field. It is important reading for everyone interested in this field, whether or not one agrees with the direction taken by neural network theoreticians.
7. In practice, of course, all sciences suffer from this same disability, so the ideal is never achieved.
8. Two of the reviewers of this paper, James T. Townsend and Stephen Grossberg, have already expressed disagreement with this assertion. I am grateful for their willingness to identify themselves in their reviews, but I am still not convinced that there is any evidence that argues strongly against it. The postulation of possible neural structures that behave comparably to organic systems is not an empirical test of the assertion. Mechanisms suggested by input-output or mathematical procedures do not exclude many other equally plausible alternatives and are examples of analogues as opposed to homologues of each other, according to Moore's theorem. The argument I make here is that this issue cannot be resolved empirically, but only in principle. Nevertheless, this is a complicated, subtle, and terribly important issue for contemporary psychology that will continue to attract many strong proponents of a view antithetical to my own.
9. I am grateful to James T. Townsend for reminding me of the fact that complex systems can produce very regular behavior and simple systems can produce very irregular behavior. We are often able to predict the behavior of complex systems and often cannot do so for simple systems. The antonyms “simple” and “complex” refer not to the number of component parts, but rather to the number of combination rules that drive the behavior of the ensemble.
10. I am well aware of the regular anatomic patterns of ocular dominance and orientation sensitivity in the visual brain and related findings from other senses. However, I do not believe that these elegant results detract from the conjecture. I am referring here to the dynamics of the individual neuron's activity, not the structural organization of these tissues. It does not seem to me that anyone has yet successfully disentangled the interactions of more than a few neurons, nor do I expect that this task will be accomplished in the general sense that I mean here.
11. A definition of these terms, as they are used here, may be found in Uttal (1981 and 1988).
12. It is also always possible that the probabilistic quantum effects of the particles that make up neurons may be reflected in their behavior. This speculation has some support from what we know of the quantum effects in visual transduction.
13. The history of learning theory in particular is filled with examples of mathematical models selected with just enough degrees of free-

NOTES

1. In using the term *formal model*, I am referring to any analysis of a system that uses the constructs, operations, rules, terminologies, and components of another field of science to propose a specific structure

dom to represent whatever function was under consideration at the moment. The various parameters in the models took on a physical and reductionistic realism that is hard to accept in retrospect.

14. I am especially grateful to both Oliver Hill and James T. Townsend for encouraging me to include the Gödel theorem in my argument after their reading of an early draft of this paper. So much controversy

has revolved around this point that I had not intended to include it. I now believe that it is very relevant.

(Manuscript received October 24, 1989;
revision accepted for publication February 5, 1990.)

Notices and Announcements

20th Annual Meeting of the Society for Computers in Psychology New Orleans, Louisiana November 15, 1990

The 20th Annual Meeting of the Society for Computers in Psychology will be held at the Hyatt Regency Hotel in New Orleans on November 15, 1990. The meeting will include presentations, discussions, and times to preview software and hardware. All areas of psychology are featured, including research, education, clinical practice, and industrial applications.

For further information, contact Sarah Ransdell, Department of Psychology, University of Maine, Orono, ME 04469 (BITNET address RANSDELL@MAINE) or C. Michael Levy, Department of Psychology, University of Florida, Gainesville, FL 32611 (BITNET address MLEVY@UFFSC).

31st Annual Meeting of the Psychonomic Society New Orleans, Louisiana November 16-18, 1990

The 31st Annual Meeting of the Psychonomic Society will be held in New Orleans, November 16-18, 1990. The meetings will begin Friday morning and continue until Sunday at noon. The headquarters hotel will be the Hyatt Regency.

The program and hotel reservation cards will be mailed to members and associates in September. A copy of the program will be published in the November issue of the *Bulletin of the Psychonomic Society*.

For further information, please contact the secretary-treasurer of the Society: Cynthia H. Null, Psychology Department, College of William & Mary, Williamsburg, Virginia 23185 (Telephone: 804-221-3882; BITNET: WPSLCHN@WMMVS).